NCEE 2009-4036

U. S. DEPARTMENT OF EDUCATION

The Enhanced Reading Opportunities Study

Findings from the Second Year of Implementation

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

The Enhanced Reading Opportunities Study

Findings from the Second Year of Implementation

NOVEMBER 2008

William Corrin
Marie-Andrée Somers
James J. Kemple
Elizabeth Nelson
Susan Sepanik
MDRC

With

Terry Salinger
Courtney Tanenbaum
American Institutes for Research

Paul Strasberg, Project Officer
Institute of Education Sciences

NCEE 2009-4036
U.S. Department of Education

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

U.S. Department of Education
Margaret Spellings
Secretary

Institute of Education Sciences
Grover Whitehurst
Director

National Center for Education Evaluation and Regional Assistance
Phoebe Cottingham
Commissioner

November 2008

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, under contract no. ED-01-CO-0111/0001 with MDRC.

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.

Contents

Appendixes

List of Exhibits

Box

Table

Table

Table

Table

Figure

Figure

Acknowledgments

This study represents a collaborative effort among the authors and the staff from the participating school districts and schools, the program developers, our colleagues at MDRC and American Institutes for Research (AIR), and Institute of Education Sciences (IES) staff. The study has benefited especially from the time, energy, and commitment put forth by staff in the participating school districts to implement the two literacy programs used in the Enhanced Reading Opportunities (ERO) study, to allow access to classrooms, and to respond to requests for data.

At the U.S. Department of Education, Paul Strasberg, Marsha Silverberg, Phoebe Cottingham, and Ricky Takai at the Institute of Education Sciences provided helpful support and guidance on the design and execution of the evaluation and in the development of the report. Braden Goetz and Valerie Randall-Walker at the Office of Elementary and Secondary Education provided invaluable support to the school districts in their efforts to implement the supplemental literacy programs and meet the demands of the evaluation.

The study's technical working group provided valuable insights on the evaluation design, data analysis, and early versions of the report. We thank Donna E. Alvermann, Donald L. Compton, Robinson Hollister, Mark W. Lipsey, Robert H. Meyer, Christopher Schatschneider, Timothy Shanahan, and Catherine Snow for their expertise and guidance.

The listed authors of this report represent only a small part of the team involved in this project. Linda Kuhn and the staff at Survey Research Management managed and conducted the follow-up testing and survey data collection effort.

At AIR, Nancy Lang, Suzannah Herrmann, Kathryn Drummond, and Courtney Zmach conducted site visits and conducted phone interviews. Christopher Doyle and Andrea Olinger coordinated data management and conducted phone interviews. Nancy Lang processed data and ensured the thoroughness of the fidelity ratings.

At MDRC, Edmond Wong assisted with data collection and provided programming and analysis support. Corinne Herlihy and Kristin Porter served as school district coordinators. Daniel Fallon oversaw the ordering of the literacy assessment. Shirley James and her staff entered data. Gordon Berlin, Alison Black, Howard Bloom, Fred Doolittle, Corinne Herlihy, John Hutchins, Robert Ivry, Janet Quint, and Pei Zhu provided substantive expertise through their thoughtful comments on, and reviews of, this report. Edmond Wong and Mario Flecha assisted with report production. Robert Weber edited the report, and Stephanie Cowell and Inna Kruglaya prepared it for publication.

The Authors

Disclosure of Potential Conflicts of Interest[1]

The research team for this evaluation consists of a prime contractor, MDRC, Inc., of New York City, NY, and two subcontractors, American Institutes for Research (AIR) of Washington, DC, and Survey Research Management (SRM) Corporation of Boulder, CO. None of these organizations or their key staff has financial interests that could be affected by findings from the evaluation of the two supplemental literacy interventions considered in this report. No one on the eight-member Expert Advisory Panel, convened by the research team once a year to provide advice and guidance, has financial interests that could be affected by findings from the evaluation. One member of the Expert Advisory Panel, Dr. Timothy Shanahan of the University of Illinois at Chicago, participated only in an early (2005) panel meeting on the study design. Subsequent to that meeting, he developed a commercial literacy intervention targeted to striving middle-school readers that might either compete with or be used along with the two programs for high school students chosen and evaluated as part of the current study. Dr. Shanahan had no role in the selection of the study programs or in the analysis of evaluation data.

[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.

Executive Summary

This report presents findings from the Enhanced Reading Opportunities (ERO) study — a demonstration and rigorous evaluation of two supplemental literacy programs that aim to improve the reading comprehension skills and school performance of struggling ninth-grade readers. The U.S. Department of Education's (ED) Office of Elementary and Secondary Education (OESE)[1] is funding the implementation of these programs, and its Institute of Education Sciences (IES) is responsible for oversight of the evaluation. MDRC — a nonprofit, nonpartisan education and social policy research organization — is conducting the evaluation in partnership with the American Institutes for Research (AIR) and Survey Research Management (SRM).

The present report — the second of three — focuses on the second of two cohorts of ninth-grade students to participate in the study and discusses the impact that the two interventions had on these students' reading comprehension skills through the end of their ninth-grade year. The report also describes the implementation of the programs during the second year of the study and provides an assessment of the overall fidelity with which the participating schools adhered to the program design as specified by the developers. While this report focuses primarily on implementation and impacts in the second year of the study, comparisons between the first and second year of the study are also provided.[2] The key findings discussed in the report include the following:

- **On average, across the 34 participating high schools, the supplemental literacy programs improved student reading comprehension test scores by 0.08 standard deviation. This represents a statistically significant improvement in students' reading comprehension (p-value = 0.042).**

- **Seventy-seven percent of the students who enrolled in the ERO classes in the second year of the study were still reading at two or more years below grade level at the end of ninth grade, relative to the expected reading achievement of a nationally representative sample of ninth-grade students.[3] One of the two interventions — Reading Apprenticeship Aca-**

[1]The implementation was initially funded by the Office of Vocational and Adult Education (OVAE), but this role was later transferred to OESE.

[2]James J. Kemple, William Corrin, Elizabeth Nelson, Terry Salinger, Suzannah Herrmann, and Kathryn Drummond, *The Enhanced Reading Opportunities Study: Early Impacts and Implementation Findings,* NCEE 2008-4015 (Washington, DC:, U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, 2008).

[3]Forty percent of ninth-graders nationally would be expected to score at two or more years below grade level on the same assessment.

demic Literacy (RAAL) — had a positive and statistically significant impact on reading comprehension test scores (0.14 standard deviation; p-value = 0.015). Although not statistically significant, a positive impact on reading comprehension (0.02 standard deviation) was also produced by the other intervention, Xtreme Reading. The difference in impacts between the two programs is not statistically significant, and thus it cannot be concluded that RAAL had a different effect on reading comprehension than Xtreme Reading.[4]

- **The overall impact of the ERO programs on reading comprehension test scores in the second year of implementation (0.08 standard deviation) is not statistically different from their impact in the first year of implementation (0.09 standard deviation), nor is each intervention's impact in the second year of implementation statistically different from its impact in the first year.**

- **The implementation fidelity of the ERO programs was more highly rated in the second year of the study than in the first year. In comparison with the first year, a greater number of schools in the second year of the study were deemed to have programs that were well aligned with the program developers' specifications for implementation fidelity (26 schools in the second year, compared with 16 schools in the first year), and fewer schools were considered to be poorly aligned (one school in the second year, compared with 10 schools in the first year).**

[4]It is important to note that the ERO study is an evaluation of a class of reading interventions, as represented by Xtreme Reading and RAAL, as well as an evaluation of each of these two programs separately. The purpose of the study is not to test the differential impact of these two interventions; while Xtreme Reading and RAAL do differ in some respects, they are both full-year supplemental literacy courses targeted at struggling adolescent readers that share many common principles, and hence there was no prior expectation that they would produce substantially different impacts. As noted below, the design of the study is such that programs are randomized to schools; however, the purpose of this randomization was to ensure that each program developer was assigned a fair draw of schools in which to implement its program, rather than to test for a differential impact between the two interventions. By this token, the statistical model chosen for the impact analysis does not utilize the school-level randomization feature of the research design; nor is the sample size large enough to detect policy-relevant differences in impacts across the two programs. Because Xtreme Reading and RAAL represent the same type of intervention, this study was designed to test their joint or overall impact. Statistical tests were used to confirm that the difference in impacts between the two programs is not statistically significant and, hence, that it is indeed appropriate to pool together the two program-specific impact estimates; these statistical tests are not appropriate for making inferences about the true difference in impacts between the two interventions.

The Supplemental Literacy Interventions

The ERO study is a test of supplemental literacy interventions that are designed as full-year courses and targeted to students whose reading skills are two or more years below grade level as they enter high school. Two programs — Reading Apprenticeship Academic Literacy (RAAL), designed by WestEd, and Xtreme Reading, designed by the University of Kansas Center for Research on Learning — were selected for the study from a pool of 17 applicants by a national panel of experts on adolescent literacy. To qualify for the project, the programs were required to focus instruction in the following areas: (1) student motivation and engagement; (2) reading fluency, or the ability to read quickly, accurately, and with appropriate expression; (3) vocabulary, or word knowledge; (4) comprehension, or making meaning from text; (5) phonics and phonemic awareness (for students who could still benefit from instruction in these areas); and (6) writing. The overarching goals of both programs are to help ninth-grade students adopt the strategies and routines used by proficient readers, improve their comprehension skills, and be motivated to read more and to enjoy reading. Both programs are supplemental in that they consist of a yearlong course that replaces a ninth-grade elective class, rather than a core academic class, and in that they are offered in addition to students' regular English language arts classes.

The primary differences between the two literacy interventions selected for the ERO study lie in their approach to implementation. Implementation of RAAL is guided by the concept of "flexible fidelity" — that is, while the program includes a detailed curriculum, the teachers are trained to adapt their lessons to meet the needs of their students and to supplement program materials with readings that are motivating to their classes. Teachers have flexibility in how they include various aspects of the RAAL curriculum in their day-to-day teaching activities, but they have been trained to do so such that they maintain the overarching spirit, themes, and goals of the program in their instruction.

Implementation of Xtreme Reading is guided by the philosophy that the presentation of instructional material — particularly the order and timing with which the lessons are presented — is of critical import to students' understanding of the strategies and skills being taught. As such, teachers are trained to deliver course content and materials in a precise, organized, and systematic fashion designed by the developers. Xtreme Reading teachers follow a prescribed implementation plan, following specific day-by-day lesson plans in which activities have allotted segments of time within each class period. Teachers also use responsive instructional practices to adapt and adjust to student needs that arise as they move through the highly structured curriculum.

Overview of the Study

Interventions. Reading Apprenticeship Academic Literacy (RAAL) and Xtreme Reading — supplemental literacy programs designed as full-year courses to replace a ninth-grade elective class. The programs were selected through a competitive applications process based on ratings by an expert panel.

Study sample. Two cohorts of ninth-grade students from 34 high schools and 10 school districts (2,916 students in Cohort 1 and 2,679 students in Cohort 2). Districts and schools were selected by ED's Office of Vocational and Adult Education through a special Small Learning Communities grant competition. Students were selected based on reading comprehension test scores that were between two and five years below grade level.

Research design. Within each district, high schools were randomly assigned to use either the RAAL program or the Xtreme Reading program during two school years (2005-2006 and 2006-2007). Within each high school, students were randomly assigned to enroll in the ERO class or to remain in a regularly scheduled elective class. A reading comprehension test and a survey were administered to students in the spring of eighth grade or at the start of ninth grade, prior to random assignment, and again at the end of ninth grade. Classroom observations in the first and second semester of the school year were used to measure implementation fidelity.

Outcomes. Reading comprehension and vocabulary test scores, reading behaviors, student attendance in the ERO classes and other literacy support services, implementation fidelity.

The ERO Evaluation

The supplemental literacy programs were implemented in 34 high schools from 10 school districts across the country. The districts were selected through a special grant competition organized by the U.S. Department of Education's Office of Vocational and Adult Education (OVAE). Experienced, full-time English/language arts or social studies teachers were self-selected and approved by ED, the districts, and the schools to teach the programs for a period of two years.

The ERO evaluation utilizes a two-level random assignment research design. First, within each district, eligible high schools were randomly assigned prior to the first year of program implementation to use one of the two supplemental literacy programs: 17 of the high schools were assigned to use RAAL, and 17 schools were selected to use Xtreme Reading. Each school implemented the same program in two school years: 2005-2006 and 2006-2007. In the second stage of the study design, eligible students within each of the participating high schools and in each year of the study were randomly assigned either to enroll in the ERO class

(the "ERO group") or to take one of their school's regularly offered elective classes (the "non-ERO group").

During the second year of the study, the participating high schools identified 2,679 ninth-grade students with baseline test scores indicating that they were reading two to five years below grade level (an average of 79 students per school). Approximately 57 percent of these students were randomly assigned to enroll in the ERO class, and the remaining students make up the study's control group and were enrolled in or continued in a regularly scheduled elective class.

Evaluation data were collected with the Group Reading Assessment and Diagnostic Examination (GRADE) reading comprehension and vocabulary tests and a survey.[5] Both instruments were administered to students at two points in time: a baseline assessment and survey in the spring of eighth grade and a follow-up assessment and survey at the end of ninth grade.[6] Follow-up test scores are available for 2,171 (81 percent) of the students in the study sample. To learn about the fidelity of program implementation, the study also includes observations of the supplemental literacy classes during the first and second semester of the school year.

Second-Year Implementation

Each ERO teacher (one per school) was responsible for teaching four sections of the ERO class. Each section accommodated between 10 and 15 students. Classes were designed to meet for a minimum of 225 minutes per week and were scheduled as a 45-minute class every day or as a 75- to 90-minute class that met every other day.

- **Of the 34 teachers who participated in the second year of the study, 25 had taught the entire first year of the study, and two had taught a portion of the first year (having replaced a teacher midyear). Seven teachers were new to the ERO programs at the start of the second year.**

During the second year of the project, the developers for each of the ERO programs provided three types of training and technical assistance to both new and returning ERO teachers: a three-day summer training institute in July or August 2006, booster training sessions during the 2006-2007 school year, and three 2-day coaching visits during the 2006-2007 school year. Prior to the summer institute, teachers new to the ERO programs also attended additional

[5]American Guidance Service, *Group Reading Assessment and Diagnostic Evaluation: Teacher's Scoring and Interpretive Manual, Level H;* and *Technical Manual* (Circle Pines, MN: American Guidance Service, 2001a, 2001b).

[6]In four of the 34 participating schools, baseline testing occurred in the fall of ninth grade rather than the spring of eighth grade.

training sessions at which they were taught the central strategies of the program being implemented in their school.

The study team assessed the overall fidelity with which the ERO programs were implemented in each school during the second year of the project. In the context of this study, "fidelity" refers to the degree to which the observed operation of the ERO program in a given high school was aligned with the intended learning environment and instructional practices that were specified by the model's developers. The analysis of implementation fidelity in the second year of the study is based on two field research visits to each of the 34 high schools — one during the first semester and one during the second semester of the 2006-2007 school year. The classroom observation protocols used in the site visits provided a structured process for observers to rate the characteristics of the ERO classroom learning environments and the use of ERO instructional strategies by teachers. The instrument included ratings for six characteristics (referred to as "constructs" from here forward) that are common to both programs, as well as ratings for seven program-specific constructs. For each construct, a category rating of 1 ("poorly aligned"), 2 ("moderately aligned"), or 3 ("well aligned") was given.

The analysis of the classroom observation ratings sought to capture implementation fidelity on two key overarching dimensions of both programs: the classroom learning environment and the teacher's use of instructional strategies focused on reading comprehension. A composite measure of implementation fidelity was calculated for each of these two dimensions by averaging across the relevant characteristics in the observation protocol. A composite rating of 2.0 or higher indicates that the school's ERO program was well aligned with the developers' implementation specifications; a rating of 1.5 to 1.9 means that the program was moderately aligned; and a rating of 1.0 to 1.4 means that it was poorly aligned. Following is a summary of key findings.

- **At the spring site visit, implementation fidelity in 26 of the 34 schools was classified as well aligned on both program dimensions. In seven schools, implementation was classified as moderately aligned with the program model on at least one of the two key program dimensions and as moderately or well aligned on the other dimension. In one school, implementation was deemed to be poorly aligned with the program models.**

The overall implementation of the ERO program in a given school was classified as well aligned if both the classroom environment and the comprehension instruction dimension were rated as being well aligned. According to the protocols used for the classroom observations, teacher behaviors and classroom activities in these schools were consistently rated as being well developed and reflective of the behaviors and activities specified by the developers. At the fall site visit, the implementation of the ERO programs in 20 of the 34 schools was classi-

fied as well aligned on both program dimensions, and, at the spring site visit, 26 schools had attained this benchmark. Because implementation fidelity in the majority of the study schools was deemed to be well aligned to the models, the study team also examined the number of schools whose implementation of the programs was "very well aligned" to developers' specifications (defined here as a composite score of 2.5 or higher on both program dimensions). At the spring site visit, implementation in 13 schools could be classified as such.

Conversely, a school's overall implementation fidelity was judged to be poorly aligned with the program model if the composite rating for either the classroom learning environment dimension or the comprehension instruction dimension was rated as poorly aligned. The ERO programs in these schools were not representative of the activities and practices intended by the respective program developers and were found to have encountered serious implementation problems on at least one of the two key program dimensions during the second year of the study.[7] At the fall site visit, implementation of the ERO programs in three of the 34 schools was classified as poorly aligned with the program models on at least one of the two program dimensions. At the spring site visit, implementation at one school was considered to be poorly aligned with the program models.[8]

- **The number of schools considered to be well aligned with the program developers' specifications for implementation fidelity was greater in the second year of the study than in the first year (26 schools in the second year, compared with 16 schools in the first year).**

At the spring site visit in the second year of the study, the ERO programs in 33 of the 34 schools reached an overall level of implementation fidelity that was at least moderately aligned to the program models (of these, 26 were considered to be well aligned). This is an improvement over the first year of the study, when 24 of the 34 schools had reached a moderate level of alignment at the spring site visit (of these, 16 schools were deemed to be well aligned). Also, during the spring site visit of the second year, only one school's implementation of the program was poorly aligned to the developers' specifications. This is lower than what was found during the first-year spring site visit, when 10 schools were ranked as poorly aligned on at least one of the two key program dimensions.

[7]In particular, poorly aligned implementation for a given dimension means that the classroom observers found that at least half of the classroom characteristics were not aligned with the behaviors and activities specified by the developers and described in the protocols.

[8]In the second year of the study, implementation-fidelity ratings were similar for the 25 schools where the ERO teacher taught two full years of the program and for the nine schools where the ERO teacher had replaced another teacher at some point during the study (an average rating of 2.5 for returning teachers and 2.4 for replacement teachers, out of a maximum of score 3).

Student Enrollment and Attendance in the ERO Classes and Participation in Literacy Support Activities

The study team collected data on the duration of the ERO classes as well as the frequency with which students attended the ERO classes and participated in other classes or tutoring services that aimed to improve their reading and writing skills.

ERO classes in the second year began an average of 2.3 weeks after the start of the school year and operated for an average of nine months. Eighteen schools started the ERO program on the first day of school, and five more schools started within the first two weeks that classes were in session. The remaining eleven started their ERO programs an average of seven weeks after the start of the school year. Among the students randomly assigned to the ERO group, 91 percent enrolled in the ERO classes, and 87 percent were still attending the classes at the end of the school year.

- **Students in the ERO group attended 79 percent of the scheduled ERO classes, and they received an average of 11 hours of ERO instruction per month.**
- **Students who were randomly assigned to the study's ERO group reported a higher frequency of participation in supplemental literacy services than students who were assigned to the non-ERO group.**

The ERO classes served as the primary source of literacy support services for students in the study sample. Although the largest difference in the use of supplemental literacy supports between the study's ERO and non-ERO groups occurred in students' participation in a supplementary school-based literacy class (an average of 75 yearly sessions for ERO students and 17 yearly sessions for non-ERO students), ERO students were also significantly more likely to report working with a tutor in school (an average of 30 yearly sessions, compared with 12 yearly sessions for non-ERO students).

Impact Findings

The GRADE assessment was used to measure students' reading achievement prior to random assignment (at "baseline") and then again in the spring at the end of their ninth-grade year (at "follow-up"). The GRADE is a norm-referenced, research-based reading assessment that is used widely to measure performance and track the growth of an individual student and groups of students. Because the two ERO programs focus primarily on helping students use contextual clues to understand the meaning of words, the reading comprehension subtest of the GRADE is the primary measure of reading achievement in this study, while the GRADE vocabulary subtest is a secondary indicator of the programs' effectiveness. Performance levels and

impacts on both subtests are presented in standard score units; students with a standard score of 100 points are considered to be reading at grade level.[9]

Following is a summary of the study's impact findings.

- **When analyzed jointly, the ERO programs produced an increase of 0.8 standard score point on the GRADE reading comprehension subtests. This corresponds to an effect size of 0.08 standard deviation and is statistically significant. The overall impact of the programs in the second year of implementation is not statistically different from their overall impact in the first year of implementation (0.09 standard deviation).**

The top panel of Table ES.1 shows the impacts on spring follow-up reading comprehension and vocabulary test scores across all 34 participating high schools in the second year of the study. The first row of data in the table shows that, on average, the reading comprehension test scores of students in the ERO group are 0.8 standard score point higher than the scores of students in the non-ERO group, which represents a statistically significant impact (its p-value is less than or equal to 5 percent).[10] Expressed as a proportion of the overall variability of test scores for students in the non-ERO group, this estimated impact represents an effect size of 0.08 (or 8 percent of the standard deviation of the non-ERO group's test scores).

Figure ES.1 places this impact estimate in the context of the actual and expected change in the ERO students' reading comprehension test scores on the GRADE from the beginning of ninth grade to the end of ninth grade. The bottom section of the bar shows that students in the ERO group achieved an average standard score of 84.6 at the start of their ninth-grade year. This corresponds, approximately, to a grade equivalent of 4.9 (the last month of fourth grade) and indicates an average reading level at the 14th percentile for ninth-grade students nationally.

The middle section of the bar shows the estimated growth in test scores experienced by the non-ERO group. At the end of the ninth-grade year, the non-ERO group was estimated to have achieved an average standard score of 89.3, which corresponds to a grade equivalent of 6.0 and an average reading level at the 23rd percentile for ninth-grade students nationally. This

[9]Based on the national norms used to calculate these scores, a standard score of 100 on the GRADE reading comprehension or vocabulary test is average for a representative group of students at the end of their ninth-grade year. The standard deviation of the standard score for both tests is 15.

[10]The impact estimates in Table ES.1 are regression-adjusted using ordinary least squares (OLS), controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The Enhanced Reading Opportunities Study

Table ES.1

Impacts on Reading Achievement, Cohort 2 Follow-Up Respondent Sample

Outcome	ERO	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
All schools					
Reading comprehension					
Average standard score	90.1	89.3	0.8 *	0.08 *	0.042
Corresponding grade equivalent	*6.1*	*6.0*			
Corresponding percentile	*25*	*23*			
Reading vocabulary					
Average standard score	93.5	93.5	0.0	0.00	0.986
Corresponding grade equivalent	*7.8*	*7.8*			
Corresponding percentile	*32*	*32*			
Sample size	1,264	907			
Reading Apprenticeship Academic Literacy schools					
Reading comprehension					
Average standard score	90.2	88.9	1.4 *	0.14 *	0.015
Corresponding grade equivalent	*6.1*	*5.9*			
Corresponding percentile	*25*	*23*			
Reading vocabulary					
Average standard score	93.4	93.8	-0.4	-0.04	0.428
Corresponding grade equivalent	*7.7*	*7.8*			
Corresponding percentile	*32*	*33*			
Sample size	645	470			
Xtreme Reading schools					
Reading comprehension					
Average standard score	90.0	89.7	0.2	0.02	0.672
Corresponding grade equivalent	*6.1*	*6.0*			
Corresponding percentile	*25*	*24*			
Reading vocabulary					
Average standard score	93.5	93.1	0.4	0.04	0.468
Corresponding grade equivalent	*7.8*	*7.7*			
Corresponding percentile	*32*	*31*			
Sample size	619	437			

(continued)

Table ES.1 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study follow-up GRADE assessment.

NOTES: The follow-up GRADE assessment was administered in the spring of 2007 near the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form B). No statistical tests or arithmetic operations were performed on these reference points.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (reading comprehension = 10.035; reading vocabulary = 9.827).

A two-tailed t-test was applied to the impact estimate. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

growth of 4.7 standard score points for the non-ERO group provides the best indication of what the ERO group would have achieved during their ninth-grade year had they not had the opportunity to attend the ERO classes.

The top section of the bar shows the estimated impact of the ERO programs on reading comprehension test scores. At the end of the ninth-grade year, the ERO group achieved an average standard score of 90.1, which corresponds to a grade equivalent of 6.1 and an average reading level at the 25th percentile for ninth-grade students nationally. This means that the ERO group experienced a growth of 5.5 points in their reading comprehension skills over the course of ninth grade, which is 0.8 point higher than the growth achieved by the non-ERO group. Thus, the impact of the ERO programs (0.8 standard score point) represents a 17 percent improvement over and above the growth that the ERO group would have experienced if they had not had the opportunity to attend the ERO classes (4.7 points).[11]

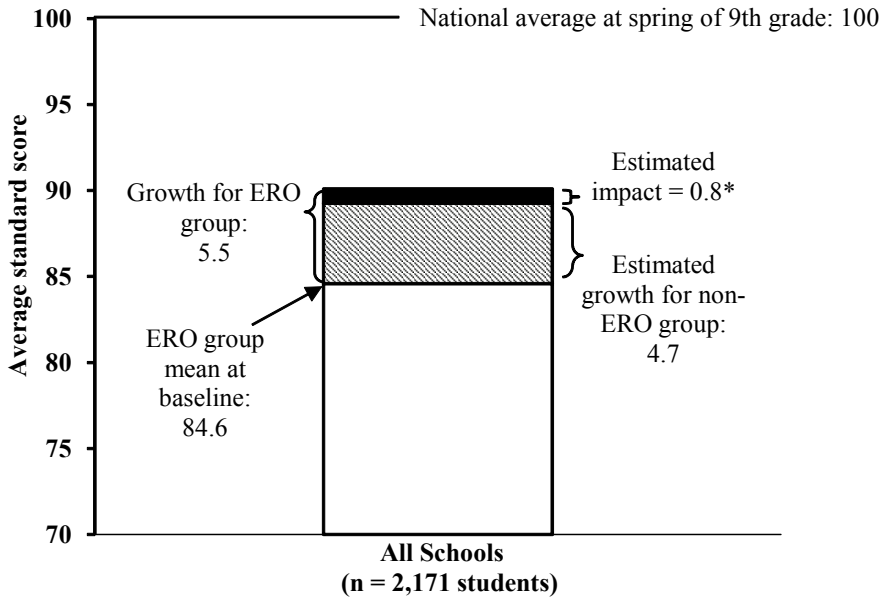
The solid line at the top of Figure ES.1 shows the national average (100 standard score points) for students at the end of ninth grade, in the spring. Students scoring at this level are considered to be reading at grade level. Thus, the ERO group's reading comprehension scores

[11]The value of 17 percent was calculated by dividing the impact (0.8 standard score point) by the average improvement of the non-ERO group (4.7 standard score points).

The Enhanced Reading Opportunities Study

Figure ES.1

Impacts on Reading Comprehension,
Cohort 2 Follow-Up Respondent Sample



SOURCES: MDRC calculations from the Enhanced Reading Opportunities Study baseline and follow-up GRADE assessments.

NOTES: The baseline GRADE assessment was administered in the fall of 2006 at the start of students' ninth-grade year and prior to their random assignment to the ERO and non-ERO groups. The follow-up GRADE assessment was administered in the spring of 2007 near the end of students' ninth-grade year.

The ERO group growth at follow-up is calculated as the difference between the unadjusted ERO group mean at baseline and the unadjusted ERO group mean at follow-up. The impact was estimated using ordinary least squares and adjusted to account for the blocking of random assignment by school and to control for random differences between the ERO and non-ERO groups in baseline reading comprehension test scores and age at random assignment. The expected ERO group growth at follow-up is the difference between the actual ERO group growth and the impact.

A two-tailed t-test was applied to the impact estimate. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

The national average for standard score values is 100, and its standard deviation is 15.

Rounding may cause slight discrepancies in calculating sums and differences.

still lagged nearly 10 points below the national average. In fact, 77 percent of students who participated in the ERO classes scored two or more years below grade level at the end of their ninth-grade year,[12] which means that they would still be eligible for the ERO programs were these programs again made available to them.[13]

- **The RAAL program increased students' reading comprehension test scores by a statistically significant amount (0.14 standard deviation). Although not statistically significant, an impact of 0.2 standard score point on reading comprehension (0.02 standard deviation) was produced by the Xtreme Reading program. The difference in impacts between the two programs is not statistically significant, and thus it cannot be concluded that RAAL had a different effect than Xtreme Reading. Nor is there a statistically significant difference between each program's impact in the second year of implementation and its impact in the first year of implementation.**

The ERO student follow-up survey was administered to students at the same time as the follow-up GRADE assessment and includes additional information on students' reading behaviors and attitudes. Responses to the follow-up survey were used to derive measures for three reading behaviors that are intended to be affected by the ERO programs: the number of times during the prior month that a student read different types of text in school or for homework, the number of times during the prior month that a student read different types of text outside of school, and students' reported use of the reading strategies and techniques that the ERO programs try to teach. The overall impact of the programs on students' reading behaviors is not statistically significant.[14]

The Relationship Between Impacts and Second-Year Implementation

This report also includes an exploratory analysis that investigates the relationship between school-level impacts and various aspects of implementation in the second year of the

[12]Forty percent of ninth-graders nationally would be expected to score two years or more below grade level on the GRADE administered in the spring of ninth grade.

[13]Furthermore, 87 percent of the students in the ERO group had reading comprehension scores that were below grade level at the end of ninth grade.

[14]The analysis also examines the extent to which impacts on reading comprehension test scores vary across schools. The impact estimates for each school range from a negative impact of 3.7 standard score points to a positive impact of 6.2 standard score points. However, the variation in observed school-level impacts is not statistically significant, indicating that the observed school-to-school variation in impacts may be due to estimation error and may not truly vary across schools.

study. Specifically, this analysis examines whether there are differences in impacts between subgroups of schools defined by teachers' experience with the ERO program (that is, schools whose ERO teacher taught two full years of the program versus schools whose ERO teacher did not teach two full years of the program); overall implementation fidelity during the spring site visit (that is, very well-aligned, well-aligned, moderately aligned, or poorly aligned implementation); and the number of weeks between the start of the school year and ERO program startup (schools that started operating their ERO program within two weeks versus those whose program startup was delayed by two weeks or more). The exploratory analysis also examines whether there are differences in impacts between schools whose implementation of the programs was particularly exemplary (that is, schools that started operating their programs within two weeks *and* whose implementation was very well aligned to the program models) and schools that did not meet these two criteria.[15] *Based on these exploratory analyses, one cannot conclude that the programs were more effective in schools with more experienced ERO teachers, with implementation better aligned with the program models, or with early program startup.* That is, one cannot infer with certainty that these particular implementation characteristics are related to program impacts because the difference in impacts between the groups of schools within each of the three measured categories of implementation — teacher experience teaching the ERO classes, the alignment of the programs as implemented to the program models, and the efficiency of program startup — is not statistically significant. Impacts for the groups of schools with the most promising implementation characterizations are positive and statistically significant (that is, for the 25 schools whose ERO teacher returned in the second year, having taught the entire first year of the program; the 13 schools where the ERO programs were rated as very well aligned to the program models; and the 23 schools where the ERO programs began within the first two weeks of school).[16] Impacts for the related groups of schools with less promising implementation characterizations are smaller and not statistically significant (that is, for the 9 schools whose teachers taught ERO for less than two full years, the 21 schools where there was weaker implementation fidelity, and the 11 schools with program startup that took longer than two weeks). The difference in impacts between the groups of schools within each of the three categories of implementation is not statistically significant.

[15]It is important to note that these analyses are exploratory and are not able to establish causal links between these aspects of implementation and variation in program impacts across sites, because other school characteristics and implementation factors may confound the association between school-level impacts and the implementation factors included in the exploratory analysis.

[16]The impacts on reading comprehension test scores for each of these three groups of schools are as follows: in the 25 schools whose ERO teacher had returned having taught all of the first year of the program, the effect size is 0.09 standard deviation (p-value = 0.050); in the 13 schools where implementation was rated as very well aligned to the program models, the effect size is 0.13 standard deviation (p-value = 0.047); and in the 23 schools where the programs began within the first two weeks of school, the effect size is 0.10 standard deviation (p-value = 0.048).

Next Steps for the ERO Study

The ultimate goal of the two ERO programs is to improve students' academic performance during high school and to keep them on course toward graduation. With this in mind, the final report from the evaluation — scheduled for 2009 — will examine the impact of the programs on the achievement and attainment outcomes of both cohorts of students as they progress through high school. The outcomes examined in the report will include students' performance in core academic classes, their performance on the high-stakes tests required by their states, their grade-to-grade promotion rates, and whether they are on track to graduate from high school.

Chapter 1

Introduction

This is the second of three reports from the Enhanced Reading Opportunities (ERO) study — a demonstration and rigorous evaluation of supplemental literacy programs targeted to ninth-grade students with limited literacy skills.[1] It focuses on the second year of implementation of the two supplementary literacy programs being tested and the impact they had on the reading skills of a second cohort of ninth-grade students who entered the study sample during the 2006-2007 school year.

The ERO demonstration involves 34 high schools from 10 school districts that are implementing one of two supplemental literacy programs: Reading Apprenticeship Academic Literacy (RAAL), designed by WestEd, or Xtreme Reading, designed by the University of Kansas Center for Research on Learning. The programs are supplemental in that they consist of a year-long course that replaces a ninth-grade elective class rather than a core academic class. They aim to help striving adolescent readers develop the strategies and routines used by proficient readers and to motivate them to read more and to apply these strategies to a wide range of texts.

The evaluation is assessing the impact of the two supplemental literacy programs on students' reading comprehension skills and on their general performance in high school, including achievement on standardized tests, course completion, and progress toward graduation. MDRC — a nonprofit, nonpartisan social policy research organization — is conducting the evaluation in partnership with the American Institutes for Research (AIR) and Survey Research Management (SRM).

The study's first report provided findings from the initial year of program implementation and examined the effects of the programs on reading comprehension skills for the first of two cohorts of ninth-grade students who are participating in the study.[2] The key findings discussed in that report include the following:

- **On average, across the 34 participating high schools, the supplemental literacy programs improved student reading comprehension test scores by 0.09 standard deviation.** This impact estimate is statistically significant (p-value = 0.019). Seventy-six percent of the students who enrolled in the

[1]The ERO study is known more formally as "An Evaluation of the Impact of Supplemental Literacy Interventions in Freshman Academies."

[2]See Kemple et al. (2008).

ERO classes were still reading at two or more years below grade level at the end of ninth grade.

- **Although they are not statistically significant, the magnitudes of the impact estimates for each literacy intervention were the same as those for the full study sample.**

- **Impacts on reading comprehension were larger for the 15 schools where (1) the ERO programs began within six weeks of the start of the school year and (2) implementation was classified as "moderately" or "well aligned" with the program model (0.17 standard deviation; p-value = 0.002), compared with impacts for the 19 schools where at least one of these conditions was not met (0.01 standard deviation; p-value = 0.811).** The difference in impacts on reading comprehension between these two groups of schools is statistically significant (0.16 standard deviation; p-value = 0.035). It is important to note, however, that these two factors did not necessarily cause the differences in impacts and that other factors may be also associated with differences in estimated impacts across schools.

The findings presented in the first report reflect implementation challenges that arose from the delayed start-up of the programs in the study schools and from the need for teachers to learn new curricula and instructional strategies. In anticipation of these challenges, the U.S. Department of Education (ED) extended the demonstration and evaluation to include a second cohort of ninth-grade students who would be exposed to the programs during their second year of operation. In the second year of the study, most of the schools did not experience the start-up delay that they encountered in the first year. Also, 27 of the 34 ERO teachers from the first year of the study returned for the second year; 25 of them had taught the entire first year of the program, while two of them had taught part of the year, having replaced an ERO teacher midway through the first year of implementation. Thus, in most of the participating schools, findings for the second cohort of students reflect their exposure to a full year of program operation and to teachers in their second year of implementing the programs.

The remainder of this chapter provides background on the ERO demonstration and on the research design being used to assess the impact of the two supplemental literacy programs selected for the project.

Overview of the ERO Study

The ERO study is both a demonstration of two supplemental literacy interventions across a range of contexts and a rigorous evaluation of the interventions' impact on students'

reading comprehension skills and their academic performance as they move through high school. The U.S. Department of Education's Office of Elementary and Secondary Education (OESE) is providing direct support for implementation to the participating schools and districts, while its Institute of Education Sciences (IES) is overseeing the design and execution of the evaluation effort. Following is a brief overview of the demonstration and evaluation components of the ERO study.

A Demonstration of Supplemental Literacy Interventions

The ERO study tracks the implementation of two established supplemental literacy interventions that were developed for high school students whose reading skills are two or more years below grade level as they enter high school. Both programs — Reading Apprenticeship Academic Literacy (RAAL), designed by WestEd, and Xtreme Reading, designed by the University of Kansas Center for Research on Learning — were selected for the study from a pool of 17 applicants by a national panel of experts on adolescent literacy.[3] Each program focuses instruction in the following areas: (1) student motivation and engagement; (2) reading fluency, or the ability to read quickly, accurately, and with appropriate expression; (3) vocabulary, or word knowledge; (4) comprehension, or making meaning from text; (5) phonics and phonemic awareness (for students who could still benefit from instruction in these areas); and (6) writing. The overarching goals of both programs are to help ninth-grade students adopt the strategies and routines used by proficient readers, improve their comprehension skills, and be motivated to read more and to enjoy reading.

Each intervention was part of a larger and more comprehensive high school reform initiative. For the purposes of the ERO study, the programs were modified somewhat and adapted for implementation as a supplemental class that would replace another elective class for ninth-grade students. For the purposes of this demonstration, the programs' developers tailored their professional development and coaching strategies to meet the special needs of high school teachers who did not have reading instruction credentials. Both RAAL and Xtreme Reading are supplemental in that they consist of a yearlong course that replaces a ninth-grade elective class, rather than a core academic class, and in that they are offered in addition to students' regular English language arts classes. Each program is a full-year course and is scheduled for a minimum of 225 minutes of instruction per week. They are both designed to accommodate class sizes of 12 to 15 students.

[3]For an overview of research related to RAAL, see Schoenbach, Greenleaf, Cziko, and Hurwitz (1999). For an overview of research related to Xtreme Reading and the Strategic Instruction Model, see Schumaker and Deshler (2003, 2004).

While the two programs share core goals and many instructional strategies, they differ primarily in their approach to implementation. Implementation of RAAL is guided by the concept of "flexible fidelity" — that is, while the program includes a detailed curriculum, the teachers are trained to adapt their lessons to meet the needs of their students and to supplement program materials with readings that are motivating to their classes. Teachers have flexibility in how they include various aspects of the RAAL curriculum in their day-to-day teaching activities, but they have been trained to do so such that they maintain the overarching spirit, themes, and goals of the program in their instruction.

Implementation of Xtreme Reading is guided by the philosophy that the presentation of instructional material — particularly the order and timing with which the lessons are presented — is of critical import to students' understanding of the strategies and skills being taught. As such, teachers are trained to deliver course content and materials in a precise, organized, and systematic fashion designed by the developers. Xtreme Reading teachers follow a prescribed implementation plan, following specific day-by-day lesson plans in which activities have allotted segments of time within each class period. Teachers also use responsive instructional practices to adapt and adjust to student needs that arise as they move through the highly structured curriculum.

The supplemental literacy programs were implemented in 34 high schools from 10 school districts across the country during the 2005-2006 and 2006-2007 school years. The districts were selected through a special grant competition organized by the U.S. Department of Education's Office of Vocational and Adult Education (OVAE).[4] Experienced, full-time English/language arts or social studies teachers volunteered to teach the programs for a period of two years. It should be noted that the participating sites were not selected to be representative of all districts and schools across the country. As a result, findings from the ERO study cannot be generalized statistically to the full population of districts and high schools or to urban districts and schools. At the same time, the participating sites reflect much of the diversity of midsize and large urban school districts that serve low-income and disadvantaged populations of students. Thus, the findings will be applicable and relevant to districts and high schools that are struggling to meet the needs of ninth-graders who lack the literacy skills required for academic success.

[4]For a complete application package for the special competition, see U.S. Department of Education (2005). The special grant competition was part of OVAE's Smaller Learning Communities initiative and was designed to provide extra funding to qualifying districts for the implementation of the supplemental literacy programs and participation in the ERO evaluation. The grants also included funds for general support of the Small Learning Communities initiatives under way in the districts. In 2006, responsibility for the Smaller Learning Communities initiative and for the special ERO grants was moved from OVAE to OESE.

A Rigorous Impact Evaluation

The ERO evaluation will extend over a five-year period and will address the following questions:

- What are the short-term impacts of these two supplemental literacy interventions, together and separately, on ninth-grade students' reading skills and behaviors?
- For which subgroups of students are supplemental literacy interventions most or least effective?
- What factors promote or impede successful implementation of the supplemental literacy interventions? In what ways are implementation fidelity and quality associated with program impacts (or lack of impacts) on reading achievement and other outcomes?
- What are the longer-term impacts on other academic outcomes, such as achievement on high-stakes standards-based assessments, performance in academic courses, and progress toward graduation? What is the nature of the relationship between the impacts on reading skills and the impacts on these other outcomes?

Like the first report, the current report provides an assessment of the first three of these questions as reflected in the second year of implementation. The final report will address the questions about longer-term impacts.

The ERO evaluation utilizes a two-level random assignment research design. First, within each district, eligible high schools were randomly assigned to use one of the two supplemental literacy programs. This feature of the design allows a direct comparison of the effectiveness of the two programs and avoids confounding the effect of purposeful or self-selection of schools to use the two programs with a true difference in the programs' impact on student achievement.

In the second stage of the design, eligible and appropriate students within each of the participating high schools were randomly assigned either to enroll in a supplemental literacy class (the "ERO group") or to remain in one of the regular elective classes available to ninth-grade students (the "non-ERO group"). Each high school was asked to identify at least 100 ninth-grade students who were reading at least two years below grade level. Approximately 55 percent of these students were randomly assigned to enroll in the ERO class, and the remaining students make up the study's control group and enrolled in or continued in a regularly scheduled elective class. This feature of the design is possible because there were more eligible and appro-

priate students in each high school than the 50 to 60 students that the literacy programs are able to serve. Students in both groups take the regular English/language arts classes offered by their schools as well as other core academic and elective classes required of or offered to ninth-graders. The study includes two cohorts of ninth-grade students: one cohort that was enrolled in the study at the beginning of the 2005-2006 school year and one cohort that was enrolled in the study starting in the 2006-2007 school year. The ERO evaluation taps a variety of data sources to measure students' reading achievement and school performance and to assess the fidelity of program implementation.

It is important to note that the ERO study is an evaluation of a class of reading interventions, as represented by Xtreme Reading and RAAL, as well as an evaluation of each of these two programs separately. The purpose of the study is not to test the differential impact of these two interventions; while Xtreme Reading and RAAL do differ in some respects, they are both full-year supplemental literacy courses targeted at struggling adolescent readers that share many common principles, and hence there was no prior expectation that they would produce substantially different impacts. As noted above, the design of the study is such that programs are randomized to schools; however, the purpose of this randomization was primarily to ensure that each program developer was assigned a fair draw of schools in which to implement its program, rather than to test for a differential impact between the two interventions. By this token, the statistical model chosen for the impact analysis does not utilize the school-level randomization feature of the research design;[5] nor is the sample size large enough to detect policy-relevant differences in impacts across the two programs.[6] Because Xtreme Reading and RAAL represent the same type of reading intervention, this study was designed to test their joint or overall impact. Statistical tests were used to confirm that the difference in impacts between the two programs is not statistically significant and, hence, that it is indeed appropriate to pool together the two program-specific impact estimates; these statistical tests are not appropriate for making inferences about the true difference in impacts between the two interventions.

Overview of This Report

The chapters in this report focus on the study design and the implementation and impact findings as they pertain to the second year of implementation and the second cohort of students

[5]The analysis is based on a school fixed-effects model, which means that the findings in this report represent the estimated impact of the ERO programs *given the assignment of these two programs to the schools* in the study. Had the purpose of the study been to test for a differential impact, the analysis would have had to allow for variability in the assignment of programs to schools (in which case, a school random-effects model would have been used). Chapter 2 and Appendix E provide greater detail on the impact model used in the analysis.

[6]Statistical power is discussed in Chapter 2.

in the study sample. While there are several references to relevant information provided in the first report, each chapter concludes with a brief comparison between design issues or findings from the second year of implementation and those from the first year. Chapter 2 describes the sample of schools and the second cohort of students who are participating in the study. Chapter 3 presents a description of the two supplemental literacy programs and their implementation during the second year of the study. Chapter 4 examines student enrollment and attendance in the ERO classes and looks at the rate at which students in the study's non-ERO sample participated in supplemental literacy services both in and outside school. Chapter 5 reports on the impacts of the literacy interventions for the second cohort of students in the study sample — pooled across both interventions and then for each intervention separately.

This report provides information about the operation and impact of the supplementary literacy interventions under conditions that reflect experiences that the program developers, the districts, the schools, and the teachers gained from the first year of implementation. The developers incorporated lessons from the first year into their follow-up training and coaching activities, while the districts and schools were able to begin operating the programs and supply the required materials earlier in the 2006-2007 school year. Twenty-seven of the 34 teachers from the first year returned to teach the ERO classes in the second year (25 of them having taught the entire first year and two of them having taught part of the year), and the seven new teachers received additional training and coaching to help them implement the program. Based on these factors and others discussed in the report, the second year of program implementation appeared to provide more conducive conditions for successful implementation of the ERO programs than the first year and, thus, a stronger test of the ERO programs' effectiveness at improving the reading comprehension skills of ninth-grade students. Moreover, the story of the ERO programs' potential effectiveness or lack of effectiveness is still unfolding. The third and final report will focus on the programs' longer-term impacts on students' academic achievement and attainment outcomes as they progress through high school, including their performance on high-stakes state tests and their progress toward graduation.

Chapter 2

Study Sample and Design

The Enhanced Reading Opportunities (ERO) study is based on a research design that randomizes both schools and students to supplemental literacy interventions. In the first stage of the design, 34 high schools from 10 school districts were randomly assigned to implement one of two supplemental literacy programs, Reading Apprenticeship Academic Literacy (RAAL) or Xtreme Reading. In the second stage of the design, eligible and consenting ninth-grade students in each high school were randomly assigned either to enroll in an ERO class (the "ERO group") or to remain in one of the regular elective classes available to ninth-grade students (the "non-ERO group"). The literacy programs were implemented by the participating high schools during two academic years (2005-2006 and 2006-2007), resulting in two cohorts of study participants.

As described in the first ERO report, two factors compromised the timely identification and recruitment of eligible students in the first year of the study (2005-2006).[7] First, it was not possible to start the student recruitment process until the school year had already begun, because support grants from the U.S. Department of Education (ED) were not awarded until the summer before the first year of implementation. Hence, the student study sample was not identified until several weeks into the school year, and students selected for the ERO classes were forced to withdraw from an elective course that they had already begun to attend. Second, the study team encountered difficulties in recruiting the desired number of students from the intended target population. The two supplemental literacy interventions being tested are primarily designed for ninth-grade students whose reading skills are two to four years below grade level; however, several of the schools had an insufficient number of students reading in this range. In addition, all of the schools faced challenges in getting eligible students to return signed consent forms. In response to these difficulties, program eligibility was expanded to include students reading one to five years below grade level, and the study sample for the impact analysis was expanded to include students reading two to five years below grade level. In the end, all participating high schools were able to recruit the prescribed number of students for the study sample. However, due to the delayed start of the student recruitment process and the early difficulties in identifying and recruiting a sufficient number of students, this process was not completed until an average of six weeks into the school year.

The present chapter describes the student recruitment and randomization process in the second year of the ERO study (2006-2007), which was modified in response to some of the

[7]Kemple et al. (2008).

challenges encountered in the first year. This chapter also provides descriptive information on the second cohort of students to participate in the study — the group on which the impact analyses in this report focus — and describes various other issues relevant to the study's research design: a description of the schools involved in the study, sources of data, the measures created from these data, student response rates during follow-up data collection, and the analytic methods used to assess program impacts. The chapter concludes with a discussion of the key differences between the study samples in the first and second years of the study.

The following key points pertaining to the second-cohort study sample are discussed:

- The second cohort of students to participate in the study includes 2,679 students with baseline reading test scores that fell between two and five years below grade level. Fifty-seven percent of these students were randomly assigned to the ERO group and scheduled into the ERO classes, while the non-ERO control group enrolled in a regular elective class.

- Approximately 81 percent of the students in the second-cohort study sample (a total of 2,171 students) completed the follow-up reading assessment and survey. Among respondents, overall differences in background characteristics between the ERO and non-ERO groups are not statistically significant.

The chapter also describes the following key differences between the first and second years of the study:

- Given the difficulty of recruiting students in the first year of the study, student recruitment in the second year of the study was begun much earlier, during the spring of students' eighth-grade year. This offered a longer timeline for testing and recruiting students, and it increased the likelihood that the ERO classes would start on the first day of school.

- The most notable difference between the first and second cohort of study participants is with respect to their reading achievement levels prior to random assignment. In particular, students in the second cohort had lower reading comprehension levels at baseline than students in the first cohort. That said, this difference is consistent with the fact that most students in the second cohort were tested in the eighth grade, while all students in the first cohort were tested in the ninth grade.

School Sample

The school districts participating in this study were selected through a special grant competition run by the Office of Vocational and Adult Education (OVAE) within ED.[8] As an extension of the Smaller Learning Communities (SLCs) grant program, this competition sought to provide funding for the implementation of two supplemental ninth-grade literacy programs in selected high schools and to sustain and enhance existing SLCs in these high schools.

In June 2005, ED selected 10 grantee school districts encompassing 34 high schools from a pool of 33 applicant districts.[9] The 10 grantee districts encompass 65 high schools, with the smallest district having four high schools and the largest having 22 high schools. Seven of the grantee districts included four of their high schools in the study, and the remaining three districts included two high schools. Grantee districts will receive approximately $1.25 million over five years for each participating high school. From their SLC grants, districts were required to set aside $250,000 per high school over the first two years of their grant period to cover the costs of implementing the supplemental reading programs, including costs associated with teachers' salaries and benefits, teacher-training activities, coaching and materials to be provided by the program developers, classroom computers, and other resources.

Random Assignment of Schools

Following the selection of grantee districts to participate in the ERO study, the study team randomly assigned the participating schools to implement one of the two literacy programs. Within each district, half the participating schools were randomly assigned to use the RAAL program, and half were randomly assigned to use the Xtreme Reading program.

The random assignment of schools to programs makes it possible to draw valid inferences about the differential impact of the two programs. By randomly assigning schools to one of the two supplemental literacy interventions, the study ensured that the two interventions were implemented in groups of schools that were similar on average. As a result, any difference in impacts that may emerge between the two groups of schools can be attributed to differences in their effectiveness. Had districts or developers been allowed to choose one of the two interventions, the two literacy programs could potentially have been implemented in two very different groups of schools. This is because the decisions of districts and developers could have been based on any number of characteristics that might have made one school a better candidate than

[8]U.S. Department of Education (2005). The implementation was initially funded by the Office of Vocational and Adult Education (OVAE), but this role was later transferred to the Office of Elementary and Secondary Education (OESE).

[9]The number of applicants for the special SLC Grant Competition was reported to the study team by OVAE staff.

another for a successful implementation of the program. For example, developers may have favored schools with a higher level of readiness for their program or schools with fewer existing reading supports for their students. Similarly, schools may have selected the literacy intervention that they believed would be more appropriate or more effective in their school. While these may be sound choices from a practitioner's or a developer's perspective, the resulting differences in school context cannot be measured and would present a threat to the inference that one of the programs is more effective than the other. The randomization of schools to interventions ensures that the difference in impacts between the two groups of schools is due to a difference in the effectiveness of the programs and not to a difference in the characteristics of the schools in which the programs are operating.

Characteristics of Schools Selected for the ERO Project

Table 2.1 presents the characteristics of the 34 high schools participating in the ERO study in the year the OVAE grant was awarded (2004-2005).[10] Overall, ERO programs were implemented in schools located predominantly in large and midsize cities, with some of the schools in each of these categories being listed as "urban fringe." As specified by the OVAE grant requirements, all schools enrolled more than 1,000 students in grades 9 through 12, averaging 1,685 students per school. The schools enrolled an average of 570 ninth-grade students, ranging from 320 to 939 ninth-grade students per school. Table 2.1 shows the average "promoting power" for the participating schools, which can serve as a proxy for the likely longitudinal graduation rate.[11] It indicates that the twelfth-grade class is 59 percent of the size of the ninth-grade class three years earlier, suggesting that roughly 41 percent of students have left the schools between the ninth and twelfth grades. The table also shows that 38 percent of the students in the participating schools were eligible for Title I services and that 47 percent of the students were approved for free or reduced-price lunch.

Overall, Table 2.1 indicates that there is a high degree of similarity between the schools randomly assigned to use RAAL and the schools assigned to use Xtreme Reading. RAAL schools have slightly lower promoting power and somewhat higher percentages of students eligible for Title I services, whereas the Xtreme Reading schools have slightly higher percentages of students eligible for free or reduced-priced lunch. However, none of the differences between

[10]The characteristics of these schools were also examined for the 2005-2006 school year. It was found that the schools had changed very little in 2005-2006; in addition, there was still a high degree of similarity between schools implementing RAAL and Xtreme Reading.

[11]Balfanz and Legters (2004) developed this measure of "promoting power" to approximate a school's graduation rate. It is calculated as the ratio of the number of twelfth-grade students in a given school year to the number of ninth-grade students from three years prior.

The Enhanced Reading Opportunities Study

Table 2.1

Characteristics of ERO Schools and Average Schools in the United States (2004-2005)

Characteristic	All ERO Schools	Reading Apprenticeship Schools	Xtreme Reading Schools	Average U.S. Schools[a]
Average number of students	1,685	1,687	1,683	1,866
Average number of students in grade 9	570	566	574	556
Average number of students in grade 10	432	436	429	478
Average number of students in grade 11	358	359	358	424
Average number of students in grade 12	317	312	322	382
Average promoting power[b] (%)	59.1	56.7	61.6	75.4
Students eligible for free or reduced-price lunch (%)	46.9	44.5	49.2	30.0
Race/ethnicity (%)				
Hispanic	25.1	24.6	25.6	19.3
Black	41.1	41.9	40.4	19.7
White	31.2	31.0	31.5	53.5
Other	2.6	2.6	2.6	7.0
Eligible for Title 1 (%)	38.2	41.2	35.3	26.0
Locale (%)				
Large city[c]	52.9	52.9	52.9	61.2
Midsize city[d]	47.1	47.1	47.1	38.8
Sample size	34	17	17	3,727

SOURCES: U.S. Department of Education, National Center for Education Statistics, Common Core of Data (CCD), "Public Elementary/Secondary School Universe Survey Data," 2004-2005 and 2001-2002.

NOTES: This table provides information on 34 ERO schools from 10 districts in the school year that the OVAE grant was awarded (2004-2005).

Rounding may cause slight discrepancies in calculating sums and differences.

[a]"Average U.S. Schools" includes schools that have more than 1,000 total students, have more than 100 students in each grade during 2004-2005, have at least 125 students in the ninth grade during 2001-2002, are noncharter schools, are located in a large or midsize city or in the urban fringe of a large or midsize city, are defined as "regular" schools by the Common Core of Data, and are operational at the time of the Common Core of Data report.

[b]"Promoting power" is calculated as the ratio of twelfth-grade students in 2004-2005 to ninth-grade students in 2001-2002.

[c]"Large city" is defined as a city having a population greater than or equal to 250,000. Schools in this category also include the urban fringe of a large city.

[d]"Midsize" city is defined as a city having a population less than 250,000 but greater than 50,000. Schools in this category also include the urban fringe of a midsize city.

RAAL and Xtreme Reading schools are statistically significant (none have p-values that are less than or equal to 5 percent).

Table 2.1 also includes information about all high schools across the country that, like those selected for the ERO study, are located in large and midsize cities, served over 1,000 students in grades 9 through 12, and did not select students based on past achievement or performance. This national sample of high schools provides a reference point that helps contextualize and describe the ERO high schools. In comparison with the national sample, the schools selected for the ERO study include a higher proportion of students with characteristics associated with low performance. The ERO schools have lower levels of student promotion, higher percentages of students eligible for free and reduced-price lunch, and higher eligibility for Title 1 funding. Additionally, the populations at ERO schools comprise higher percentages of minority students than the national sample.

Student Sample

Recruitment and Random Assignment of Students

Based on power calculations, the central goal for random assignment was to identify approximately 110 eligible students from each participating high school and to obtain parental consent and background information for these students. Random assignment was to be conducted such that 60 of these students would be selected to enroll in the ERO classes. The ERO programs were designed to accommodate between 12 and 15 students per class, and each high school was required to offer four ERO class sections. The 50 remaining students were to be assigned to remain in one of the regular elective classes available to ninth-grade students. Ideally, schools would have a pool of more than 110 eligible and consenting students, which would allow for some students to be assigned to a nonresearch waiting list and to be admitted to an ERO class if enrollment levels fell below the desired minimum of 12 students due to attrition over the school year.[12]

Due to the difficulties associated with student recruitment in the first year of the study, it was decided that the second cohort of study participants should be recruited using the same "expanded" eligibility criteria as in the first year. Specifically, the primary target population for the study and the impact analysis consists of students whose reading skills are two to five years below grade level. However, eligibility for participation in the ERO programs includes students

[12]Students assigned to the nonresearch waiting list were not included in the analysis, even if they were later scheduled into ERO classes.

reading one to five years below grade level.[13] Hence, while sites were encouraged to recruit as many students as possible in the target population (two to five years below grade level), they were also asked to recruit students whose reading skills were one to two years below grade level. In the event that a site was unable to recruit 110 students reading two to five years below grade level, the study team would fill out the ERO sections and the study sample with students reading one to two years below grade level, which would enable schools to run their ERO classes at capacity.[14,15]

As a result of the first-year recruitment challenges, it was also decided that student recruitment for the second year of the study should begin much earlier, during the spring semester of students' eighth-grade year. Recruiting in the spring allowed for a longer timeline for testing students and obtaining parental consent forms, and it afforded districts additional flexibility insofar as scheduling testing at times that did not conflict with local or state assessments. Also, eighth-grade recruitment made it possible to randomly assign and schedule students for the ERO classes prior to the first day of the 2006-2007 school year, thus increasing the likelihood that ERO classes would start on the first day of school or as close to that as possible.

While the study team felt that eighth-grade recruitment would be a successful approach for recruiting students into the study sample, this did present the potential drawback that student mobility over the summer could result in study participants' not arriving at their expected ERO high school in the fall. Two strategies were used to minimize the loss of study sample members due to mobility over the summer. First, schools were encouraged to allow the study team to delay random assignment for as long as possible, particularly if postponing random assignment would make it possible to verify which students would attend an ERO high school in the fall. The study team was then able to confine random assignment to students whose fall enrollment in an ERO high school had been ascertained.[16] Second, the study team developed a system of oversubscription for the ERO classes. In schools with more than 110 eligible and consenting

[13]It should be noted that English Language Learning (ELL) and special education students who required specific classroom, instructional, or testing accommodations were not eligible for the ERO classes. The ERO programs were not designed to accommodate the special needs of these students nor the potential scheduling conflicts with other services that the students were likely to receive.

[14]For each site, random assignment was conducted separately for students reading two to five years behind grade level (the analytical target population) and students reading one to two years behind grade level.

[15]After comprehensive baseline reading achievement testing, it became apparent that some schools did not have a sufficient number of students scoring between one and five years below grade level to meet the sample-size requirements. In these instances, the eligibility range was expanded to include students reading between 0.1 year and 5 years below grade level.

[16]The timing of random assignment was a function of two opposing considerations: *later* random assignment was preferred by the study team because it would maximize the number of students enrolled in an ERO school in the fall, while *earlier* random assignment was preferred by schools because it allowed them more time to schedule students into ERO sections.

students, the maximum class size was increased from 15 to 18 students, and a greater number of students were randomly assigned to the ERO group. Thus, even with mobility away from ERO high schools over the summer, ERO classes would have a higher likelihood of starting at capacity on the first day of school, which would reduce the need to have waitlist students withdraw from one of their elective courses after the first day of school.

Districts began recruiting students for the second year of the ERO study in the spring semester of the 2005-2006 school year. In order to identify eligible students, districts administered the Group Reading Assessment and Diagnostic Examination (GRADE) to eighth-grade students enrolled in the district's primary feeder middle schools, with the exception of one school district where students were tested in the fall of ninth grade.[17] Districts adopted one of three approaches for testing students and obtaining parental consent for participation in the study:

1. Comprehensive testing using the GRADE, followed by targeted collection of parental consent forms from students scoring in the eligible range

2. Broad collection of parental consent forms, followed by comprehensive GRADE testing

3. Collection of parental consent forms from students likely to be eligible for the ERO programs based on local assessments, followed by targeted GRADE testing

Students who were one to five years below grade level on the GRADE reading comprehension subtests at the time of testing were considered eligible for the ERO classes.[18] Once a signed affirmative consent form and a complete baseline survey were obtained from eligible students, they were entered into MDRC's random assignment database. While the recruitment of eligible students required the assistance of school and district staff members in communicating with parents and students and collecting consent forms, computerized random assignment of students was conducted solely by MDRC staff.

Figure 2.1 illustrates the random assignment of eligible students and charts the construction of the analytic sample from the pool of eligible students. In the second year of the study, the study team identified 3,441 eligible and consenting students from across the 34 participating high schools (on average, 101 students per school). Of these students, 1,946 (57 percent) were randomly selected to enroll in the ERO classes (the ERO group) and 1,495 (43 per-

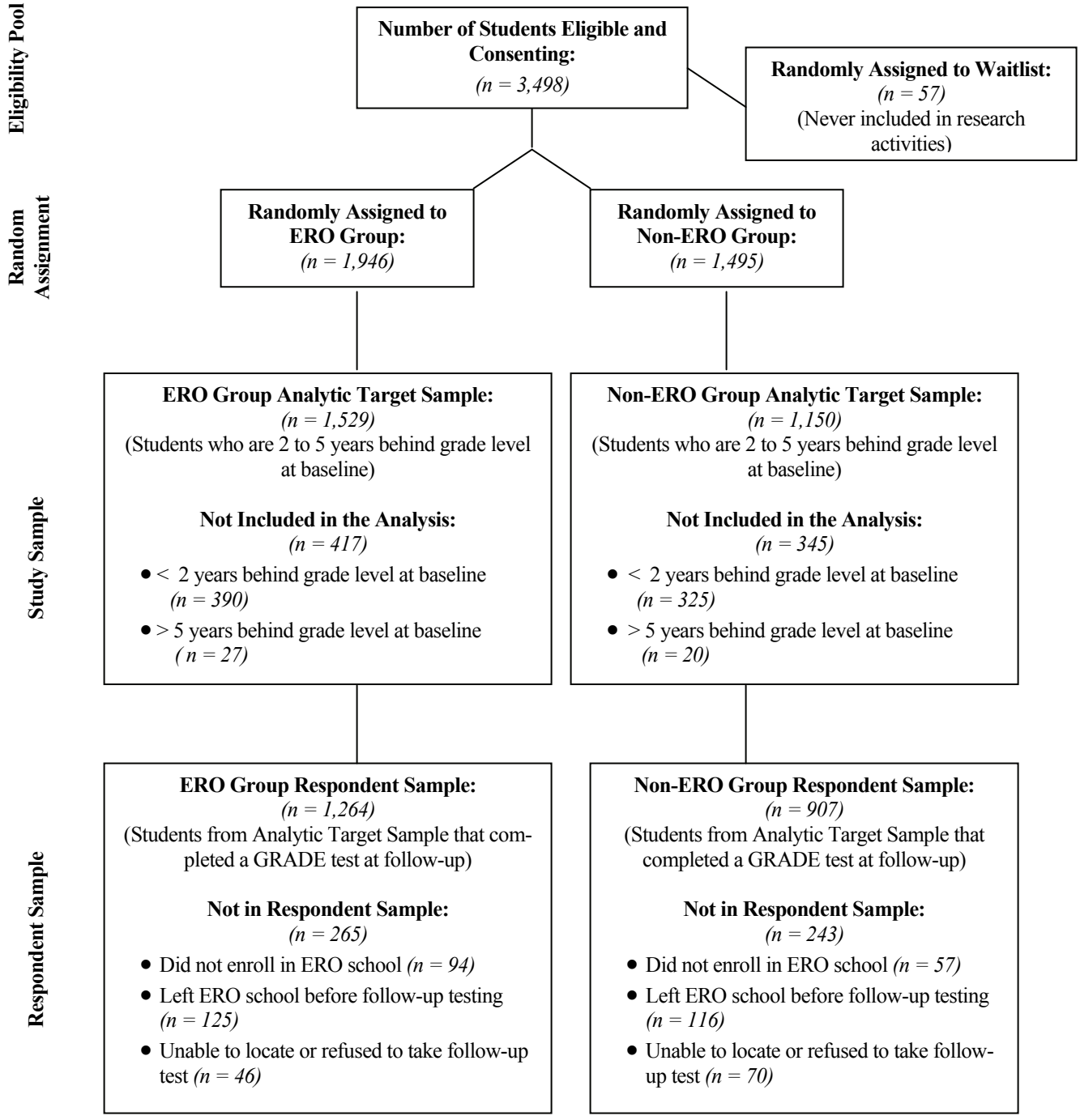
[17]This school district comprises four ERO high schools.

[18]The month in which students were tested varies by district and school, ranging from February 2006 to September 2006. Thus, program eligibility and the definition of the study sample were determined based on the number of years that a student was behind grade level *at the time of testing.* This was calculated as a student's grade level at the time of testing minus the student's grade equivalent score on the GRADE.

The Enhanced Reading Opportunities Study

Figure 2.1

Construction of the Impact Sample from the Eligibility Pool for Cohort 2



cent) were randomly assigned to the control group (the non-ERO group). As described above, eligibility for participation in the ERO programs includes students reading one to five years below grade level, but the analyses in this report focus exclusively on students reading two to five years below grade level at the time of testing. Figure 2.1 shows that there are 2,679 students in this group (78 percent of the entire study sample; on average, 79 students per school), with 1,529 (57 percent) randomly assigned to the ERO group and 1,150 (43 percent) randomly assigned to the non-ERO group.[19] All further references in this report to the "study sample" refer to students who are two to five years below grade level based on their baseline reading comprehension test scores.

Characteristics of the Study Sample

The background characteristics of students in the ERO group and the non-ERO group were compared to determine whether random assignment in the second year of the study resulted in two equivalent groups. (Box 2.1 explains how outcome levels for these two groups are calculated and presented throughout this report.) As illustrated in Table 2.2, there is a high degree of similarity between the two groups' baseline characteristics. On average, students in the second-cohort study sample had a reading comprehension composite score of just under 85 standard score points on the GRADE reading assessment at the time of baseline testing. This average corresponds to the 4.9 grade level (an average of almost four years below grade level) and to the 14th percentile nationally.[20] The study sample is over 78 percent Hispanic or black; about 49 percent of the students come from multilingual homes; and about 29 percent are overage for grade (15 years old or older at the start of ninth grade, suggesting that they were retained in grade at some point in their schooling).[21] A general F-test indicates that, overall, there is no systematic difference in the background characteristics of the ERO and non-ERO groups in the

[19]The impact analysis does not include 762 study participants because their baseline GRADE scores fall outside the analytical target range. Specifically, 578 students (17 percent of the sample) had baseline test scores that were between one and two years below grade level; 47 students were more than five years below grade level (1 percent of the sample); and 137 students were less than one year below grade level at baseline (4 percent of the sample). The latter group of students was allowed to participate in the ERO classes following a further expansion of the eligibility criteria; some schools did not have a sufficient number of students scoring between one and five years below grade level to meet the sample-size requirements, so their eligibility range was expanded to include students reading between 0.1 year and 5 years below grade level.

Impact analyses conducted on a sample that includes these 762 students produced similar estimates to the ones presented in Chapter 5.

[20]The average GRADE reading comprehension standard score for a representative group of students at the end of their ninth-grade year is 100, and the standard deviation is 15.

[21]National Center for Education Statistics (1990).

Box 2.1

Description of the Calculation and Presentation of Outcome Levels

Throughout the report, when a table is presented to report estimated program impacts, the mean outcome levels for the ERO and the non-ERO groups are reported, in order to provide context for interpreting the estimated differences. Program impacts are estimated using a regression model that uses all available observations from both the ERO group and the non-ERO group, and the mean outcome levels are calculated using the same impact regression model.

When calculating the regression-adjusted mean outcome levels for the ERO and non-ERO groups, the adjustment is made using the observed mean covariate values for the ERO group in the impact regression model. In other words, means for *both* groups are "regression-adjusted" using this common set of baseline covariate values: the *ERO group's observed means.*

By adjusting based on the observed mean covariate values for the ERO group, the tables report:

- Observed mean outcome levels for students randomly assigned to the ERO group
- Regression-adjusted mean outcome levels for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment

By presenting the observed mean outcome values for the ERO group, the discussion is based on the actual mean outcomes for the ERO group, which makes it possible to compare these actual values with those for other reference groups or for the same group of students over time. The reported mean outcome level for the non-ERO group also has a straightforward interpretation: it provides an unbiased estimate of how the ERO group students would have performed had they not been assigned to the ERO programs. In other words, it represents the "counterfactual."

In the text and tables of this report, when presenting these outcome levels, the *observed* mean level for the ERO group is referred to as the "ERO group" mean. The mean value for the counterfactual, or the *regression-adjusted* mean for the non-ERO group, is referred to as the "non-ERO group" mean. In addition, note that observed means (adjusted only for random assignment blocks) for both the ERO and the non-ERO group are presented in Appendix Tables E.1 and E.2.

The Enhanced Reading Opportunities Study

Table 2.2

Characteristics of Students in Cohort 2
Full Study Sample

Characteristic	ERO Group	Non-ERO Group	Difference	P-Value for the Difference
Race/ethnicity (%)				
Hispanic	30.3	29.9	0.3	0.779
Black, non-Hispanic	47.8	48.5	-0.7	0.640
White, non-Hispanic	15.5	14.5	1.0	0.429
Other	6.4	7.0	-0.6	0.509
Gender (%)				
Male	49.4	52.9	-3.5	0.076
Female	50.6	47.1	3.5	0.076
Average age (years)	14.8	14.8	0.0	0.735
Overage for grade[a] (%)	28.8	30.3	-1.5	0.361
Language other than English spoken at home (%)	48.2	50.0	-1.8	0.310
Language spoken at home missing (%)	1.8	2.3	-0.5	0.370
Mother's education level (%)				
Did not finish high school	18.4	17.5	0.9	0.521
High school diploma or GED certificate	24.0	25.0	-1.0	0.552
Completed some postsecondary education	30.5	32.5	-2.0	0.268
Don't know	24.3	22.6	1.7	0.300
Missing	2.7	2.4	0.3	0.618
Father's education level (%)				
Did not finish high school	15.3	18.0	-2.7	0.054
High school diploma or GED certificate	24.1	19.3	4.8 *	0.003
Completed some postsecondary education	20.5	20.8	-0.3	0.872
Don't know	36.6	39.0	-2.4	0.196
Missing	3.5	2.9	0.6	0.412
GRADE reading comprehension[b]				
Average standard score	84.4	84.7	-0.3	0.241
Corresponding grade equivalent	*4.8*	*4.9*		
Corresponding percentile	*14*	*14*		
2.0 - 3.0 years below grade level (%)	31.0	32.2	-1.2	0.495
3.1 - 4.0 years below grade level (%)	27.3	28.6	-1.2	0.477
4.1 - 5.0 years below grade level (%)	41.7	39.2	2.5	0.198
Sample size	1,529	1,150		

(continued)

Table 2.2 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities baseline data.

NOTES: Baseline data for students in 30 of 34 high schools were collected in spring 2006 (during students' eighth-grade year); baseline data for students in the remaining four schools were collected in fall 2006 (at the start of students' ninth-grade year).

The estimated differences are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO" Group values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed distribution of the ERO group across random assignment blocks (i.e., schools) as the basis for the adjustment.

A two-tailed t-test was used to test differences between the ERO and non-ERO groups. The statistical significance level is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

[a]A student is defined as overage for grade if he or she turned 15 before the start of ninth grade.

[b]The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form A). No statistical tests or arithmetic operations were performed on these reference points.

study sample.[22] The lack of a systematic difference indicates that random assignment was successful in creating two equivalent research groups at baseline. Though not presented in this report, similar results were found when examining the background characteristics of students in the study samples from the RAAL sites and the Xtreme Reading sites, separately.

Data Sources and Measures

The ERO evaluation utilizes a variety of data sources to measure students' reading achievement and reading behaviors and to assess the fidelity and quality of program implementation. Following is an overview of the current report's data sources and measures, which are the same as those used to evaluate the first year of program implementation.[23]

[22]Because many hypothesis tests are conducted in Table 2.2 (one for each baseline characteristic), there is an increased probability of concluding that a particular baseline difference is statistically significant when, in fact, it is not. (This is called a Type I error, or a "false positive.") In particular, one would expect to see a "false positive" for every 20 hypothesis tests conducted. For this reason, it is important to use an overall F-test to test for a systematic, or *overall,* difference between the characteristics of the ERO and non-ERO groups.

[23]See Kemple et al. (2008).

Group Reading Assessment and Diagnostic Examination (GRADE)

The GRADE assessment was used to measure students' reading achievement prior to random assignment (at "baseline") and then again at the end of their ninth-grade year (at "follow-up").[24] The GRADE is a norm-referenced, research-based reading assessment that can be administered to groups. It is meant to be a diagnostic tool to assess what reading skills individuals have and what skills need to be taught.[25] It is used widely to measure performance and track the growth of an individual student and groups of students from fall to spring and from year to year. The GRADE contains multiple subtests, including two reading comprehension subtests (sentence comprehension and passage comprehension), a listening comprehension subtest, and a vocabulary subtest.

The ERO impact analysis uses two measures of reading achievement based on the GRADE follow-up test:

- **Reading Comprehension.** The primary measure of reading achievement for this study is students' scores on the two GRADE reading comprehension subtests. A central objective of each of the two ERO programs is to provide students with immediate and intensive instruction in the use of strategies and skills that expert readers use to understand written texts. Thus, for the purposes of the ERO evaluation, the GRADE reading comprehension assessment serves as the primary early indicator of the programs' effectiveness.

- **Reading Vocabulary.** A secondary measure of students' reading achievement is their scores on the GRADE vocabulary subtest. Each of the two ERO programs provides some instruction aimed at helping students break down word meanings through advanced decoding skills and strategies for recognizing word structures (root words, prefixes, and suffixes). Thus, the GRADE vocabulary assessment can provide indication of whether these approaches increase the stock of words that students know. However, because the two ERO programs focus primarily on helping students use contextual clues to understand the meaning of words, the vocabulary subtest is seen as a secondary indicator of the programs' effectiveness.

The GRADE reading comprehension and vocabulary performance levels and impacts for the ERO and non-ERO groups are presented in standard score units provided by the Ameri-

[24]Prior to random assignment, the two reading comprehension subtests (Level H, Form A) were administered to all students. These two tests were administered again near the end of students' ninth-grade year (Level H, Form B), in addition to the vocabulary subtest.

[25]See American Guidance Service (2001a, 2001b) for technical information about the GRADE.

can Guidance Service, which publishes the GRADE.[26] Standard scores are a more accurate representation of a student's level of performance than raw scores because they have uniform meaning from one test period to another and from one grade level to another. Standard scores indicate how far a student's performance on the test is from the average for all students at a given grade level, and standard scores take into account the variability of scores among a nationally representative group of students in that grade. Also, standard scores on the GRADE can be compared with standard scores on other tests of reading comprehension and vocabulary.

To help the reader interpret the standard score values, the impact tables also present the national grade equivalent and national percentile that correspond most closely to the average standard score for the ERO and non-ERO groups, respectively. A grade equivalent score is the grade at which a particular raw score or standard score represents the median for the test's norming population. For example, a grade equivalent score of 9.0 refers to a median performance at the beginning of ninth grade, and a 9.9 grade equivalent indicates a median performance at the end of ninth grade.[27]

The reading comprehension and vocabulary test score impact estimates are presented both in standard score units and in effect-size units. Effect sizes provide an indication of the magnitude of the impact estimates relative to the overall variation in test scores for students in the study sample. For the purposes of the impact analysis, effect sizes are calculated as a proportion of the standard deviation of the test scores for students in the non-ERO group at the end of ninth grade.[28] The standard deviation for the non-ERO group reflects the expected variability in test scores that one would find in the absence of the ERO programs. The impact effect size,

[26]Specifically, each student's raw scores on the GRADE subtests and composite scores were converted to standard scores based on national norms for Level H, Grade 9, Spring Testing (American Guidance Service, 2001b, pp. 30-33). Based on these norms, a standard score of 100 on the GRADE reading comprehension or vocabulary test is average for a representative group of students at the end of their ninth-grade year. The standard deviation of the standard score for both tests is 15. A standard score of 85 corresponds, approximately, to the 4.9 grade equivalent.

[27]Note that grade equivalents and percentiles are not equal-interval scales of measurement. Grade equivalents indicate a student's place along a growth continuum, which may not increase at regular intervals. For example, the difference between a vocabulary grade equivalent of 1.0 and 2.0 represents a greater difference in vocabulary knowledge than the difference between a grade equivalent of 8.0 and 9.0. Percentiles indicate the percentage of students in the test's norming group who performed at or below a given student's score. As such, percentiles provide information only about the rank order of students' scores; they do not provide any information about students' actual performance. Because they do not reflect equal intervals between units of measure, neither grade equivalents nor percentiles can be manipulated arithmetically. (See American Guidance Service, 2001a, pp. 55-60.) Thus, readers should exercise caution when interpreting differences in grade equivalents or percentiles between the ERO and non-ERO groups and between the baseline and follow-up tests.

[28]The standard deviation of the reading comprehension standard score for the non-ERO group at follow-up is 10.035. The standard deviation of the vocabulary standard score for the non-ERO group is 9.827.

therefore, provides an indication of how much the ERO programs moved students along this variability in expected performance.

Student Surveys

Students in the study completed a baseline survey and a follow-up survey. The baseline survey was completed by students in the study sample prior to random assignment, and it includes background information on such items as gender, race/ethnicity, age, and parents' education. These items are used to compare the ERO and non-ERO research groups at baseline.

The follow-up survey was administered to students at the same time as the follow-up GRADE assessment, and it includes additional information on students' reading behaviors and attitudes. Responses to the follow-up survey were used to derive measures for three reading behaviors that should be affected by the ERO programs:[29]

- **Amount of School-Related Reading.** The self-reported number of times during the prior month that a student read different types of text in school or for homework: history, science, or math textbooks; literary texts; research or technical reports; newspaper or magazine articles; or workbooks (7 survey items; Cronbach's alpha = .71).[30]

- **Amount of Non-School-Related Reading.** The self-reported number of times during the prior month that a student read different types of text outside school: fictional books; plays; poetry; (auto)biographies; books about science, technology, or history; newspaper or magazine articles; or reference books (7 survey items; Cronbach's alpha = .75).

- **Use of Reflective Reading Strategies.** Students' reported use of the reading skills and techniques that the ERO programs try to teach as they read for their English/language arts class and for one other academic class.[31] Students were asked to rate their use of these strategies on a scale from 1 (strongly disagree) to 4 (strongly agree) (4 survey items; Cronbach's alpha = .77).

[29]A list of the survey items used to create these three measures and a copy of the survey instrument are presented in Appendix A.

[30]Cronbach's coefficient alpha is a statistical measure of the degree to which the individual items used to create the multi-item construct are correlated with each other (Cronbach, 1951).

[31]The follow-up survey asked students to report on reading strategies that they used in social studies, science, and mathematics classes, if they were taking these courses. The measure relies on the social studies class, if the student reported taking social studies. Otherwise, it includes science. If the student was not taking either social studies or science, the measure includes mathematics.

The impact estimates for these three measures of reading behaviors in Chapter 5 are presented both in their original metrics and in effect-size units. Effect sizes provide an indication of the magnitude of the impact estimates relative to the variation in the measures for students in the study sample who were not exposed to the ERO programs. As with the test score outcomes, effect sizes are calculated as a proportion of the standard deviation of the given outcome for students in the non-ERO group.[32] The standard deviation for the non-ERO group reflects the expected variability in the reading behavior that one would find in the absence of the ERO programs. The impact effect size, therefore, provides an indication of how much the ERO programs moved students along this variability in expected reading behavior.

The student follow-up survey was also used to derive four measures of students' participation in supplementary literacy support activities during the school year, both inside and outside school. These measures are described in greater detail in Chapter 4, where they are used to quantify the level of reading support services received by the ERO group relative to the non-ERO group.

Teacher Survey

The study team administered a teacher survey to ERO teachers prior to the first year of program implementation, during the summer training institutes held by the interventions' developers. This survey was also administered during the summer training institutes held prior to the second year of implementation, but only to ERO teachers who were new to the program in the second year. The survey has two parts: Part 1 of the survey asked teachers about their backgrounds, their experiences with professional development activities, their school environments, and their beliefs about literacy instruction. Part 2 of the survey asked teachers about their impressions of the training they attended.

Implementation Data

Classroom Observations

The analysis of ERO program implementation fidelity in the second year of the study is based on field research visits to each of the 34 high schools in both the fall and the spring semesters of the 2006-2007 school year. The primary data collection instrument for the site visits was a set of protocols for classroom observations and interviews with the ERO teachers.[33] The ob-

[32]The standard deviation of the "amount of school-related reading" for the non-ERO group is 38.322. The standard deviation of the "amount of non-school-related reading" for the non-ERO group is 32.976. The standard deviation of the "use of reflective reading strategies" for the non-ERO group is 0.592.

[33]The observation protocols can be found in Appendix D.

servation protocols provided a structured process for trained classroom observers to rate characteristics of the ERO classroom learning environments and the ERO teachers' instructional strategies. Each of these characteristics was selected for assessment because it was aligned with program elements specified by the developers and, by design, was aligned with supplemental literacy program elements that are believed to characterize high-quality interventions for struggling adolescent readers.[34] Chapter 3 provides a more detailed description of the data collection process and a description of the summary measures of implementation fidelity that were developed from the classroom observation data. Appendix D provides further background on the properties of the classroom observation data and the fidelity measures.

Teacher Interviews

During both the fall and the spring field visits, the study team interviewed the ERO teacher using a semi-structured interview protocol that focused on teachers' perceptions of aspects of the intervention, of the coaching and support that they received from the developers, of the ease of implementing the program, and of students' responses to and challenges with the program. During the spring site visit, the study team also interviewed English/language arts teachers and elective teachers in order to explore the extent to which literacy instruction may be taking place in classes other than ERO.

ERO Class Attendance Records

Each of the ERO teachers provided monthly school attendance data for all students in the study sample and ERO class attendance data for those students assigned to an ERO class.

Student Course Schedules

Each school provided the study team with copies of the schedules for all students in the study sample. One purpose of the schedule data is to confirm that ERO students were enrolled in the ERO classes and that non-ERO students were not.[35] These data allow the study team to check for possible contamination — that is, for non-ERO students receiving the ERO program.

Follow-Up Data Collection and Response Rates

The follow-up GRADE assessment and survey were administered to students in the second-cohort study sample late in the 2006-2007 school year. Overall, follow-up data are available for 81 percent of the study sample; response rates are also 81 percent for each of the

[34]Biancarosa and Snow (2004).

[35]See Chapter 4 for discussion of student schedules and enrollment in the ERO classes.

two groups of schools implementing either RAAL or Xtreme Reading. However, Table 2.3 shows that there is a statistically significant difference (p-value is less than or equal to 5 percent) in response rates between the ERO and the non-ERO groups in the full study sample and in the Xtreme Reading sample. In both of these samples, the response rate for students in the ERO group is 83 percent, compared with 79 percent for students in the non-ERO group.[36]

When response rates are less than 100 percent or when there are different response rates for the program and control groups, it is important to investigate two concerns. First, does the respondent sample differ from the full study sample and from the nonrespondent sample? Second, within the respondent sample, are the ERO group and the non-ERO group still equivalent?

The ERO study team conducted a nonresponse analysis by examining differences in background characteristics between respondents and nonrespondents in the second-cohort study sample.[37] While the respondent sample reflects the general characteristics of the full study sample, an overall F-test comparing the respondents and nonrespondents indicates that there are systematic differences between them in terms of their background characteristics. Most notably, nonrespondents are more likely than respondents to have characteristics associated with doing more poorly in school. For example, a higher percentage of nonrespondents are overage for grade (41 percent, compared with 26 percent of respondents), which means that nonrespondents are more likely to have been held back in a previous grade. On average, nonrespondents also had lower reading comprehension test scores at baseline than respondents (84 points for nonrespondents, compared with 85 points for respondents). Both of these differences are statistically significant at the 5 percent level.[38] Such overall differences between respondents and nonrespondents suggest that one should be cautious when generalizing findings from the follow-up respondent sample to the full second-cohort study sample.[39]

As noted at the beginning of this section, there is a statistically significant difference in response rates between the ERO group and the non-ERO group. This raises a concern about whether respondents in the ERO group differ systematically from respondents in the non-ERO group. Table 2.4 shows the background characteristics of all 2,171 students in the second-cohort follow-up respondent sample and provides a comparison between the ERO and non-ERO

[36]See also Appendix Table B.1 in Appendix B.

[37]See Appendix B for the results of the statistical analyses that were conducted to assess differences between respondents and nonrespondents. Results are presented for all the participating high schools together and, separately, for the groups of schools using RAAL and Xtreme Reading.

[38]See Appendix Table B.2 in Appendix B.

[39]See Appendix F for results from supplemental impact analyses that include sampling weights to account for differences between respondents and nonrespondents. These results indicate little difference between the weighted and the unweighted impact estimates.

The Enhanced Reading Opportunities Study

Table 2.3

Response Rates of Students in Cohort 2
Full Study Sample

	ERO Group	Non-ERO Group	Difference	P-Value for the Difference
All schools				
Response rate (%)	82.7	79.1	3.6 *	0.018
Sample size	1,529	1,150		
Reading Apprenticeship schools				
Response rate (%)	82.4	79.4	3.0	0.163
Sample size	783	594		
Xtreme Reading schools				
Response rate (%)	83.0	78.7	4.3 *	0.049
Sample size	746	556		

SOURCES: MDRC calculations from the Enhanced Reading Opportunities baseline data and follow-up GRADE assessment.

NOTES: This table represents the response rates for the follow-up GRADE assessment, which was administered in spring 2007 at the end of students' ninth-grade year. The follow-up student questionnaire was also administered at that time. The difference in response rates between the test and survey is negligible.

The estimated differences are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed distribution of the ERO group across random assignment blocks (i.e., schools) as the basis for the adjustment.

A two-tailed t-test was used to test differences between the ERO and non-ERO groups. The p-value is the probability that the observed difference is the result of chance and does not represent a true difference between groups. The lower the p-value, the less confidence that there is not a difference between the two groups. The statistical significance level is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

The Enhanced Reading Opportunities Study

Table 2.4

Characteristics of Students in Cohort 2 Follow-Up Respondent Sample

Characteristic	ERO Group	Non-ERO Group	Difference	P-Value for the Difference
Race/ethnicity (%)				
Hispanic	31.8	32.0	-0.2	0.912
Black, non-Hispanic	45.8	46.9	-1.0	0.509
White, non-Hispanic	16.1	14.4	1.7	0.204
Other	6.3	6.8	-0.5	0.622
Gender (%)				
Male	49.9	53.5	-3.5	0.103
Female	50.1	46.5	3.5	0.103
Average age (years)	14.7	14.7	0.0	0.749
Overage for grade[a] (%)	25.6	26.8	-1.3	0.490
Language other than English spoken at home (%)	49.1	50.9	-1.8	0.352
Language spoken at home missing (%)	1.3	1.0	0.3	0.519
Mother's education level (%)				
Did not finish high school	17.9	16.9	1.0	0.547
High school diploma or GED certificate	23.8	25.9	-2.0	0.275
Completed some postsecondary education	31.6	33.3	-1.7	0.383
Don't know	24.5	22.7	1.8	0.336
Missing	2.2	1.2	1.0	0.081
Father's education level (%)				
Did not finish high school	14.6	18.7	-4.2 *	0.008
High school diploma or GED certificate	24.2	19.3	4.9 *	0.006
Completed some postsecondary education	21.7	21.7	0.0	0.998
Don't know	36.6	38.3	-1.7	0.415
Missing	2.9	1.9	1.0	0.150
GRADE reading comprehension[b]				
Average standard score	84.6	85.0	-0.4	0.091
Corresponding grade equivalent	*4.9*	*5.0*		
Corresponding percentile	*14*	*15*		
2.0 - 3.0 years below grade level (%)	32.0	33.6	-1.7	0.408
3.1 - 4.0 years below grade level (%)	27.7	29.6	-1.9	0.338
4.1 - 5.0 years below grade level (%)	40.3	36.8	3.5	0.094
Sample size	1,264	907		

(continued)

Table 2.4 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities baseline data.

NOTES: Baseline data for students in 30 of 34 high schools were collected in spring 2006 (during students' eighth-grade year); baseline data for students in the remaining four schools were collected in fall 2006 (at the start of students' ninth-grade year).

The estimated differences are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed distribution of the ERO group across random assignment blocks (i.e., schools) as the basis for the adjustment.

A two-tailed t-test was used to test differences between the ERO and non-ERO groups. The statistical significance level is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

[a]A student is defined as overage for grade if he or she turned 15 before the start of ninth grade.

[b]The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form A). No statistical tests or arithmetic operations were performed on these reference points.

groups. Like Table 2.2 for the overall study sample, Table 2.4 shows a high degree of similarity between the respondents in the ERO and non-ERO groups across the baseline characteristics. A general F-test indicates that, overall, there is no systematic difference between the ERO and non-ERO group respondents.[40] This suggests that one may have a high degree of confidence that differences in outcomes between the two groups reflect impacts of the ERO programs rather than preexisting differences in students' background characteristics.

The characteristics displayed in Table 2.4 indicate that the typical follow-up respondent sample member in the second year of the study was reading well below grade level in the spring of eighth grade and that many respondents have characteristics associated with a risk of doing poorly in school. On average, students had a reading comprehension composite score of about 85 standard score points at baseline testing, corresponding to the 4.9 grade level (an average of four years below grade level) and to the 14th percentile nationally. Also, 78 percent of the students in the follow-up respondent sample are Hispanic or black, and 50 percent reported that a language other than English is spoken in their homes. Tables 2.5 and 2.6 present analogous

[40]See Appendix B for the results of the statistical analyses that were conducted to assess differences between the ERO and non-ERO groups in the respondent sample.

The Enhanced Reading Opportunities Study

Table 2.5

Characteristics of Students in Cohort 2 Follow-Up Respondent Sample, Reading Apprenticeship Schools

Characteristic	ERO Group	Non-ERO Group	Difference	P-Value for the Difference
Race/ethnicity (%)				
Hispanic	31.0	31.6	-0.6	0.780
Black, non-Hispanic	46.7	48.7	-2.1	0.366
White, non-Hispanic	15.2	13.1	2.1	0.256
Other	7.1	6.6	0.5	0.724
Gender (%)				
Male	50.1	52.2	-2.1	0.490
Female	49.9	47.8	2.1	0.490
Average age (years)	14.7	14.7	0.0	0.723
Overage for grade[a] (%)	27.4	27.7	-0.3	0.911
Language other than English spoken at home (%)	46.8	50.2	-3.4	0.206
Language spoken at home missing (%)	1.2	0.9	0.4	0.539
Mother's education level (%)				
Did not finish high school	18.6	17.3	1.4	0.558
High school diploma or GED certificate	25.0	26.2	-1.2	0.638
Completed some postsecondary education	30.5	30.4	0.1	0.962
Don't know	23.7	24.5	-0.7	0.774
Missing	2.2	1.7	0.5	0.552
Father's education level (%)				
Did not finish high school	14.9	16.5	-1.6	0.466
High school diploma or GED certificate	25.7	21.3	4.4	0.082
Completed some postsecondary education	20.9	20.0	1.0	0.689
Don't know	35.3	39.8	-4.5	0.124
Missing	3.1	2.4	0.7	0.515
GRADE reading comprehension[b]				
Average standard score	84.7	85.0	-0.3	0.321
Corresponding grade equivalent	*4.9*	*5.0*		
Corresponding percentile	*14*	*15*		
2.0 - 3.0 years below grade level (%)	33.8	34.1	-0.3	0.926
3.1 - 4.0 years below grade level (%)	26.0	28.3	-2.3	0.401
4.1 - 5.0 years below grade level (%)	40.2	37.6	2.5	0.393
Sample size	645	470		

(continued)

Table 2.5 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities baseline data.

NOTES: Baseline data for students in 30 of 34 high schools were collected in spring 2006 (during students' eighth-grade year); baseline data for students in the remaining four schools were collected in fall 2006 (at the start of students' ninth-grade year).

The estimated differences are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed distribution of the ERO group across random assignment blocks (i.e., schools) as the basis for the adjustment.

A two-tailed t-test was used to test differences between the ERO and non-ERO groups. The statistical significance level is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

[a]A student is defined as overage for grade if he or she turned 15 before the start of ninth grade.

[b]The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form A). No statistical tests or arithmetic operations were performed on these reference points.

results for students in the follow-up respondent samples from the RAAL schools and from the Xtreme Reading schools, respectively.

The similarity between the student characteristics of the follow-up respondent sample and the full study sample — as well as the lack of a systematic difference between the ERO and non-ERO groups in the follow-up respondent sample — indicate that the second-cohort follow-up respondent sample preserves the balance that was achieved with random assignment for the full study sample. This balance was also preserved within each of the groups of schools using the two supplemental literacy programs.

Analytic Methods and Procedures

When examining the effectiveness of the ERO programs in improving students' reading achievement and behaviors, it is important to distinguish between measures of program "outcomes" and measures of program "impacts." *Outcomes* refer to the measures of student performance, behaviors, achievement, and attitudes — in this case, reading achievement and reading behaviors at the end of the ninth-grade year. An *impact* is the effect that the ERO programs have on an outcome. The average outcome levels for students in the ERO group alone provide potentially misleading conclusions. Reading achievement and behaviors are likely to change for students for reasons not related to a special intervention like the ERO programs. In order to determine the net effect, or "value added," of the ERO programs, it is necessary to compare the experiences of a group of students who were exposed to the ERO classes with a similar group of students who also applied but were not selected to enroll. As discussed above in this chapter, the

The Enhanced Reading Opportunities Study

Table 2.6

Characteristics of Students in Cohort 2 Follow-Up Respondent Sample, Xtreme Reading Schools

Characteristic	ERO Group	Non-ERO Group	Difference	P-Value for the Difference
Race/ethnicity (%)				
Hispanic	32.6	32.4	0.3	0.888
Black, non-Hispanic	44.9	44.9	0.0	0.991
White, non-Hispanic	17.1	15.8	1.3	0.505
Other	5.3	7.0	-1.6	0.270
Gender (%)				
Male	49.8	54.8	-5.1	0.103
Female	50.2	45.2	5.1	0.103
Average age (years)	14.7	14.7	0.0	0.928
Overage for grade[a] (%)	23.6	25.9	-2.3	0.377
Language other than English spoken at home (%)	51.5	51.6	-0.1	0.983
Language spoken at home missing (%)	1.3	1.1	0.2	0.761
Mother's education level (%)				
Did not finish high school	17.1	16.6	0.6	0.802
High school diploma or GED certificate	22.6	25.5	-2.9	0.276
Completed some postsecondary education	32.6	36.3	-3.7	0.198
Don't know	25.4	20.9	4.5	0.092
Missing	2.3	0.7	1.6 *	0.050
Father's education level (%)				
Did not finish high school	14.2	21.1	-6.9 *	0.002
High school diploma or GED certificate	22.6	17.2	5.4 *	0.033
Completed some postsecondary education	22.5	23.5	-1.0	0.696
Don't know	38.0	36.7	1.2	0.684
Missing	2.7	1.4	1.3	0.147
GRADE reading comprehension[b]				
Average standard score	84.5	85.0	-0.5	0.160
Corresponding grade equivalent	*4.9*	*5.0*		
Corresponding percentile	*14*	*15*		
2.0 - 3.0 years below grade level (%)	30.0	33.2	-3.1	0.269
3.1 - 4.0 years below grade level (%)	29.4	30.9	-1.5	0.608
4.1 - 5.0 years below grade level (%)	40.5	35.9	4.6	0.128
Sample size	619	437		

(continued)

Table 2.6 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities baseline data.

NOTES: Baseline data for students in 30 of 34 high schools were collected in spring 2006 (during students' eighth-grade year); baseline data for students in the remaining four schools were collected in fall 2006 (at the start of students' ninth-grade year).

The estimated differences are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed distribution of the ERO group across random assignment blocks (i.e., schools) as the basis for the adjustment.

A two-tailed t-test was used to test differences between the ERO and non-ERO groups. The statistical significance level is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

[a]A student is defined as overage for grade if he or she turned 15 before the start of ninth grade.

[b]The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form A). No statistical tests or arithmetic operations were performed on these reference points.

ERO and non-ERO groups participating in this study were determined through a random assignment process. The non-ERO group serves as a benchmark, or "counterfactual," for how students in the ERO group would have performed if they had not had access to the programs. Therefore, the impacts (differences in outcomes between the ERO and the non-ERO groups) represent the effects that the ERO programs had on students' reading achievement and other outcomes over and above what the students would have achieved had they stayed in their regularly scheduled elective class.

This section of the chapter discusses several technical issues that lie at the heart of the evaluation's capacity to produce valid and reliable estimates of the literacy interventions' impacts on student reading achievement and other outcomes. It first reviews the study's sample sizes and the implications for statistical power (that is, the precision with which the analysis can measure program impacts). The section then briefly describes the statistical model being used to estimate the impacts of ERO, and it discusses the standards used for indicating statistical significance (that is, the confidence one may have that the impact estimates are not zero).

Sample Sizes and Statistical Power

To ensure that the ERO impact evaluation could produce valid and reliable findings, several design features were put in place to enable the study to measure program effects (if they exist) that are large enough to be both meaningful in students' lives and relevant to policy de-

bates about the efficacy of supplemental literacy interventions.[41] The number of schools and the number of student sample members are crucial factors that determine the degree to which the impacts on student achievement and other outcomes can be estimated with enough precision to reject with confidence the hypothesis that the program had no effect. In general, larger sample sizes provide more precise impact estimates.

An important goal for the design of the ERO study was to ensure that the sample sizes would be sufficient to allow for estimates of even "small" impacts on reading test scores and other outcomes both overall and for each of the supplemental literacy programs separately.[42] The overall study sample is equipped to detect impacts as small as 0.06 standard deviation units (referred to as "effect sizes"). These pooled impact estimates provide insight into the impact of the family of interventions that share characteristics with RAAL and Xtreme Reading. The samples for each of the two supplemental reading programs are equipped to detect impacts with effect sizes as small as approximately 0.09 standard deviation.

Calculations of statistical power also indicate that the study can detect a differential impact between the two programs of 0.30 standard deviation, which is three times larger than the minimum detectable effect size (MDES) for the program-specific impacts.[43] However, as noted in Chapter 1, the purpose of this study is not to test for a differential impact between RAAL and Xtreme Reading, given the many similarities that exist between these two programs.[44]

Statistical Model for Estimating Impacts and Statistical Significance

The ERO impact analysis uses ordinary least squares (OLS) regression to estimate the difference in outcomes between the ERO and the non-ERO groups, adjusted for the blocking of

[41]Appendix C provides a more detailed assessment of the statistical power of the ERO study's impact design and discusses the role of other design features and assumptions, including the use of pre-random assignment characteristics to improve precision and assumptions about fixed versus random effects.

[42]There are no universally agreed-upon standards for what constitutes "small" versus "large" impacts. Some attempts have been made to examine the range of effects that have been found across an array of evaluations and to divide this range into segments that reflect the higher, middle, and lower categories of effects (see Lipsey, 1990). More recent work has begun to examine actual year-to-year rates of growth on a variety of achievement measures for students in a range of school districts and with a variety of background characteristics (see Bloom, Hill, Black, and Lipsey, 2006). These analyses provide additional background for interpreting the impact of interventions like those in the ERO study within the context of the expected growth in student outcomes nationally and under similar conditions.

[43]The MDES for the difference in impacts between programs is larger than the MDES for the program-specific impacts because the standard error for the difference in impacts is a combination of the uncertainty around each of the two program-specific impact estimates.

[44]The actual precision of estimated impacts may differ somewhat from those calculated for the analyses of statistical power presented in Appendix C. These differences are due to such factors as actual variation in samples sizes, random assignment ratios, pretest scores, and outcome levels across sites.

random assignment by school. In order to improve the precision of the impact estimates, the analysis also controls for random differences between the ERO and the non-ERO groups in their GRADE reading comprehension test score at baseline and whether they are overage for grade (and likely to have been retained in a prior grade). Another key feature of the impact model is that it utilizes a school fixed-effects specification; this means that the findings presented in this report should be interpreted as the estimated impact of the ERO programs in the set of schools in which these interventions were implemented. In other words, the impact estimates are not generalizable to other schools or school districts, nor to alternate allocations of the two interventions (RAAL, Xtreme Reading) to the study schools.[45] Further details on the impact model are presented in Appendix E. As noted earlier, the impact analysis in this report is based on the second cohort of study participants only.

The statistical significance of the impact estimates is assessed using a two-tailed t-test. *Statistical significance* is a measure of the degree of certainty one may have that some nonzero impact actually occurred. If an impact estimate is statistically significant, then one may conclude with some confidence that the program really had an effect on the outcome being assessed. If an impact estimate is not statistically significant, then the nonzero estimate is more likely to be a product of chance. For the purposes of this report, statistical significance is indicated in the tables by an asterisk (*) when the p-value of the impact estimate is less than or equal to 5 percent.

When making judgments about statistical significance, however, it is important to recognize the potential problems associated with conducting multiple hypothesis tests. Specifically, when hypothesis tests are conducted for estimated impacts on several different outcomes and for many subgroups of students and schools, this increases the likelihood of concluding that a given impact estimate is statistically significant when, in fact, it is not. (As noted above in discussing Table 2.2, this is a Type I error, or a false positive.) While it is important to avoid making conclusions based on such errors, the analysis should also not be so conservative with respect to producing false positive results that it unduly increases the likelihood of missing true impacts when they exist (that is, of relying on false negative results).

This study uses two sets of safeguards to attenuate the risk of drawing inappropriate conclusions about program effectiveness on the basis of statistically significant results that may have

[45] As noted in Chapter 1, a fixed-effects model was chosen because the purpose of this study is not to test for a differential impact between Xtreme Reading and RAAL; hence, the distribution of programs to schools can be assumed as given. Had the purpose of the study been to test for a differential impact, the analysis would have had to allow for variability in the assignment of programs to schools, which would have required that a school random-effects model be used.

occurred by chance.[46] The first safeguard is to confine the analysis to a parsimonious list of outcome measures and subgroups. The shorter this list, the fewer the number of hypothesis tests and, thus, the less exposed the analysis will be to "spurious statistical significance" as a result of having tested multiple hypotheses. The primary evidence of overall ERO program effectiveness for this report is reflected by estimates of program impacts on the second cohort's reading comprehension test scores (expressed in standard score values) for the full study sample and for each of the two ERO programs being evaluated. Vocabulary knowledge and student reading behaviors, while targets of the interventions and important to students' literacy development, are considered secondary indicators of program effectiveness. Similarly, subgroups of students and subgroups of schools provide useful information about the relative impact of supplemental literacy programs, but they too are considered secondary indicators of effectiveness in this report.

The second safeguard uses composite statistical tests to "qualify" or call into question multiple hypothesis tests that are statistically significant individually but that may be due to chance in the context of mixed results.[47] These statistical tests are applied in cases where impacts are estimated for more than one outcome in a given measurement domain (for example, the three survey measures that attempt to capture students' reading behaviors) or for subgroups of the full study sample. In general, these qualifying statistical tests estimate impacts on composite indices that encompass all of the measures in a given domain, or estimate the overall variation in impacts across subgroups.[48] If the results of these tests are not statistically significant, this indicates that the statistical significance of the associated individual impact estimates *may have* occurred by chance. In these cases, the discussion of the impacts includes cautions or qualifiers about the robustness of the individual findings.

Comparison of Year 1 and Year 2

The number of students who participated in the ERO programs was similar in both years of the study. In the first year of implementation, 3,339 eligible and consenting students were recruited into the study, compared with 3,441 in the second year of the study. In both cohorts, 57 percent of eligible students were randomly assigned to enroll in ERO classes (1,911 in the first year and 1,946 in the second year). That said, the number of students reading in the tar-

[46]See Appendix E for a more detailed discussion of the approach used to address the risks associated with multiple hypothesis testing.

[47]Measurement of overall effects has its roots in the literature on meta-analysis (see O'Brien, 1984; Logan and Tamhane, 2003; and Hedges and Olkin, 1985). For a discussion of qualifying statistical tests to account for the risk of Type I error, see Duflo, Glennerster, and Kremer (2007). Other applications of these approaches are discussed in Kling and Liebman (2004) and in Kling, Liebman, and Katz (2007).

[48]See Appendix E for a more detailed description of the method used to conduct these qualifying statistical tests. Appendix E also includes tables with the results of these tests.

get range of two to five years below grade level — and therefore included in the study sample for the impact analysis — was somewhat larger in the first year of the study (2,916 students) than in the second year of the study (2,679 students).

Response rates on the spring follow-up reading comprehension test are similar for both cohorts of students. In the first year of implementation, 82.8 percent of students in the study sample completed the follow-up GRADE assessment, compared with 81 percent in the second cohort of students. The difference in response rates between the two cohorts is not statistically significant. As noted above, however, the study sample was larger in the first year of the study. Consequently, the respondent sample used in the impact analysis was larger in the first year of the study (2,413 students) than in the second year of the study (2,171 students). In each year of the study, random assignment resulted in statistically equivalent ERO and non-ERO groups, both in the full study sample and in the respondent sample used for the impact analysis.

Table 2.7 compares the background characteristics of the respondent sample in the first and second years of the study. The most notable difference between the two cohorts is in terms of their reading achievement at baseline. Specifically, the average baseline reading comprehension score of the first cohort of students is 1.1 standard score points higher than that of the second cohort of students. This difference is consistent with the fact that the two cohorts of students were tested in different grades: all students in the first cohort were tested in the fall of their ninth-grade year, while most students in the second cohort were tested in the spring of their eighth-grade year.[49] A more useful metric for comparing the achievement of the two cohorts is the amount by which students were below grade level at the time they were tested. In this respect, Table 2.7 shows that the distribution of students across three categories of this metric does not differ by a statistically significant amount between the two cohorts.

The second notable difference between the two cohorts is with respect to the quantity of missing data in respondents' background characteristics. In particular, Table 2.7 shows that, when completing the baseline survey, students in the second cohort were more likely than students in the first cohort to provide information on their parents' education and the language spoken in their home. This cross-cohort difference in nonresponse makes it difficult to compare the parental education and linguistic composition of the two cohorts of students, because what appears to be a difference between the two cohorts with respect to the education of their parents or the language spoken in their home may, in fact, be due to a more accurate assessment of students' background characteristics in the second year. For example, Table 2.7 shows that, in the

[49]In one school district (four ERO high schools), students in the second cohort were tested in the fall of ninth grade.

The Enhanced Reading Opportunities Study

Table 2.7

Characteristics of Students in Cohort 1 and Cohort 2 Follow-Up Respondent Sample

Characteristic	Cohort 1	Cohort 2	Difference	P-Value for the Difference
Race/ethnicity (%)				
Hispanic	31.2	31.6	-0.4	0.649
Black, non-Hispanic	45.6	46.6	-1.0	0.354
White, non-Hispanic	16.8	15.2	1.6	0.093
Other	6.4	6.6	-0.2	0.826
Gender (%)				
Male	50.6	51.3	-0.7	0.630
Female	49.4	48.7	0.7	0.630
Average age as of September 1st (years)	14.7	14.7	0.0	0.249
Overage for grade[a] (%)	26.6	26.2	0.4	0.748
Language other than English spoken at home (%)	45.1	49.7	-4.6 *	0.001
Language spoken at home missing (%)	7.2	1.2	6.0 *	0.000
Mother's education level (%)				
Did not finish high school	16.8	17.4	-0.6	0.577
High school diploma or GED certificate	25.2	24.6	0.6	0.631
Completed some postsecondary education	30.0	32.3	-2.2	0.093
Don't know	20.1	23.9	-3.8 *	0.002
Missing	7.9	1.8	6.1 *	0.000
Father's education level (%)				
Did not finish high school	16.3	16.2	0.2	0.879
High school diploma or GED certificate	23.1	22.2	0.9	0.467
Completed some postsecondary education	19.7	21.6	-1.9	0.100
Don't know	32.1	37.5	-5.4 *	0.000
Missing	8.8	2.5	6.3 *	0.000
GRADE reading comprehension at time of testing[b]				
Average standard score	85.8	84.7	1.1 *	0.000
Corresponding grade equivalent	*5.1*	*4.9*		
Corresponding percentile	*16*	*14*		
2.0 - 3.0 years below grade level (%)	34.7	32.5	2.3	0.108
3.1 - 4.0 years below grade level (%)	28.0	28.5	-0.5	0.722
4.1 - 5.0 years below grade level (%)	37.2	39.0	-1.8	0.215
Sample size	2,413	2,171		

(continued)

Table 2.7 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities baseline data.

NOTES: Baseline data for students in Cohort 1 were collected at the start of students' ninth-grade year (fall 2005). For students in Cohort 2, baseline data in 30 of 34 high schools were collected in spring 2006 (during students' eighth-grade year); baseline data in the remaining four schools were collected in fall 2006 (at the start of students' ninth-grade year).

The estimated differences are regression-adjusted using ordinary least squares, controlling for indicators of school. The values in the column labeled "Cohort 2" are the observed means for students in the second cohort of study participants. The "Cohort 1" values in the previous column are the regression-adjusted means for students in the first cohort of study participants, using the observed distribution of Cohort 2 students across schools as the basis for the adjustment.

A two-tailed t-test was used to test differences between Cohort 1 and Cohort 2. The statistical significance level is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

[a]A student is defined as overage for grade if he or she turned 15 before the start of ninth grade.

[b]The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form A). No statistical tests or arithmetic operations were performed on these reference points.

second year of the study, a larger percentage of respondents came from a home where a language other than English is spoken; however, if nonresponse in the first year was not random (for example, if students who did not answer this survey item in the first year were from multilingual homes), then the higher percentage of students from multilingual homes in the second year is partly explained by a higher response rate among these students in the second year and does not entirely represent a true difference in the linguistic composition of the first and second cohorts.[50] Hence, while a general F-test indicates that there is a systematic difference between the first and second cohorts of respondents in terms of their background characteristics, many of the background differences are consistent with changes in the timing of student testing and an improvement in the quality of student survey data in the second year.[51]

[50]More generally, if nonresponse in the first year of the study was not random (that is, if nonresponse was a function of students' home language or the education of their parents), then, by definition, the average parental education and linguistic composition of students in the second year will change relative to the first year, even if there is no true difference between the two cohorts with respect to these characteristics. Hence, assuming that nonresponse is, in fact, nonrandom, then some portion of the between-cohort difference in parental education and linguistic composition is due to better data quality in the second year, rather than to a true difference in the background characteristics of the two cohorts of students.

[51]Differences between the first and second cohorts of students were also examined by program (RAAL and Xtreme Reading). These results follow the same patterns as those described for the overall study sample.

Chapter 3

Implementing the Supplemental Literacy Programs

This chapter describes the two supplemental literacy programs used in the Enhanced Reading Opportunities (ERO) study and assesses the fidelity of their implementation during the study. The chapter first provides an overview of the process used to select the programs at the start of the study, and it describes the programs' core elements as presented in the proposals submitted by their developers and in other literature and materials associated with the programs. Then the chapter focuses on the second year of program implementation, presenting the background characteristics of the ERO teachers and describing the training activities and technical support that they received. The chapter next discusses findings on the fidelity with which each of the supplemental literacy programs was implemented in the participating high schools. The chapter concludes with a comparative discussion of factors affecting the first year of implementation and how the second year of implementation was different.

Briefly revisiting here the findings regarding the first year of implementation helps provide context for the chapter's focus on the second year. In the spring of the first year of the study, a site visit was conducted by the study team to assess program implementation fidelity in the 34 ERO high schools. Based on this visit, the study team found that ERO programs in 16 of the 34 high schools were considered "well aligned" with the requisite components, as outlined by the developers, for successful implementation of the program models. Eight of the programs were found to be "moderately aligned" with the models, and 10 of the schools had enough implementation problems for the programs to be considered "poorly aligned." Based on ongoing interactions with the ERO high schools and the site visits conducted in the first year of the study, the study team identified three challenges to implementation: the programs started later than expected; some schools did not have the required program materials and resources when the programs began; and the programs themselves were completely new to the schools and teachers, thus presenting a steep learning curve that had to be surmounted in a short amount of time. These first-year findings and challenges are revisited at the end of the chapter.

The chapter makes several key points about program selection and design that pertain to both years of implementation:

- The two programs evaluated were selected by an independent national panel of adolescent literacy experts from among 17 proposals through a competitive process.
- Both programs focus on establishing a positive learning environment in the classroom to facilitate the delivery of instruction in reading comprehension

processes and strategies. The comprehension instruction seeks to make explicit the processes used by capable readers, teaching less proficient students to pay attention to *how* they read so that they can improve their comprehension of *what* they read.

- Teachers were identified by schools as appropriate for teaching ERO and subsequently self-selected to teach the programs. They were then approved by the districts and the U.S. Department of Education (ED). They held a high school teaching license or certificate and had an average of over 11 years of teaching experience.

This chapter also raises key points about implementation fidelity in the second year of the study and compares implementation fidelity findings across the two years of the study:

- Of the 34 teachers who participated in the second year of the study, 25 had taught all of the first year, and two had taught a portion of the first year (having replaced a teacher midyear). Seven teachers were new to the programs and the study at the outset of the second year. Teachers in their first year of teaching all received the requisite training to familiarize them with the programs and the study before they attended the second-year summer trainings with the teachers who taught during the first year. All teachers taught ERO the entire school year.

- Site visits were conducted by the study team in the middle of the fall and the middle of the spring of the second year of the study, to measure the implementation fidelity of the two ERO programs. Based on the fall site visit, the implementation of the ERO programs in 20 of the 34 participating high schools was deemed to be well aligned with the respective program models' requirements for high-fidelity learning environments and comprehension instruction. The level of implementation was found to be moderately aligned at 11 of the schools and poorly aligned at three of the schools with the developers' implementation specifications. Based on the spring site visit, implementation of the programs was considered well aligned in 26 of the schools, moderately aligned in seven of the schools, and poorly aligned in one school.[52]

[52]Statistical tests of the difference in average fidelity ratings between subgroups of schools and across implementation years were not conducted, because fidelity ratings are available for all schools in the sample. Hence, on the days during which teachers were observed and given this allocation of programs to schools, observed differences in fidelity ratings are real differences in the ratings. In other words, there is no error in the (continued)

- Average composite scores of implementation fidelity in the second year of the study are similar at the 25 schools with returning teachers who had taught all of the first year and at the nine schools with replacement teachers (average ratings of 2.4 for returning teachers and 2.3 for replacement teachers at the spring site visits, out of a maximum score of 3).[53]

- In comparison with the first year of implementation, the study team found in the second year that more schools' ERO programs were deemed to be well aligned with the program developers' specifications for implementation fidelity (26 schools in the second year, compared with 16 schools in the first year). In addition, fewer schools' programs were considered to be poorly aligned with the program models (one school in the second year, compared with 10 schools in the first year).

- The study team found that, on average, the 25 teachers who returned to teach the ERO programs after having taught all of the first year had higher implementation ratings in the second year than in the first year. They established learning environments that were better aligned to the program models (average ratings of 2.5 in the second year and 2.4 in the first year). They had higher implementation-fidelity ratings in terms of their comprehension instruction in their second year teaching the ERO programs than in their first year (an average rating of 2.3 in the second year, compared with 2.0 in the first year).

- The study team also found that, on average, the nine replacement teachers had higher implementation-fidelity ratings in terms of both the learning environment established in their classrooms and their comprehension instruction (average ratings of 2.4 and 2.3, respectively) than the teachers they replaced (average ratings of 1.7 and 1.8, respectively).

Characteristics of the Supplemental Literacy Programs: Reading Apprenticeship Academic Literacy and Xtreme Reading

The supplemental literacy programs were selected through a competitive proposal process that was managed by the study team and guided by a panel of seven nationally known

---

average fidelity ratings due to the sampling of schools or due to the selection of observation days, given that the fidelity data presented in this chapter are not used to make generalizations or inferences about the implementation fidelity that would have been observed at other points in time or based on a different allocation of programs to schools.

[53]Refer to Appendix Table D.5.

experts in adolescent literacy research and program development. Because the intent of the study is not simply to evaluate a specific literacy program but, rather, a *type* of literacy program, the process was designed to select two programs to test for effectiveness. While each of the two programs could be tested individually, together they could be tested as representative of a class of intervention. A request for proposals (RFP) was advertised in a wide range of education publications and was disseminated to over 40 organizations that develop and implement high school curricula.[54] The RFP specified that prospective supplemental literacy programs must be research-based, high-quality programs that provide instruction in the areas that experts increasingly agree are necessary for effective adolescent literacy instruction, as outlined in *Reading Next,* but that were not yet rigorously tested.[55] The prospective programs were to have been developed already (that is, not be new programs) and to be ready for systematic use in multiple schools and districts.

Seventeen proposals were submitted in response to the RFP. After a review of the research base presented in the proposals for each program, the proposals were rated by the panel of adolescent literacy experts. The developers of four of the proposed programs were invited to give oral presentations before the panel, staff from ED, and the ERO study team. Based on the presentations and subsequent discussion, the panelists recommended and ED accepted two programs for inclusion in the study: WestEd's Reading Apprenticeship Academic Literacy (RAAL) and the University of Kansas Center for Research on Learning's (KU-CRL) Xtreme Reading.

Overall Goals and Approach

The overarching goals of both RAAL and Xtreme Reading are to help students adopt the strategies and routines used by proficient readers, improve their comprehension skills, and motivate them to read more and enjoy reading. Both programs emphasize the importance of establishing a specific type of learning environment in the classroom that is conducive to the effective delivery of the core instructional strategies by the teacher and to facilitate student and teacher interactions around the reading skills that are being taught and practiced. They both use a "cognitive apprenticeship" approach to instruction in which the teacher initially takes the lead in modeling the strategies that proficient readers use and then gradually increases the responsibility of the students to demonstrate and apply these strategies. The teachers seek to make explicit *how* proficient readers read, and they support their students in recognizing and using the strategies or methods used by stronger readers. That is, both programs focus students' attention on *how* they read (a metacognitive process) to help the students better understand *what* they

[54]American Institutes for Research (2004).
[55]Biancarosa and Snow (2004).

read (understanding content). Also, both programs integrate direct, whole-group instruction with small-group and individualized instruction.[56]

Key Components

The key components of RAAL and Xtreme Reading are discussed categorically below. This discussion is based on information provided by the two program developers. Table 3.1 also presents these components by category. These components are the specific aspects of the programs' instructional approaches that the developers expect to improve the literacy skills of high school students.[57]

Developers' Implementation Philosophy

In implementing RAAL, teachers are guided by the concept of "flexible fidelity." That is, while the program includes a detailed curriculum, the teachers are trained to adapt their lessons to meet the needs of their students and to supplement program materials with readings they expect to be motivating to their classes. Teachers have flexibility in how they include various aspects of the RAAL curriculum in their day-to-day teaching activities, but they have been trained to do so such that they maintain the overarching spirit, themes, and goals of the program in their instruction.

Xtreme Reading was developed with the philosophy that the presentation of instructional material — particularly the order and manner in which the material is presented — is of critical import to the students' understanding of it, and, as such, teachers are trained to deliver course content and materials in a precise, organized, and systematic fashion designed by the developers. Xtreme Reading teachers follow a prescribed implementation plan, following specific day-by-day lesson plans in which activities have allotted segments of time within each class period. However, there are opportunities in the Xtreme Reading instructional program for teachers to use responsive instructional practices to adapt and adjust to student needs that arise as they move through the highly structured curriculum.

[56]Additional information about the Reading Apprenticeship Academic Literacy course is available on the Internet at http://www.wested.org/cs/we/view/serv/111; information about the Xtreme Reading course is available at http://www.xtremereading.org/. Furthermore, the descriptive material about the program-specific observation rating scales in Appendix D provides more information specific to each program.

[57]The proposals submitted by the two developers, WestEd (2004) and University of Kansas (2004), contain information about the key components of their programs. These proposals are unpublished and cannot be released based on the rules of the competition through which the programs were selected.

The Enhanced Reading Opportunities Study

Table 3.1

Key Components of the ERO Programs

	WestEd/Reading Apprenticeship	KU-CRL/Xtreme Reading
Developer's Implementation Philosophy	"Flexible fidelity" guided by the instructional and behavioral/social needs of the students	Prescribed daily lesson plans and time limits on classroom activities
Role of Teacher	Instructor as "master reader," apprenticing students in various literacy competency areas and drawing on variety of materials	Instructor explicitly teaches seven reading strategies using a prescriptive eight-stage instructional approach with step-by-step instructional materials
Curriculum Design	<u>Learning Environment</u> Establish "social reading community" early in program <u>Comprehension Instruction</u> Five curricular strands of classroom instruction: 1. Metacognitive Conversation 2. Silent Sustained Reading 3. Language Study 4. Content/Theme 5. Writing	<u>Learning Environment</u> Focus at beginning of course on teaching social and behavioral skills and strategies aimed to develop a productive and positive classroom learning environment <u>Comprehension Instruction</u> Focus of rest of course on developing literacy skills through seven learning strategies: 1. LINCS Vocabulary Routine 2. Word Mapping 3. Word Identification 4. Self-Questioning 5. Visual Imagery 6. Paraphrasing 7. Inferencing
Teaching Strategies	Instructors usually use one or two of the following routines during class period: 1. Think aloud 2. Talking to the text 3. Metacognitive logs/journals 4. Preambles (daily warm-ups)	Each strategy is taught using a prescribed eight-stage instructional methodology: 1. Describe 2. Model 3. Verbal practice 4. Guided practice 5. Paired practice 6. Independent practice 7. Differentiated instruction 8. Integration and generalization
Program Type	Supplemental course, like an elective	Supplemental course, like an elective
Duration	One school year	One school year

Role of Teacher

Both RAAL and Xtreme Reading are grounded in the principle of a cognitive apprenticeship. That is, the teacher assumes the role of reading expert whose task is to share expertise in explicit ways with the students and then to support their development of those skills and nurture their increased independence in using them. The process is one that starts off as teacher-centered and gradually transitions to being more student-centered, as students assume more responsibility for monitoring their own comprehension and adjust their use of skills as needed. In RAAL — where the teacher is considered the "master reader" for the students, who are the "reading apprentices" — the transition is facilitated through the teacher's integration of the four dimensions of classroom life (personal, social, cognitive, and knowledge-building; described below), which he or she links together through ongoing metacognitive conversations (thinking internally and talking externally about reading processes).

For the Xtreme Reading teacher, this transitional process is a specific eight-stage instructional model through which seven specific literacy strategies are taught. In Xtreme Reading classes, the expectation is that the learning of each strategy begins with specific teacher-directed instruction and that control is relinquished to students incrementally as they progress through the stages. By the eighth stage, students are working independently and have an understanding of the application of the strategy outside the Xtreme Reading classroom.

Curriculum Design and Teaching Strategies

As discussed above, the two programs are attentive to both the learning environment in the classroom and the nature of the literacy instruction, particularly around reading comprehension. The curriculum design and the teaching strategies of the two ERO programs reflect these two priorities. Table 3.1 provides an overview of the key elements of each ERO program. The developers' curriculum designs both highlight the equal importance of creating a conducive classroom learning environment and focusing instruction on strategies that promote reading comprehension skills and proficiency.

The core of the RAAL program is the integration of four dimensions: social, personal, cognitive, and knowledge-building. The social and personal dimensions reflect the attention of the program to the *learning environment* for the class. The social dimension refers to adolescents' interests in peer interaction and in larger social, political, and cultural issues. The personal component addresses students' own goals for reading and for reading improvement. These aspects of the program are combined in the establishment of a social reading community, a classroom environment that allows for the respectful, open exchange of ideas considered essential for the program to have effective comprehension instruction.

The cognitive and knowledge-building dimensions are the *instructional components* of the RAAL program. They address students' needs to increase both their repertoire of comprehension strategies and their background knowledge — expanding their knowledge base through reading and providing knowledge about aspects of strong reading, such as word construction, vocabulary, text structure, and figurative language. The instructional components are delivered across the following three major thematic units during the school year: "Who Am I as a Reader?" "Reading History," and "Reading Science and Technology." Within each unit, the teacher incorporates the five key curricular strands of the program:

- **Metacognitive conversations.** The students and the teacher think and talk about the thinking processes that are engaged when reading.
- **Silent sustained reading.** The student reads a book of his or her choice for 20 to 25 minutes at least twice a week to build reading fluency, comprehension, motivation, and stamina.
- **Language study.** The teacher and the students routinely practice strategies and learn skills at the word, sentence, and text levels to enhance language development.
- **Content/theme.** The teacher uses the majority of instructional time to address one of the three thematic units of the curriculum so that students are able to apply what they are learning in the classroom to their other classrooms and relate what they are learning to contexts other than RAAL.
- **Writing.** The teacher provides opportunities for the students to write and provides new knowledge of writing processes and strategies as needed.

The curriculum strands are taught and reinforced through the use of four teaching strategies: *think alouds, talking to the text, metacognitive logs,* and *daily preambles.* These strategies offer teachers and students opportunities to interact around what they are reading and how they are reading.

The Xtreme Reading program also emphasizes creating a positive *learning environment* in the classroom. The program aims to create a structured classroom climate with explicit social and behavioral expectations and regular routines for both students and teachers. The main tenet of classroom management is time-on-task behavior; this is essential to successful implementation of the instructional sequence. Student motivation and engagement are encouraged through several activities that help students set short- and long-term goals for their learning and through the availability and sharing of high-interest novels about students who have overcome academic obstacles. Teachers seek to help students to set real purposes for learning and to link their learning to personal goals.

The program's *literacy instruction* involves both a systematic component (driven by the curriculum) and a responsive component (driven by student needs). The systematic component involves teaching seven reading strategies following lesson plans provided by the developer that map out daily instruction. Two strategies focus explicitly on vocabulary: LINCS and Word Mapping. Five strategies focus more directly on comprehension: Word Identification, Self-Questioning, Visual Imagery, Paraphrasing, and Inferencing. Each strategy is taught using an eight-stage model that starts off being highly teacher-centered (the teacher describes and models the strategy in the first two stages), then entails shared work between the teacher and the students (verbal and guided practice), and becomes more and more the responsibility of the students (paired practice between students and independent student practice). The seventh stage is differentiated instruction, allowing those struggling with the strategy to receive additional support and those who have been successful learning the strategy more and varied opportunities for practice. The eighth stage, integration and generalization, involves students' taking the strategy beyond the Xtreme Reading classroom and materials and applying it to reading in other classes. The responsive instruction component focuses on assessing and addressing individual student needs as they arise. The responsive instruction component represents where flexibility enters into Xtreme Reading instruction.

Both ERO programs were developed from preexisting programs prior to implementation in the ERO study. The program developers adapted their already existing curricula to create programs that would be supplemental, yearlong reading classes. The Reading Apprenticeship Academic Literacy curriculum combined elements of two WestEd programs, Reading Apprenticeship and Academic Literacy. These programs had been the focus of most of the work within WestEd's Strategic Literacy Instruction initiative. Instruction in Reading Apprenticeship helps students identify weaknesses in their reading skills and improve them through mastering and then consciously applying advanced reading strategies. Academic Literacy is usually woven into content-area instruction so that students learn to apply subject-specific skills and strategies in areas such as science and social studies. The curriculum used in this study offered instruction in strategic reading within three themed units, two of which emphasized content-area reading. The Xtreme Reading curriculum combined the components of the Strategic Instruction Model (SIM) for reading improvement that has been developed, studied, and refined at the University of Kansas Center for Research on Learning for close to 30 years. SIM content consists of six specific reading processes, such as vocabulary identification and strategies for making inferences from the text. Previous implementation of SIM had followed the eight-stage instructional model used in Xtreme Reading but had not combined the six reading strategies into a full-year curriculum for use in self-contained intervention classes. Further, two versions of this curriculum were developed to accommodate both 45- and 90-minute instructional blocks.

The ERO Teachers and Their Preparation for the ERO Programs

Teachers play a key role in both programs selected for the study. The study sought to have experienced, core-content-area teachers implement the programs and to provide adequate training and support for them. The teachers were nominated by their schools on the grant applications submitted to the Office of Vocational and Adult Education (OVAE) at ED.[58] Additionally, participating districts and schools committed to make these teachers available for professional development activities prior to the start of the school year and on an ongoing basis during the study.

Teacher Characteristics

The Request for Proposals from OVAE to which school districts responded in their application for grant funding and participation in this study specified that teachers selected to teach the ERO classes at each high school should have at least two years of experience and be certified core-content-area teachers — specifically, English or social studies teachers — and not necessarily reading specialists. The project sought to target content-area teachers rather than reading teachers to teach the classes in order to enhance the replicability of the interventions if they proved to be effective. First, the study sought to demonstrate that if content-area teachers could be trained to deliver a literacy program, schools and districts that later chose to pursue this type of intervention may have a more realistic chance to identify staff to teach it without being restricted to reading specialists. Second, one of the goals of both interventions is transference — helping students use the literacy skills that they develop in their content-area classes. Thus, it was hoped that involving content-area teachers would help facilitate this.

Table 3.2 provides a list of background characteristics for the teachers in each of the two ERO programs.[59] The average number of years of previous experience for ERO teachers was 11.9 years, although prior teaching experience ranged from student teaching to over 30 years as a regular classroom teacher. Over two-thirds (68 percent) of the teachers had graduate-level degrees, and all held high school-level certification. The majority of the teachers (85 percent) were certified in English/language arts, and nearly 9 percent held social studies certification while 6 percent held certification in some other area. Teachers reported attending an aver-

[58]The implementation was initially funded by the Office of Vocational and Adult Education (OVAE), but this role was later transferred to the Office of Elementary and Secondary Education (OESE).

[59]Information in Table 3.2 is drawn from the survey that teachers completed at the beginning of the ERO training or at the beginning of their tenure as an ERO teacher.

The Enhanced Reading Opportunities Study

Table 3.2

Background Characteristics of ERO Teachers

Characteristic	All Schools	Reading Apprenticeship Schools	Xtreme Reading Schools
Race/ethnicity (%)			
Black	20.6	29.4	11.8
White	70.6	58.8	82.4
Other	8.8	11.8	5.9
Gender (%)			
Male	11.8	0.0	23.5
Female	88.2	100.0	76.5
Total time teaching (years)	11.9	11.3	12.4
Total time teaching at current school (years)	5.4	5.2	5.6
Total time teaching at current level (years)[a]	8.0	8.0	8.0
Total time teaching English/language arts or social studies (years)[a]	10.3	10.1	10.6
Master's degree or higher at the start of ERO program (%)	67.7	64.7	70.6
Holds high school-level teaching certification (%)	100.0	100.0	100.0
Subject matter certification (%)			
Certified in English/language arts	85.3	76.5	94.1
Certified in social studies	8.8	11.8	5.9
Certified in other subject	5.9	11.8	0.0
Number of professional development workshops attended in the last two years prior to becoming an ERO teacher[a]	4.2	4.7	3.7
Number of hours spent in professional development workshops during the last two years prior to becoming an ERO teacher[b]	41.4	35.5	48.2
Taught the ERO class for the full school year (%)	100.0	100.0	100.0
Taught the ERO class in its first year (%)	79.4	76.5	82.4
Sample size	34	17	17

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study baseline teacher survey.

NOTES: This table contains data from the baseline teacher survey for Cohort 1 and Cohort 2 administered to teachers at their point of entry into the ERO demonstration.

Rounding may cause slight discrepancies in calculating sums and differences.

[a]Missing data: One teacher did not respond.

[b]Missing data: Six teachers did not respond.

age of 41.4 hours of professional development in the two years prior to the beginning of the ERO program.[60]

Of the 34 ERO teachers in the second year of program implementation, 25 had returned after having taught all of the first year of the program. Two of the teachers replaced first-year teachers in the middle of the first year. Seven teachers replaced first-year teachers who left at the end of the first year.[61] Given that the replacement teachers came to the second year of implementation with no experience with the ERO programs, there were differences in their training and support compared with that of the returning teachers. These differences are discussed in the next section of the chapter. Also, potential differences in implementation findings for the nine replacement teachers, the teachers they replaced, and the teachers who participated in both years of the study are discussed below in the chapter.

Training and Technical Assistance

Training and technical assistance for the second year of implementation were delivered to the ERO teachers in the following ways: Both new and returning RAAL teachers attended one 3-day summer training institute as well as two 2-day booster training sessions during the 2006-2007 school year. Prior to the summer training, new RAAL teachers participated in a national Reading Apprenticeship Institute[62] to learn the central strategies and philosophies of the Reading Apprenticeship program. All RAAL teachers also received ongoing support through three 2-day coaching visits during the second year and access to a special online listserv that was set up by the developer for the project. New RAAL teachers were also offered up to five extra days of coach support if the teacher, coach, and the district coordinator determined that supplemental technical assistance was needed. New Xtreme Reading teachers attended a two-day training immediately prior to a three-day training for all (new and returnees from the first year) Xtreme Reading teachers during the summer before school started in the 2006-2007

[60]Differences between teachers in each ERO program were not tested for statistical significance. There is one ERO teacher per school, which means that teacher characteristics are also school characteristics. As discussed below in the chapter, the impact analysis accounts for differences across school characteristics (and, thus, across teachers) by including regression covariates for each school.

[61]Replacement teachers were identified by the schools and the districts. Their resumes were submitted to the study team for initial screening. The study team screened the resumes for a teaching certification and at least two years of teaching experience. Resumes satisfying those criteria were forwarded to ED for final approval.

[62]The Reading Apprenticeship Academic Literacy course being implemented in the ERO Study is an adaptation of the preexisting Reading Apprenticeship program on which the national workshops were focused. While at the national workshops, these two ERO teachers received additional training that addressed aspects of Reading Apprenticeship that are specific to the ERO Study. A specific supplemental training that focused entirely on RAAL was not offered to teachers because there were very few teachers that were replaced at the end of Year 1 and the economies of scale found in the first-year start up conditions could not be replicated.

school year. All Xtreme Reading teachers also attended one 2-day booster training during the year. They also received three 2-day on-site coaching visits. As was the case in the first year of implementation, district coordinators were again invited to observe (and some did attend) the trainings to familiarize them with the programs in case they had to provide technical assistance or other support to ERO teachers. Table 3.3 summarizes the activities provided to the ERO teachers by each of the developers for the 2006-2007 school year. Figure 3.1 provides a timeline for both years of program implementation, including teacher selection, teacher replacement, and teacher training.

The Enhanced Reading Opportunities Study

Table 3.3

Training and Technical Assistance Provided During the 2006-2007 School Year, by ERO Program

	Summer Training	School-Year Booster Training	Additional Supports
Reading Apprenticeship	One 3-day training for new and returning teachers (August)	Two 2-day trainings (October; March)	Three 2-day on-site coaching visits Weekly e-mail and phone calls Listserv
Xtreme Reading	One 2-day training for new teachers, followed by a 3-day training for both new and returning teachers (July-August)	One 2-day training (January)	Three 2-day on-site coaching visits Weekly e-mail and phone calls Additional technical assistance for replacement teachers

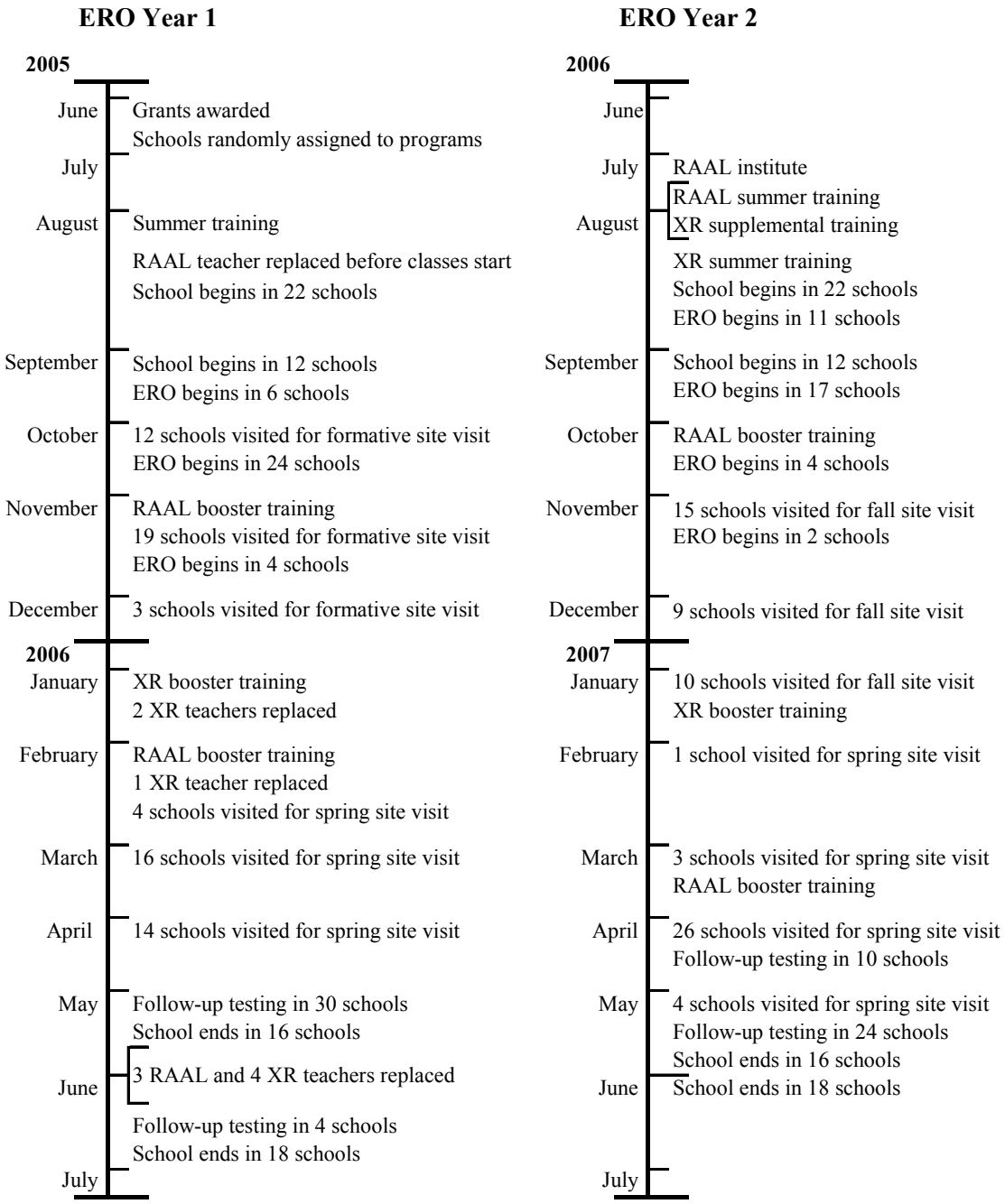
Summer Trainings

The summer teacher training institutes for both programs were conducted in late July or early August 2006. The RAAL training was conducted by the program developer, by experienced RAAL teachers (including some of returning teachers from the study who taught all of the first year and were asked to lead presentations on particular units, as well as teachers who are employees of WestEd), and by the coach who would work with the ERO teachers throughout the year. The training focused on those program areas that were identified by returning ERO teachers or the developers as needing additional support and instruction. The staff at WestEd determined areas of teacher weakness during their coaching visits to the RAAL classrooms throughout the first year.

The Enhanced Reading Opportunities Study

Figure 3.1

Study Timeline



SOURCES: MDRC and AIR project records and documentation.

The trainings included modeling and whole-group discussions led by the developers, the coach, and selected returning teachers. There were also small-group activities, such as roundtable discussions on particular subjects and individual planning periods. Teachers were able to meet with the RAAL coach individually during this training, as well. The summer training was attended by 14 of the 17 RAAL teachers.

The Xtreme Reading training was conducted by the program developers, research staff from the University of Kansas Center for Research on Learning, a behavioral specialist, and the coaches who would work with the teachers throughout the second year. The training mainly addressed challenges faced by ERO teachers during the first year of implementation, though the teaching of new strategies was also a part of the agenda. KU-CRL staff worked with ERO teachers during the Year 1 booster training to identify the challenges that would be discussed at the summer training. Training methods included modeling, discussion, and formal presentations as well as large-group and small-group activities. Teachers also had time to meet with the coaches with whom they would be working during the year. All of the 17 Xtreme Reading teachers attended the summer training.

Booster Trainings

The booster trainings during the school year (two for RAAL and one for Xtreme Reading) were conducted in a similar format to the summer training institutes and were two days each in duration. The program developers used these trainings to address challenges that the teachers were facing with the classes, to refine teaching approaches, and to introduce new material and curriculum changes. Each of the trainings also provided time for the teachers to meet with their coaches and opportunities for the teachers and developers to discuss any issues with the implementation of the program that had come up during the first part of the year. The fall booster training session included 15 of the 17 RAAL teachers, and 14 of them attended the spring session. Of the 17 Xtreme Reading teachers, 15 attended the booster training session in person, and two teachers participated by telephone.

Ongoing Technical Assistance

Both programs provided on-site coaching and electronic and telephone communication among teachers and their coaches. RAAL also made a listserv available to teachers. The RAAL and Xtreme Reading coaches made three 2-day visits to each of the teachers. During these visits, they observed classes, modeled instruction, and in some cases co-taught lessons, in addition to working through issues that each teacher was experiencing.

Implementation Fidelity

The prior sections of the chapter provide descriptive background about the ERO programs, the teachers teaching them, and the preparation and support for those teachers. The discussion now turns to the implementation of the programs. This section of the chapter examines the fidelity with which the two supplemental literacy programs — RAAL and Xtreme Reading — were implemented during the second year of the study. In particular, it defines the method by which composite measures of implementation fidelity were computed for each school, based on classroom observations conducted by study team members during two separate site visits (one in the fall and one in the spring) in the second year of implementation. In the context of this study, "fidelity" refers to the degree to which the observed operation of the ERO program in a given high school approximated the intended learning environments and instructional practices that were specified by the model's developers.

Overall ratings of the implementation fidelity of the ERO programs at each school provide a context for interpreting the study's impact findings and offer information to policymakers and practitioners about factors that they may wish to consider if they are establishing these programs or ones like them in high schools.

Data Sources and Measures

The analysis of ERO program implementation fidelity in the second year of the study is based on two field research visits to each of the 34 high schools — one during the first semester and one during the second semester of the 2006-2007 school year.[63] In the second year of the study, site observations were conducted by one researcher (a senior staff member of the study team) at each school. The site visitors observed 160 to 180 minutes of instruction during each school visit, capturing a "snapshot" of instruction at a particular time in the school year across all 34 participating high schools. They measured implementation fidelity using the same research protocol as used in the first year. The study team felt that using the same instrument in both years of the study was appropriate, given that modifications to the curricula were minor. The classroom observation protocols used in the site visits provided a structured process for observers to rate characteristics of the ERO classroom learning environments and the ERO teachers' instructional strategies. The instrument included ratings for six characteristics (referred to as "constructs" from here forward) that are common to both programs, as well as ratings for seven program-specific constructs. For each construct, a category rating of 1 ("poorly aligned"), 2 ("moderately aligned"), or 3 ("well aligned") was given.

[63] Appendix D provides a more detailed description of the site visits.

The analysis of the classroom observation ratings sought to capture the implementation fidelity of two key overarching dimensions of both programs: the classroom learning environment and the instructional strategies that focused on reading comprehension. A composite measure of implementation fidelity for each dimension was calculated from the average ratings for both general and program-specific constructs. An average rating of 2.0 or higher means that the school's ERO program was well aligned with the developers' implementation specifications; a rating of 1.5 to 1.9 means that the program was moderately aligned; and a rating of 1.0 to 1.4 means that it was poorly aligned.

Table 3.4 provides a list of the constructs that were combined to create composite ratings for the learning environment and comprehension instruction dimensions, respectively, for the ERO programs in each high school. The *learning environment composite* was calculated as the average of ratings on two general constructs and ratings of one or two program-specific constructs for RAAL and Xtreme Reading, respectively. The *comprehension instruction composite* was calculated as the average of ratings on two general constructs and ratings of five program-specific constructs.[64] The composite measures ranged from 1 to 3 and were rounded to the nearest tenth of a point. Based on the composite ratings for each of the two program dimensions — learning environment and comprehension instruction — the implementation fidelity for each dimension was classified as well aligned, moderately aligned, or poorly aligned with the models specified by the program developers.[65]

The ratings and resulting categories indicate whether the programs reflected the characteristics of the classroom learning environments and instructional strategies intended by the developers. While it is reasonable to expect that higher-fidelity programs could produce stronger impacts than programs where the fidelity was only a limited reflection of the intended model, other factors could intervene to make higher-fidelity programs ineffective or to make limited- or inadequate-fidelity programs effective.Findings

Tables 3.5a and 3.5b provide summaries of the findings regarding implementation fidelity for the Year 2 fall and spring site visits, respectively.[66] The top two panels of the tables provide a summary of the number of schools whose composite ratings for each individual site visit

[64]Note that, for Xtreme Reading, the program-specific component comprises two subcomponents: curriculum-driven, or systematic, instruction; and needs-driven, or responsive, instruction. Appendix D provides a detailed description of the method used to average the ratings on individual constructs to create the composites for the two overarching program dimensions.

[65]Appendix D provides a more detailed description of the fidelity-rating classifications.

[66]For a combined summary measure of the spring and fall site visits, refer to Appendix Table D.3, which gives average ratings for the two site visits.

The Enhanced Reading Opportunities Study

Table 3.4

Dimensions and Component Constructs of Implementation Fidelity, by ERO Program

Dimension	Component	Reading Apprenticeship	Xtreme Reading
Learning Environment	**General Instructional Constructs**	Classroom climate On-task participation	Classroom climate On-task participation
	Program-Specific Constructs	Social reading community	Classroom management Motivation and engagement
Comprehension Instruction	**General Instructional Constructs**	Comprehension Metacognition	Comprehension Metacognition
	Program-Specific Constructs	Metacognitive conversations Silent sustained reading Content/theme integration Writing Integration of curriculum strands	Curriculum-driven (systematic) instruction • Structured content • Research-based methodology • Connected, scaffolded, informed instruction Needs-driven (responsive) instruction • Student accommodations • Feedback to students

The Enhanced Reading Opportunities Study

Table 3.5a

Number of ERO Classrooms with Well-, Moderately, or Poorly Aligned Implementation to Program Models on Each Implementation Dimension, by ERO Program — Year 2 Fall

Implementation Dimension	All Schools	Reading Apprenticeship Schools	Xtreme Reading Schools
Learning environment			
Well-aligned implementation (composite rating is 2.0 or higher)	30	14	16
Composite rating is 2.5 or higher (very well-aligned implementation)	20	11	9
Composite rating is 2.0 - 2.4	10	3	7
Moderately aligned implementation (composite rating is 1.5-1.9)	3	3	0
Poorly aligned implementation (composite rating is less than 1.5)	1	0	1
Comprehension instruction			
Well-aligned implementation (composite rating is 2.0 or higher)	20	11	9
Composite rating is 2.5 or higher (very well-aligned implementation)	10	4	6
Composite rating is 2.0 - 2.4	10	7	3
Moderately aligned implementation (composite rating is 1.5-1.9)	11	5	6
Poorly aligned implementation (composite rating is less than 1.5)	3	1	2
Combined dimensions			
Well-aligned implementation on both dimensions	20	11	9
Moderately aligned implementation on at least one dimension and moderately or well-aligned implementation on the other dimension	11	5	6
Poorly aligned implementation on at least one dimension	3	1	2
Sample size	34	17	17

(continued)

Table 3.5a (continued)

SOURCES: MDRC and AIR calculations from classroom observation data.

NOTES: Implementation with a composite score of less than 1.5 for a given dimension was deemed to be at the beginning stages of development. The implementation for these dimensions was designated as poorly aligned with the program models.

Implementation with composite scores between 1.5 and 1.9 for a given dimension exhibited at least moderate development in some areas while being at the beginning stages of development in other areas. The implementation for these dimensions was designated as moderately aligned.

Implementation with scores of 2.0 or higher for a given dimension exhibited well-developed fidelity on several areas and at least moderate development in most other areas. The implementation for these dimensions was designated as well aligned.

Implementation with scores of 2.5 or higher for a given dimension exhibited well-developed fidelity on most areas and at least moderate development on most other areas. The implementation for these dimensions was designated as very well aligned.

There were 7 schools rated as very well aligned on both program dimensions in the fall site visit, of which 2 were Reading Apprenticeship schools and 5 were Xtreme Reading schools.

on the classroom learning environment and comprehension instruction dimensions fell into the well-aligned, moderately aligned, and poorly aligned categories of fidelity. The bottom panel of the tables categorizes schools in terms of their overall implementation fidelity, based on their ratings across both implementation dimensions after each site visit. The discussion that follows focuses first on each implementation dimension and then turns to overall fidelity, which accounts for the importance of the implementation of both dimensions to the ERO programs.

Fidelity, by Implementation Dimension

As described above in the chapter, one focus of both RAAL and Xtreme Reading is the learning environment in the classroom. This involves setting expectations for the organization of the classroom, for how students should interact with the teacher and with their peers, and for the daily and weekly schedules of classroom activities. These same expectations are reinforced in each of the subsequent curriculum units. Table 3.5a shows that, based on the ratings from the observations conducted during the fall site visit, the ERO programs in 30 of the 34 high schools reached a level of implementation that was well aligned with the program models in terms of classroom learning environment dimension. The programs at three schools were rated as demonstrating moderate alignment on this dimension, and the program at one school was rated as demonstrating poor alignment. Table 3.5b shows that, by the spring of the second year of implementation, 31 of the schools had implemented programs at a level of well-aligned fidelity on the learning environment dimension; two schools had programs that were considered moderately aligned; and one school's program was rated as poorly aligned on this dimension.

The Enhanced Reading Opportunities Study

Table 3.5b

Number of ERO Classrooms with Well-, Moderately, or Poorly Aligned Implementation to Program Models on Each Implementation Dimension, by ERO Program — Year 2 Spring

Implementation Dimension	All Schools	Reading Apprenticeship Schools	Xtreme Reading Schools
Learning environment			
Well-aligned implementation (composite rating is 2.0 or higher)	31	17	14
Composite rating is 2.5 or higher (very well-aligned implementation)	21	13	8
Composite rating is 2.0 - 2.4	10	4	6
Moderately aligned implementation (composite rating is 1.5-1.9)	2	0	2
Poorly aligned implementation (composite rating is less than 1.5)	1	0	1
Comprehension instruction			
Well-aligned implementation (composite rating is 2.0 or higher)	28	13	15
Composite rating is 2.5 or higher (very well-aligned implementation)	17	8	9
Composite rating is 2.0 - 2.4	11	5	6
Moderately aligned implementation (composite rating is 1.5-1.9)	6	4	2
Poorly aligned implementation (composite rating is less than 1.5)	0	0	0
Combined dimensions			
Well-aligned implementation on both dimensions	26	13	13
Moderately aligned implementation on at least one dimension and moderately or well-aligned implementation on the other dimension	7	4	3
Poorly aligned implementation on at least one dimension	1	0	1
Sample size	34	17	17

(continued)

Table 3.5b (continued)

SOURCES: MDRC and AIR calculations from classroom observation data.

NOTES: Implementation with a composite score of less than 1.5 for a given dimension was deemed to be at the beginning stages of development. The implementation for these dimensions was designated as poorly aligned with the program models.

Implementation with composite scores between 1.5 and 1.9 for a given dimension exhibited at least moderate development in some areas while being at the begninning stages of development in other areas. The implementation for these dimensions was designated as moderately aligned.

Implementation with scores of 2.0 or higher for a given dimension exhibited well-developed fidelity on several areas and at least moderate development in most other areas. The implementation for these dimensions was designated as well aligned.

Implementation with scores of 2.5 or higher for a given dimension exhibited well-developed fidelity on most areas and at least moderate development on most other areas. The implementation for these dimensions was designated as very well aligned.

There were 13 schools rated as very well aligned on both program dimensions in the spring site visit, of which 7 were Reading Apprenticeship schools and 6 were Xtreme Reading schools.

The other focus of the two programs is providing skills and strategies to enhance reading comprehension. As shown in Table 3.5a, program implementation was rated as well aligned on the comprehension instruction dimension for the ERO programs in 20 of the schools, based on observations conducted at the second site visit. Eleven schools had programs that demonstrated moderate alignment, and three schools had programs that demonstrated poor alignment in terms of comprehension instruction. Table 3.5b shows that, by the spring site visit in the second year, program implementation on the comprehension instruction dimension was well aligned at 28 of the schools and moderately aligned at the remaining six schools.

As Tables 3.5a and 3.5b indicate, at both site visits in the second year of implementation, more than half of the schools were considered well aligned to the program models on both the learning environment dimension (88 percent of schools in the fall and 91 percent of schools in the spring) and the comprehension instruction dimension (59 percent of schools in the fall and 82 percent of schools in the spring). The numbers of schools whose programs were well aligned on the two implementation dimensions provided an opportunity for the study team to break apart the "well-aligned" category further and identify a subgroup of "very well-aligned" schools.[67] Schools where program implementation was rated as very well aligned have compo-

[67]During the first year of implementation, there were a few notable challenges to program implementation (discussed in the final section of this chapter), and more schools were rated as poorly aligned to the implementation models outlined by the program developers. As a result, the first-year report focused on understanding problematic implementation. In the second year of the study, the implementation challenges observed in the first year were diminished. There was a shift from having 10 schools with poorly aligned implementation in the first year to having one school with poorly aligned implementation in the second year.

site ratings of 2.5 or higher on either program dimension, suggesting implementation that is closely representative of the expectations of the program developers. Twenty schools had reached a level of very well-aligned implementation on the learning environment dimension by the fall site visit in the second year. By the spring, this number had increased to 21. For the comprehension instruction dimension, 10 schools had achieved very well-aligned implementation at the fall site visit; by the spring site visit, 17 schools were rated as very well aligned on this dimension.

Differences in Fidelity, by Implementation Dimension

The pattern of findings shown in Tables 3.5a and 3.5b indicate that, at both site visits, more programs reached a level of well-aligned implementation fidelity on the learning environment dimension (30 schools in the fall, 31 schools in the spring) than on the comprehension instruction dimension (20 schools in the fall, 28 schools in the spring). Two hypotheses — generated by the study team and based in part on the team's understanding of the programs as outlined by the developers — offer potential explanations for this observed difference in the fidelity achieved by schools on these two dimensions. While presented separately, these hypotheses should not be considered as mutually exclusive:

- First, this difference may reflect how these programs evolve during their implementation. The continuous and mutually reinforcing way that the elements of the classroom learning environment dimension are situated in the curriculum presents ongoing opportunities for teachers to refine their implementation of this dimension's elements and reach alignment with the program model. Comprehension instruction elements evolve over the course of the year and vary across curriculum units; instructional strategies focusing on metacognition and content are part of all the curriculum units, but new strategies get incorporated by teachers throughout the year. Thus, teachers may not have the same continuous opportunity to refine their implementation of each instructional element.

- Second, the difference in fidelity achieved on the two implementation dimensions might be explained by the difference in teachers' experience with teaching reading as opposed to developing a positive classroom environment. The teachers in the study classrooms had been trained as content-area teachers (primarily English and social studies), and so the procedures for teaching reading and reading comprehension strategies of the programs were in some respects new to the teachers at the beginning of the study, particularly since they had not been reading teachers previously. However, the principles behind the learning environment dimension of the program models reflect prin-

ciples often advocated for classrooms across subject areas, such as respect between individuals and creating a safe space for sharing opinions and ideas.

The program developers emphasize the importance of both program dimensions, but it is useful for policymakers and practitioners to be aware that, in the implementation of these programs or similar ones, different aspects of the programs may develop more quickly than others. The learning environment dimension may develop more quickly than the comprehension instruction dimension, given that high school teachers have familiarity with the teaching of adolescents and with managing adolescent classrooms. The comprehension instruction dimension may take longer to align with the program models because the classes most likely have students at different reading levels. Furthermore, the comprehension instruction dimension may take longer to develop, as teachers are likely new to teaching reading comprehension as defined by RAAL or Xtreme Reading.

Based on these fall and spring observation ratings, the study team found that, during the fall site visit of the second year, most of the sites had already established respectful learning environments — characterized by the open exchange of ideas between the students and the teacher in a safe and nurturing environment — as measured using scales that reflected the programs' definitions of positive teacher-student interactions.[68] The classrooms were also "print-rich," in that they had ample supplies of books and other supporting materials, as measured by program-specific environmental checklists.[69] During the spring site visit, field observers found that overall the high schools improved their program implementation compared to the fall — with eight more schools rated as having programs well aligned to the program models in terms of comprehension instruction and one more school rated as having a well-aligned program in terms of learning environment.

Rating the Overall Fidelity of ERO Program Implementation

The bottom panels of Tables 3.5a and 3.5b cluster schools based on their levels of implementation fidelity across both the classroom learning environment and the comprehension instruction dimensions. Because the classroom learning environments and comprehension instruction activities were designed to be interdependent and mutually reinforcing, the implementation of the ERO program in a given school was deemed to be well aligned with the program model overall only if both of these dimensions were rated as well aligned. In the fall of the second year of implementation, the ERO programs in 20 of the 34 schools were found to have

[68]Please see Appendix D for details about the specific measurement scales.

[69]As a part of the learning environment dimension, the implementation-fidelity scales used during site visit observations contain a general construct that measures classroom climate and social support for learning. These scales reflect, among other things, the level of respect in ERO classrooms. Refer to Appendix D for the observation scales.

reached the well-aligned level of implementation on both the classroom learning environment and the comprehension instruction dimension. Of these 20 schools, 7 were considered very well aligned to the program models. By the spring, 26 programs reached the level of well aligned on both dimensions, and 13 of these 26 schools had rated as very well aligned to the models. The schools rated as very well aligned on both implementation dimensions represent places where the ERO programs as implemented came closest to matching the models as intended by the developers. While there is variation among these schools, the assessment of their implementation fidelity revealed that, after the fall and spring site visits, schools considered to be well aligned to program models were rated as poorly aligned on no more than one of the general instruction and program-specific constructs. These 26 schools include 13 RAAL schools and 13 Xtreme Reading schools. For the 13 schools that were considered very well aligned to the program models at the spring site visit, seven were RAAL schools, and six were Xtreme reading schools.

After the fall site visit, the implementation of the ERO program was rated as moderately aligned with the program model for at least one of the two key program dimensions in 11 high schools. In these schools, neither of the dimensions was rated as poorly aligned. By the spring site visit, this number had decreased: seven schools had programs whose implementation fidelity ranked as moderately aligned on one dimension and moderately or well aligned on the other dimension. Schools identified as having especially problematic program implementation were those schools whose average fidelity rating on either the classroom learning environment dimension or the comprehension instruction dimension was classified as implementation poorly aligned to the program models. The bottom panel of Table 3.5a also shows that, at the fall site visit, three of the 34 high schools were found to have encountered serious implementation problems on at least one of the two key program dimensions during the second year of the study. Of these, one RAAL program and two Xtreme Reading programs had poorly aligned comprehension instruction scores. One of the two Xtreme Reading programs also was rated as poorly aligned on the learning environment dimension in the fall. The bottom panel of Table 3.5b indicates that, by the spring site visit, one Xtreme Reading program had a poorly aligned learning environment. This program was well aligned on the comprehension instruction dimension.

As mentioned, there were nine teachers who replaced teachers during or after the first year of the study, seven of whom replaced teachers after the conclusion of the first year. These nine replacement teachers fared comparably — having similar average implementation-fidelity scores for both dimensions — to those returning teachers who taught all of the first year based on their implementation rating scores for the second year (an average composite rating of 2.4 for returning teachers and 2.3 for replacement teachers at the spring site visit). Refer to Appendix D for a comparison of the second-year implementation ratings for the replacement teachers and the teachers who had taught all of the first year.

Comparison of Year 1 and Year 2

During the first year of the implementation of the ERO programs, three challenges to implementation arose that were addressed systematically during the second year of the study, thereby providing some explanation for the difference in implementation rankings between the two years of the study. One challenge to implementation during the first year was the delayed start (an average of six weeks after the first day of school) of ERO classes in all of the schools. As is discussed in Chapter 4, in the second year of the study, ERO classes started an average of two weeks into the school year, and, in 18 of the schools, ERO classes started on the first day of school. The earlier start in the second year in most study schools reduced disruptions in students' class schedules and allowed the teachers a longer amount of time to cover the year's worth of anticipated curricula.

A second challenge during the first year of implementation was the delayed acquisition of some prescribed program materials and resources.[70] In the second year, all but one teacher had the program materials needed (having acquired them in Year 1) to begin teaching on the first day of school. In cases where teachers needed to replenish materials from the first year, the study team worked with the school administrative staff to acquire them.

The third challenge to implementation during the first year was the newness of the programs to the schools and the ERO teachers. For the 25 teachers who taught all of the first year of implementation and continued to teach the second year, the programs were no longer so new to them, as they had taught the full curriculum once. Additionally, during the first year, two of the teachers left during the middle of the school year, adding another challenge to implementation, as the schools had to find and train replacements while classes were in session.[71] In the second year of implementation, all of the teachers taught for the entire year. Thus, there was no challenge presented by having to find and train replacement teachers in the middle of the school year.

In comparing overall implementation across both program dimensions (learning environment and comprehension instruction) in the spring of Year 2 with overall implementation in the spring of Year 1, more schools were rated as well aligned to the program models, and fewer

[70]During Year 1, the study team visited the ERO schools approximately four weeks after the start of classes to ensure that the rooms were equipped with the materials required by the program developers (a library, a file cabinet, a flip chart, an overhead projector, two computers, and a printer/scanner). These materials were to be purchased by the district using Smaller Learning Communities (SLC) grant funds, discussed in Chapter 2. The study team found that one or more of these classroom components were missing in 23 of the 34 schools, and they communicated with district grant coordinators about securing these items for the classrooms.

[71]Training for midyear replacement teachers was arranged and coordinated by the study team, the program developers, and the school districts.

schools were rated as poorly aligned to the program models.[72] Based on ratings from the spring site visit of the second year of implementation, the ERO programs in 26 of the 34 schools reached an overall level of implementation that was well aligned with the program models. This is an improvement over the first year of the study, when 16 of the 34 schools were rated as well aligned with the program models. Also, during the spring site visit of the second year, implementation in only one school was found to be problematic on either dimension. This is lower than what was found in the first year, when 10 schools ranked as poorly aligned on at least one of the two key program dimensions.

Figure 3.2 presents the schools' composite scores (using a scale from 1 to 3) for the learning environment dimension over the course of the three site visits (one occurring in the spring of Year 1 of the study and two occurring in Year 2 — in the fall and in the spring). Disaggregating the three site visits and presenting them using average composite scores provides a sense of the change that occurred over time. Figure 3.2 shows that overall learning environment scores for both programs increased from the spring of the first year to the spring of the second year. In the spring of the first year of implementation, 26 schools were rated as well aligned on the learning environment dimension. By the spring of the second year, the number of schools considered well aligned to the program models increased to 31.

Figure 3.3 illustrates that average fidelity ratings for composite comprehension instruction at both the RAAL and the Xtreme Reading schools increased from the spring of the first year to the spring of the second year, as was the case with the learning environment dimension. At the spring site visit during the first year of implementation, 16 schools were considered as having well-aligned fidelity of implementation for the comprehension instruction dimension. The number of schools that had well-aligned implementation for comprehension instruction increased by the spring of the second year; 28 schools were rated as well aligned by the second-year spring site visit.

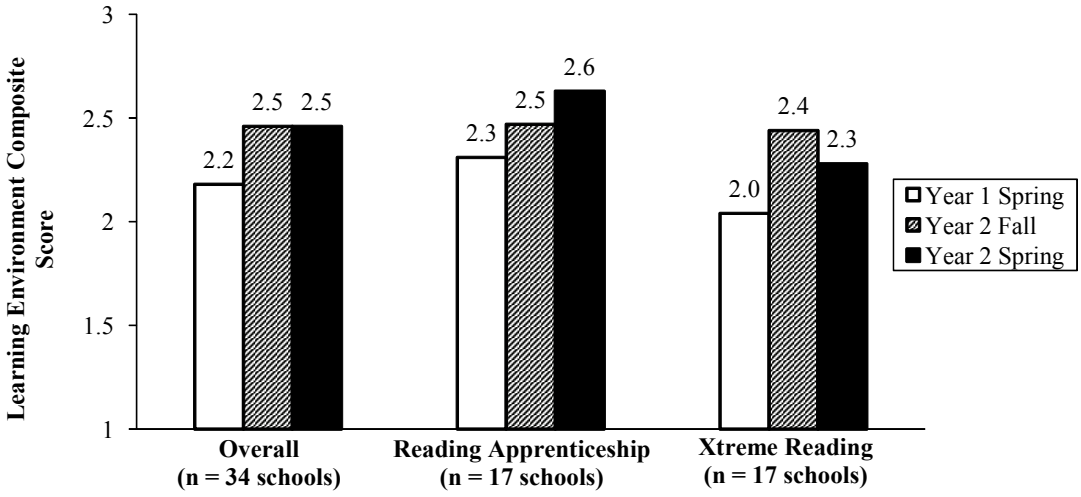
Figure 3.4 plots the composite fidelity scores (averages of the learning environment and comprehension instruction ratings) across the two spring site visits from the two years of the study. The diagonal line indicates no change in implementation fidelity from Year 1 to Year 2. Implementation improved from Year 1 to Year 2 in schools above the diagonal line, and it declined in schools below the diagonal line. The figure illustrates that these overall fidelity scores in the second year were the same as or higher than the first-year scores for 28 of the 34 schools (that is, the scores are on or above the diagonal line). Figure 3.4 also shows that while there were nine replacement teachers at the beginning of the second year of the study, only one of the schools where there was a replacement teacher did not have a higher rating in the second year than in the

[72]For the comparisons of Years 1 and 2, implementation data were gathered from the two spring site visits in order to examine data from the same time frame in both years of the study.

The Enhanced Reading Opportunities Study

Figure 3.2

Learning Environment Composite Scores, by ERO Program



SOURCES: MDRC and AIR calculations from classroom observation data.

NOTES: Implementation with a composite score of less than 1.5 for a given dimension was deemed to be at the beginning stages of development. The implementation for these dimensions was designated as poorly aligned with the program models.

Implementation with composite scores between 1.5 and 1.9 for a given dimension exhibited at least moderate development in some areas while being at the beginning stages of development in other areas. The implementation for these dimensions was designated as moderately aligned.

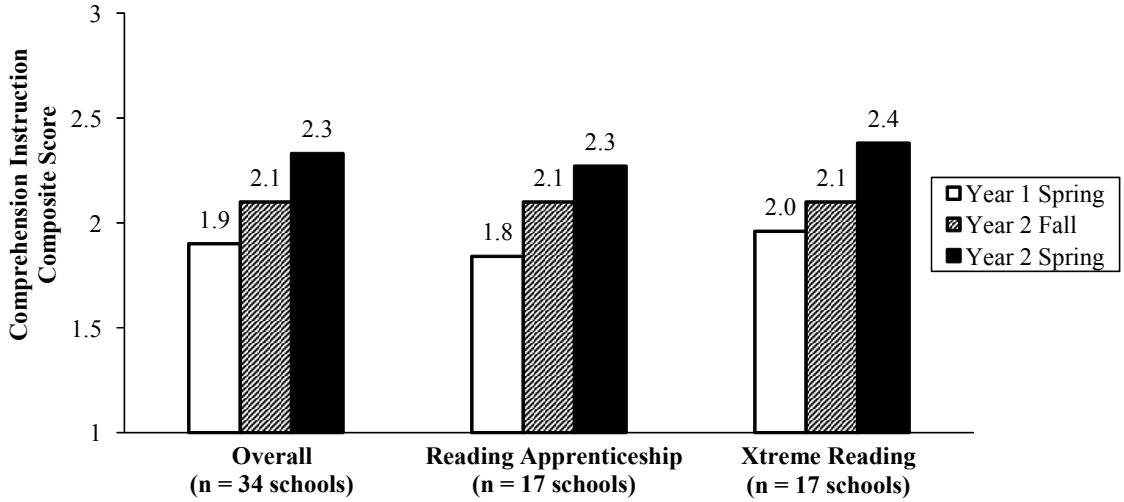
Implementation with scores of 2.0 or higher for a given dimension exhibited well-developed fidelity on several areas and at least moderate development in most other areas. The implementation for these dimensions was designated as well aligned.

Implementation with scores of 2.5 or higher for a given dimension exhibited well-developed fidelity on most areas and at least moderate development on most other areas. The implementation for these dimensions was designated as very well aligned.

The Enhanced Reading Opportunities Study

Figure 3.3

Comprehension Instruction Composite Scores, by ERO Program



SOURCES: MDRC and AIR calculations from classroom observation data.

NOTES: Implementation with a composite score of less than 1.5 for a given dimension was deemed to be at the beginning stages of development. The implementation for these dimensions was designated as poorly aligned with the program models.

Implementation with composite scores between 1.5 and 1.9 for a given dimension exhibited at least moderate development in some areas while being at the beginning stages of development in other areas. The implementation for these dimensions was designated as moderately aligned.

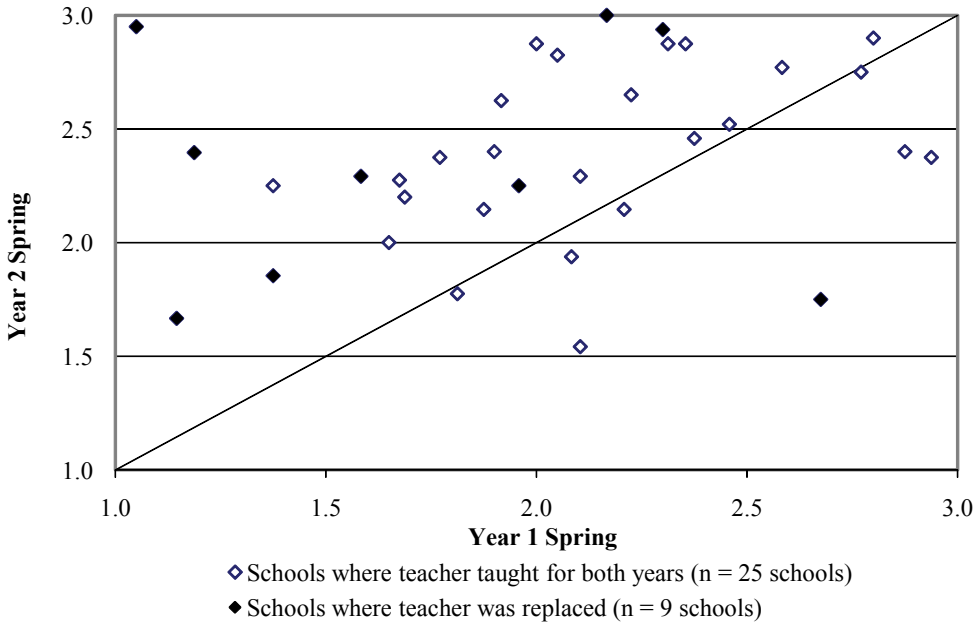
Implementation with scores of 2.0 or higher for a given dimension exhibited well-developed fidelity on several areas and at least moderate development in most other areas. The implementation for these dimensions was designated as well aligned.

Implementation with scores of 2.5 or higher for a given dimension exhibited well-developed fidelity on most areas and at least moderate development on most other areas. The implementation for these dimensions was designated as very well aligned.

The Enhanced Reading Opportunities Study

Figure 3.4

Composite Fidelity Scores, by Site Visit



SOURCES: MDRC and AIR calculations from classroom observation data.

NOTES: Implementation with a composite score of less than 1.5 for a given dimension was deemed to be at the beginning stages of development. The implementation for these dimensions was designated as poorly aligned with the program models.

Implementation with composite scores between 1.5 and 1.9 for a given dimension exhibited at least moderate development in some areas while being at the beginning stages of development in other areas. The implementation for these dimensions was designated as moderately aligned.

Implementation with scores of 2.0 or higher for a given dimension exhibited well-developed fidelity on several areas and at least moderate development in most other areas. The implementation for these dimensions was designated as well aligned.

Implementation with scores of 2.5 or higher for a given dimension exhibited well-developed fidelity on most areas and at least moderate development on most other areas. The implementation for these dimensions was designated as very well aligned.

The diagonal line represents points where the Year 1 fidelity score is equal to the Year 2 fidelity score. For schools located above the diagonal line, the Year 2 fidelity score is greater than the Year 1fidelity score. For schools located below the diagonal line, the Year 2 fidelity score is lower than the Year 1 fidelity score.

first year (its score is below the diagonal line). In comparison with the first year of implementation, all other schools where there were replacement teachers had higher ratings in the second year on both the learning environment and the comprehension instruction dimension. Based on the classroom observations conducted during the spring site visit in the second year, the nine programs taught by replacement teachers had average scores of 2.4 for learning environment and 2.3 for comprehension instruction. In the first year, the ratings based on the spring observations for these programs were 1.7 for learning environment and 1.8 for comprehension instruction.[73] Thus, schools where the teachers were replaced had higher implementation-fidelity ratings, on average, in the second year, with the replacement teachers teaching the ERO programs.

Figure 3.4 also shows that at 19 of the 25 schools where the ERO teachers returned after having taught all of the first year, the average implementation composite ratings were the same as or higher than in the second year. The average learning environment rating at these 25 schools was 2.5 in the second year of implementation, compared with 2.4 in the first year; their average comprehension instruction rating was 2.3 in the second year, compared with 2.0 in the first year. In addition, the implementation of the ERO programs in terms of both the learning environment and the comprehension instruction dimension was well aligned at more of these 25 schools in the second year than in the first.[74] Thus, implementation-fidelity ratings were higher in the second year of implementation of the study than in the first year, both at schools where replacement teachers taught the programs in the second year and at schools where the programs were taught for a second full year by the same teacher. Also, in the second year of implementation, more schools than in the first year achieved ratings of "well aligned" on both key program dimensions, and fewer received "poorly aligned" ratings.

[73]One of the nine replacement teachers was also observed during the spring of Year 1 of the study. This is due to the replacement of one teacher prior to the site visit that occurred during the spring of Year 1.

[74]Appendix Tables D.6 and D.7 show the distribution of these 25 high schools across rating categories, by implementation dimension, for each year.

Chapter 4

Student Attendance in the ERO Classes, Course Enrollment, and Participation in Literacy Support Activities

In addition to examining the fidelity with which the sites participating in the Enhanced Reading Opportunities (ERO) study implemented the models of the two supplemental high school literacy programs — Reading Apprenticeship Academic Literacy (RAAL) and Xtreme Reading — the evaluation also includes an assessment of the extent to which the experience with literacy instruction of the students in the ERO group was substantially different from that of the students in the non-ERO group. To answer this question, the ERO study team measured how many hours of the programs the ERO students actually received, compared ERO and non-ERO group students' class schedules, and analyzed ERO and non-ERO group students' self-reported participation in other literacy support services either in or outside school.

Based on the ERO programs' theory of action, the impact is hypothesized to be a function, in part, of how much exposure the ERO students have to the ERO classes throughout the school year. To clarify whether or not ERO students were actually exposed to the entire program, the evaluation team collected data about the frequency with which the ERO classes met and about whether and how often students attended these courses. These data provide an indication of the overall "dosage" of the ERO interventions that students in the ERO group received.

The ERO team also hypothesized that the impact of the ERO program is a function of the contrast between the ERO dosage and the amount of literacy support available to non-ERO students. Impacts of the ERO program could be falsely minimized if the ERO course replaced another literacy-focused course taken by the majority of non-ERO students. On the other hand, impacts could be falsely enhanced if ERO group students were taking another literacy-focused course beyond ERO that non-ERO students were not taking. Even if both ERO and non-ERO students are taking similar courses outside the ERO programs, if all students are taking literacy-rich English/language arts (ELA) courses or supplemental literacy courses, the impacts for the program could be minimized because the ERO class would not be adding much value to an already literacy-rich environment.

Several types of data were collected and analyzed to measure the contrast in the literacy support available and received by ERO and non-ERO group students. The data on ERO class attendance provided an indication of whether students in the non-ERO group inadvertently enrolled in the ERO classes and thus diluted the overall contrast in literacy services received by students in the ERO and non-ERO groups. Student schedule data were used to assess

course-taking patterns of students participating in the study, providing an understanding of how the ERO courses fit into student schedules and an opportunity to explore potential contrasts in the enrollment of ERO students and non-ERO students in different types of core content and noncore content courses, including electives. ELA teachers were interviewed to investigate literacy instruction available to students through their ELA classes. Students were also surveyed about the frequency with which they participated in classes or tutoring services that aimed to improve reading and writing skills. Specifically, the student follow-up survey asked several questions about the frequency and duration with which students participated in such activities either in school or outside school. These data are available for students in both the ERO and the non-ERO group and are intended to capture participation in both the ERO classes and other literacy support programs and services. They provide a measure of the difference in exposure to supplemental literacy support services between the ERO and non-ERO groups — which is a key factor in whether the ERO programs offer a contrast to the services that would otherwise be available.

To help the reader place the second-year findings in context, a brief description of the first-year findings is provided. The ERO study team investigated these questions about the amount of and contrast in literacy services in the first year of implementation and found that ERO group students did receive substantially more literacy support services during their ninth-grade year than non-ERO group students, but the ERO students received fewer hours of treatment than was originally planned, due to implementation issues. During the first year of implementation, most of the recruitment was done at the beginning of the students' ninth-grade year, causing ERO classes to begin an average of six weeks after the start of the school year. The program was intended to run for nine months, but no schools were able to offer the full dosage to students, and, on average, the programs ran for just over seven and half months. At the same time, there was little student attrition between recruitment and the start of the ERO programs: 95 percent of the ERO group students enrolled in the program, and 91 percent were attending at the end of the first year. An analysis of student class schedules suggests that the ERO programs tended to replace elective courses that were not literacy focused, and the students randomly assigned to the study's ERO group during the first year of implementation reported a much higher frequency of participation in supplemental literacy services than non-ERO group students.[75]

This chapter first describes these same analyses and their findings for the second year of implementation of the ERO programs and ends with a comparison of the first and second years of the program. The chapter discusses the following key findings for the second year of implementation:

[75]See Kemple et al. (2008), pp. 68-70.

- The ERO classes began an average of just over two weeks after the start of the school year and operated for an average of just over nine months of the academic year.
- Among the students in the ERO group, 91 percent attended at least one ERO class, and 87 percent were still attending at the end of the school year. On average, students attended 10.8 of the 13.6 hours of ERO classes offered per month.
- ERO and non-ERO students were equally represented in English/language arts courses and were enrolled in the same variety of noncore content courses outside the ERO programs.
- Students who were randomly assigned to the study's ERO group reported a higher frequency of participation in supplemental literacy services than students in the non-ERO group.

This chapter also compares the two years of program implementation. It discusses the similarities and differences between the two cohorts of students in terms of exposure to literacy support services. The main differences between the two years can be found in the duration and enrollment of the ERO classes:

- In the second year, the ERO courses were in session for an average of one and a half months longer than the first year, but a smaller percentage of students ever attended an ERO class. Overall, the students in the second year of implementation attended more hours of ERO classes than the students in the first year.
- ERO group students in both years of implementation reported a much higher level of literacy support than non-ERO group students, and, in both years, the main difference in students' schedules for ERO and non-ERO group students was that, for most students, ERO replaced a non-literacy-focused elective class.

Student Enrollment and Attendance in the ERO Classes

The amount of ERO instruction that students receive is a function of program duration and student attendance. The longer the duration of the program, the greater the opportunity students have to participate in the ERO classes. The more often students attend, the more ERO instruction they will be exposed to. Following is an overview of findings from the analysis of program duration and attendance during the second year of the study.

Program Duration

The ERO programs were designed to operate for the full school year and to provide students with approximately nine months of supplemental literacy instruction. The ERO classes began an average of 2.3 weeks after the start of the 2006-2007 school year. Eighteen schools started the ERO program on the first day of school, and five more schools started within the first two weeks classes were in session. The remaining eleven schools started between four and just under 12 weeks into the school year. These eleven schools started their ERO programs an average of seven weeks after the start of the school year. Before an ERO class could begin, students needed to be tested, recruited for the program, randomly assigned to the program or control group, and scheduled into the course. Schools that were unable to begin the program on or near the first day of school met with challenges or delays related to one or more of these necessary preconditions for starting the class.[76] Despite these delays in the start-up of some classes, the ERO programs operated for an average of just over nine months.[77]

Student Enrollment and Attendance

As part of their responsibilities to the project, the ERO teachers were required to maintain and report to the study team daily attendance records for all students randomly assigned to the ERO group. These data, along with information about the length of ERO class periods, provided the basis for calculating several measures of ERO enrollment and attendance. These measures are displayed in Table 4.1.[78]

Among students in the ERO group, 91 percent attended at least one ERO class during the year, and 87 percent were still attending ERO classes at the end of the school year. On

[76]Schools also needed a teacher, a classroom, and supplies to begin the ERO classes on time. In interviews with the study team, 33 out of the 34 teachers reported having classrooms and supplies at the start of the year, allowing classes to start as soon as students were randomly assigned to the programs.

[77]This calculation is based on the length of the school year from start date to end date. Because it reflects how long the programs ran in months across the full span of the school calendar, it includes vacations and holidays. The number of months that schools were in session varies by district, with 10 months being the longest period.

[78]The findings presented in Table 4.1 are based on attendance data for ERO group students in the follow-up respondent sample — the same sample as is used in the impact analysis for this report. The ERO enrollment and attendance findings for these students provide an assessment of the dosage of ERO program services that is associated with the impact findings discussed in Chapter 5. Note that all measures in Table 4.1 include the 110 students from the ERO group who never attendedthe ERO classes. All measures also include students who left the program during the school year. Zero values were included for the attendance of these students during the periods when they were not enrolled in the programs.

The Enhanced Reading Opportunities Study

Table 4.1

Attendance in ERO Classes, Cohort 2 Follow-Up Respondent Sample in the ERO Group

Characteristic	All Schools	Reading Apprenticeship Schools	Xtreme Reading Schools
Ever attended an ERO class during the year (%)	91.3	92.1	90.5
Attending ERO classes at the end of the year (%)	86.9	88.2	85.6
Average daily attendance rate in ERO classes per month[a] (%)	79.3	80.0	78.6
Median student value	*91.8*	*91.7*	*92.0*
Average number of months ERO program was in operation	9.1	9.1	9.0
Median student value	*9.5*	*9.5*	*9.5*
Average number of months attending ERO classes	7.9	8.1	7.8
Median student value	*9.0*	*9.3*	*9.0*
Average number of hours ERO class met per month	13.6	13.8	13.5
Median student value	*13.5*	*13.5*	*13.5*
Average number of hours student attended ERO class per month	10.8	11.1	10.6
Median student value	*11.8*	*12.1*	*11.6*
Sample size	1,264	645	619

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study monthly attendance data.

NOTE:
[a]There were 110 students who never attended an ERO class, 51 students from Reading Apprenticeship schools and 59 students from Xtreme Reading schools. When these students are excluded, the average daily attendance rate for the remaining students who attended at least 1 ERO class is 86.9 percent for all schools, 86.9 percent for Reading Apprenticeship schools, and 86.9 percent for Xtreme Reading schools.

average, ERO group students attended 79 percent of the scheduled ERO classes each month and attended ERO classes for 7.9 months during the school year. The ERO programs were designed for an average of 3 hours and 45 minutes of class time per week (which is typically scheduled either as 45-minute classes each day or as 80- to 90-minute classes every other day). With an average of 20 days of school per month, the ERO classes were designed to provide students with approximately 15 hours of supplemental literacy instruction per month. Based on the attendance data provided by the ERO teachers, Table 4.1 shows that the ERO classes met for an average of 13.6 hours per month (nearly three and a half hours per week), with students attending an average of 10.8 hours of ERO instruction per month, or just under 2 hours and 45 minutes per week.

In the follow-up respondent sample, 110 ERO group students never attended an ERO class. The median values included in Table 4.1 help to put into perspective the effect that these missing students had on the average attendance rates also included in the table. Although the average daily attendance rate for ERO students in ERO classes is 79 percent, the median student attendance rate is 92 percent, indicating that half of the ERO students attended 92 percent or more of their ERO classes. Similarly, the average number of months students attended ERO classes is 7.9 months, but the median number of months students attended classes is 9.0 months.

Table 4.1 also shows that the number of hours of ERO classes attended by students was similar across the two literacy interventions due to comparable attendance rates and program offerings. RAAL and Xtreme Reading classes met for an average of 13.8 hours and 13.5 hours per month, respectively. Average daily attendance for RAAL was 80 percent, and it was 79 percent for Xtreme Reading. As a result, the average RAAL student attended an ERO class 11.1 hours per month, while the average Xtreme Reading student attended class 10.6 hours per month.[79]

Overall, Table 4.1 presents multiple measures of the dosage of the intervention — the amount of literacy support — that the ERO group students received through the ERO classes. In contrast, the ERO classes were not a source of literacy support for non-ERO students. Among the 907 respondent sample students in the non-ERO group, 16 were enrolled in the ERO classes.

[79]Statistical tests of the difference in enrollment/attendance between the two programs were not conducted, because the data reported in Table 4.1 are based on complete student records. Hence, for this particular group of students and given this allocation of programs to schools, observed differences in average attendance are real differences. Stated otherwise, there is no error in the data presented in Table 4.1 due to the sampling of students or schools, given that these findings are not used to make generalizations or inferences about the enrollment/attendance patterns that would have been observed had different students or schools been assigned to the ERO programs.

Student Participation in Literacy Support Activities

A requirement of the ERO grants from the U.S. Department of Education was that the participating schools would not operate other supplemental literacy programs for ninth-grade students who were not enrolled in English Language Learner or Special Education courses during the evaluation period.[80] This was to ensure that the non-ERO group students would not be participating in literacy interventions similar to RAAL or Xtreme Reading. This is important because the overall contrast in literacy support between the ERO group and the non-ERO group would be reduced to the degree that students in the non-ERO group participated in literacy interventions. Still, students in both the ERO and the non-ERO group enrolled in a variety of other elective and required courses with varying levels of literacy support. Students and parents were also free to seek out other literacy-related services on their own. In some cases, they found other adults in the school to provide tutoring; in other cases, students and their families sought out other classes or tutors outside school.

This section of the chapter examines whether the ERO programs created a contrast in ERO students' course-taking and self-reported participation in literacy support activities compared with students in the study's non-ERO group. The section first reviews the manner in which the ERO classes were inserted into students' course schedules and compares ERO and non-ERO group students' course enrollments. Then the section discusses the degree to which literacy instruction was included in students' English/language arts classes in the participating high schools. ERO and non-ERO group students' responses to survey questions regarding the amount of literacy instruction they received inside and outside school are then compared.

Student Class Schedules

Each ERO program was scheduled as one of the classes in ERO students' schedules. Non-ERO students were free to take a course other than ERO to fill that slot in their class schedules. The goal of this section is to review what classes non-ERO students took in place of ERO, to better understand whether or not non-ERO group students were receiving literacy support comparable to that offered in the ERO programs in these non-ERO classes. This is complicated by the fact that both ERO and non-ERO group students had a set of core classes — English/language arts (ELA), mathematics, history/social studies (history), and science — and noncore classes to take during the school year. In this way, the ERO programs did not replace a specific class in a student's schedule but replaced one of several possible elective courses. For this reason, students' entire schedules are explored to determine whether there were any specific types of courses that non-ERO students were taking in place of ERO.

[80]U.S. Department of Education (2005).

Participating high schools used scheduling models that allowed students to take six (2 schools), seven (7 schools), or eight (25 schools) course credits during the year. Four of these course credits were academic requirements in core content areas: ELA, mathematics, history, and science, leaving three or four slots for elective classes.[81] Even in high schools where one of those slots was filled with another required course like physical education or homeroom, there were still two or three slots open for electives. To demonstrate how ERO fits into student schedules, two examples are presented in Table 4.2. Between them, these examples represent the three most common types of variation in student schedules: the schedule model, the number of course slots within the schedule model, and the number of required courses. First, the two most commonly used schedule models in the 34 high schools were the *traditional bell schedule,* in which each class typically meets daily for 40 to 50 minutes (Example 1); and the *alternating (or A/B) block schedule,* in which each class meets for about 80 to 90 minutes every other day (Example 2).[82] Second, since the modal number of course slots in the schools' schedule models was eight slots and the mean was 7.7, Example 1 reflects a schedule with seven course slots, and Example 2 has eight course slots. Lastly, as noted above, some schools may have included another required course (for example, physical education or health) beyond the four core academic courses. Both examples show ERO and non-ERO group students scheduled in the same amount of required courses (four in Example 1 and five in Example 2) and the ERO class replacing one of the students' elective courses.

The ERO study team compared the class schedules of the ERO and non-ERO students to establish whether or not the ERO programs were truly replacing one of the two or three noncore classes in students' schedules, and not a core course, and whether or not there were any particular types of noncore courses that non-ERO students were more likely to take in place of the ERO programs. As noted in Chapter 2, the experience of the non-ERO students represents the "counterfactual," or what the experience of the ERO students would have been

[81] In some instances, students had the option to take two shorter elective courses to fill one course slot.

[82] Although most schools employed these two schedule models, there were three schools in the second year of implementation that offered a 4 x 4 block schedule: students typically take the same four 90-minute courses every day for an entire term, and then they take four different courses during the second term. One of these schools using such a block schedule implemented the Xtreme Reading class as a one-semester class that met for 90 minutes every day, with some of the ERO students taking the class the first semester and the rest taking it the second semester. The other two schools, from the same district, paired the RAAL class with a yearlong computer lab during which the ERO students used TeenBiz, the online application that is part of the RAAL program. The computer lab was not taught by the ERO teacher, and no students from the non-ERO group were enrolled. The ERO class and the computer lab were alternated every other day throughout the year, similar to the A/B block schedule model, remaining aligned with the expectations of the study and the grant that the ERO classes be full-year classes. (Time spent in this computer lab is not included in Table 4.1.) The study team conducted sensitivity testing of the overall impacts presented in Chapter 5 and determined that they are not sensitive to the inclusion of these schools, either collectively or independently.

The Enhanced Reading Opportunities Study

Table 4.2

Comparison of ERO and Non-ERO Student Schedules

Example 1: Traditional Bell Schedule, Seven Periods, Four Required Courses

Period	ERO Students	Non-ERO Students
1	English/Language Arts	English/Language Arts
2	Math	Math
3	Science	Science
4	Social Studies/History	Social Studies/History
5	***ERO***	***Elective***
6	Elective	Elective
7	Elective	Elective

Example 2: Alternating (A/B) Block Schedule, Eight Periods, Five Required Courses

	ERO Students		Non-ERO Students	
Period	Day A	Day B	Day A	Day B
1	English/Language Arts	Science	English/Language Arts	Science
2	Math	Social Studies/History	Math	Social Studies/History
3	Required course	***ERO***	Required course	***Elective***
4	Elective	Elective	Elective	Elective

NOTE: These are not actual schedules, but they represent two types of schedules in ERO high schools. They are used to demonstrate how ERO fits into student schedules.

had they not been able to take the ERO class. The goal of this analysis was to establish whether or not there was a specific noncore class or type of class being taken by non-ERO students in lieu of ERO, thus representing the counterfactual. Table 4.3 displays the course schedules of the "average" ERO and non-ERO student.[83] The table shows the number of courses per student in each of the four core subject areas as well as seven categories of noncore classes including the ERO programs.[84] Courses per student for each category are calculated by dividing the number of students taking a course in a specific category by the total number of student schedules. This tells, on average, what proportion of class periods ERO and non-ERO students devoted to each category of core and noncore classes.

[83]The table is based on 2,878 student schedules from all 34 high schools — 84 percent of the students in the study sample (the 3,441 students who were randomly assigned to the ERO or non-ERO group).

[84]Noncore courses were categorized into traditional high school departments using the department codes for the classes and the class titles and asking the district or school for a class description when necessary.

The Enhanced Reading Opportunities Study

Table 4.3

Comparison of ERO and Non-ERO Student Course Enrollment

Class Type	ERO Students	Non-ERO Students
Number of core courses per student		
English Language Arts	1.01	1.01
Mathematics	1.01	1.01
History/Social Science	0.95	0.96
Science	0.94	0.97
Total number of core courses	*3.90*	*3.95*
Number of noncore courses per student		
ERO	0.93	0.00
Careeer and Technical	0.46	0.65
Visual and Performing Arts	0.31	0.45
Physical Education and Health	0.71	0.88
Foreign Language	0.36	0.46
Support Services	0.38	0.50
Other[a]	0.79	0.83
Total number of noncore courses	*3.94*	*3.78*
Sample size	1,636	1,242

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study student schedules during the 2006-2007 school-year.

NOTES: These calculations include student schedules for 84 percent of the full study sample.

[a]"Other" includes homeroom and advisory courses that are mandatory in several schools in the study as well as noncore courses that did not fit into the other categories.

Table 4.3 confirms that ERO classes did not replace ELA or other required academic classes (mathematics, history, or science). With very few exceptions, all students, both ERO and non-ERO, were enrolled in one ELA course and one mathematics course.[85] Almost every student, regardless of participation in ERO, also took a history and science course during ninth grade.

With few exceptions, students in the non-ERO group were *not* enrolled in the ERO classes, while there were 0.93 ERO courses taken per ERO student, suggesting that 93 percent of the ERO group students were enrolled in ERO classes. Beyond this difference, students in

[85]These core subject areas have slightly more than one course taken per student because, in some cases, a student was enrolled in a second ELA or math class.

the ERO and non-ERO groups enrolled in a similar variety of noncore classes. As shown in Table 4.3, for each noncore course category, there are fewer courses taken per ERO student than per non-ERO student. For instance, there are 0.46 career and technical courses taken per ERO student and 0.65 career and technical courses taken per non-ERO student. Since ERO filled one slot, students in the ERO group have fewer class slots to devote to classes in each of the other noncore areas, but ERO students still selected classes from all of the different noncore categories. The higher level of participation in noncore classes for non-ERO group students appears consistent with the larger amount of non-ERO elective slots available for non-ERO group students.[86] To further assess whether ERO group students were more likely to be underrepresented in one of these noncore categories over the others, the ERO study team also compared how the ERO students distributed their set of available, non-ERO class slots over the six non-ERO course categories with how the non-ERO group students distributed their larger set of available class slots over these six noncore categories. ERO and non-ERO group students were found to be similarly distributed across all six non-ERO, noncore categories.[87] Since no particular elective category stands out as being "replaced" by ERO in ERO students' schedules, it can be concluded that, in general, ERO students replaced one noncore course slot with an ERO class and that this course slot is not from a particular category of noncore courses.

English/Language Arts Instruction

ELA classes offered a means for some literacy instruction to occur beyond the ERO programs. The analysis of student class schedules found that both ERO and non-ERO students were enrolled in ELA classes. Thus, all study participants had similar opportunities for literacy instruction through ELA classes. For this reason, it is unlikely that ELA courses affected the contrast in literacy instruction dosage between ERO and non-ERO group students. Still, a par-

[86]Several courses with titles and descriptions that suggest the possibility of literacy instruction (such as reading and writing supplemental courses, literature, and journalism) are included in the calculations in Table 4.3 but are spread among several of the categories, including "career and technical," "support services," and "other." When these courses are removed from the different categories and combined to create a separate literacy-focused category, the distributions of ERO and non-ERO group students are consistent with the distributions of other non-ERO noncore courses. Of these types of courses, 0.25 course was taken per ERO student, and 0.36 course was taken per non-ERO student.

[87]This was done by dividing the noncore courses per student (found in Table 4.3) in each non-ERO noncore course category by the total courses per student in all six of these categories. The distributions of ERO students for each of the first five noncore course categories were no more than 2 percentage points different than the distributions of non-ERO students. The percentage distributions of ERO and non-ERO student enrollments, respectively, in each area were 15 percent and 17 percent in Career and Technical, 10 percent and 12 percent in the Arts, 23 percent for both groups in Physical Education and Health, 11 percent and 12 percent in Foreign Language, and 13 percent for both groups in Support Services. The distribution of ERO students (28 percent) in the "other" category was slightly higher than for non-ERO students (24 percent) because this category includes some mandatory homeroom and advisory classes.

ticularly literacy-rich environment could weaken the value added by the ERO programs because all students would already be receiving literacy instruction similar to the ERO programs. To better understand the literacy content of the ELA classes, a sample of 39 teachers who taught study sample students were interviewed about the ELA classes they teach.[88]

Overall, these interviews suggest that ELA programs had some similarities to the ERO programs. For instance, ELA courses were mostly taught by ELA-certified teachers who had received professional development in literacy instruction.[89] Also, all of the teachers said that students completed at least some reading assignments in class, and many offered time for students to read self-selected texts, had a classroom library, and included differentiated instruction — all components of the ERO programs.[90]

Still, the teachers who were interviewed suggested that ELA classes tended to encompass a variety of subject matter in their curricula, including reading fluency and comprehension, vocabulary, grammar, writing, literary analysis and appreciation, and critical thinking. Strengthening basic and higher-order reading skills was a common objective for the ELA teachers who were interviewed, but it tended to be one of many objectives. Only three of the teachers suggested that strengthening reading skills was their only objective. In short, the ELA courses that are taken by both ERO and non-ERO students offer some similar literacy support to the ERO programs using some similar teaching methods, but the interviews with ELA teachers suggest that ELA courses include a variety of subject matter and therefore do not offer reading instruction at the same level of intensity as the ERO programs, in which reading instruction is the primary objective.[91]

[88]These interviews include ELA teachers teaching across all districts and at 27 of the 34 schools. This is not a random sample. The teachers were chosen because their classes were representative of the ninth-grade ELA courses at their schools with the highest enrollments of study sample students.

[89]Of these 39 teachers, 38 teachers (97 percent) were certified in ELA instruction, and 35 of these teachers (90 percent) had experienced some type of literacy professional development in the last two years, although the subject matter and setting of this professional development varied.

[90]Of the 39 teachers, 17 teachers (44 percent) offered time for students to read self-selected texts; 27 teachers (69 percent) had a classroom library; and 22 teachers (56 percent) included differentiated instruction in their classroom activities.

[91]The four high schools participating in the study in one district offered a supplementary English course to students who were struggling. The intent of this course was to provide additional time for these students to master the ELA content. Of the students whose schedules the study team received from this district, 61 percent were enrolled in this course, with similar representation from both the ERO and the non-ERO group (62 percent of the ERO students and 60 percent of the non-ERO students). Thus, this supplementary course served similar proportions of ERO and non-ERO students.

Student Participation in Supplemental Literacy Support Activities

The student follow-up survey included items aimed at determining the amount of extra literacy support that students received during the school year, beyond their regular ELA class. The survey asked about four categories of extra literacy help: classes in school, classes outside school, an adult tutor in school, and an adult tutor outside school. The first category describes such supports as the ERO courses. This item essentially provides an opportunity for ERO students to report on their attendance in the ERO classes and for non-ERO students to report on their participation in literacy support activities that would be most similar to or "competitive" with ERO. The other three categories of activities cover other ways in which students might receive help with their reading and writing skills.

The survey questions asked all students about how long (duration) and how often (frequency) they participated in each of the four categories of activities. For example, a student who attended a "help" session every day for the full school year was projected to have attended approximately 180 sessions (about 20 days per month for nine months, or the typical number of days in a school year). Similarly, a student who reported attending twice per week for a semester was projected to have attended about 36 sessions (eight days per month for about four and a half months).

Table 4.4 provides the average levels of student participation in these four types of supplemental literacy support activities and estimates of the differences in participation between the ERO and non-ERO groups. Overall, ERO group students reported receiving more supplemental literacy services than non-ERO group students. Reflecting their participation in the ERO program, students in the ERO group participated in a school-based literacy class four and a half times more frequently than students in the non-ERO group.[92] It should be noted, however, that students in the non-ERO group did report receiving some exposure to a literacy class in school during the year (16.5 sessions, on average), though only 16 of the non-ERO students across all the high schools ever enrolled in an ERO class. Table 4.4 also shows that students in the ERO group reported higher levels of participation in school-based tutoring sessions (an average of 29.5 sessions, compared with 11.8 sessions for the non-ERO students). These differences are statistically significant. Although the differences between ERO and non-ERO group students in literacy support activities outside school are not statistically significant, students in the ERO-group reported slightly higher participation than students in the non-ERO group (on average, 2.0 more sessions in a class setting and 2.3 sessions with a tutor).

[92]This was computed by dividing the amount of participation in school-based literacy classes reported by ERO group students (75.0 sessions) by the amount of participation in school-based literacy classes reported by non-ERO group students (16.5 sessions).

The Enhanced Reading Opportunities Study

Table 4.4

Participation in Supplemental Literacy Support Activities, Cohort 2 Follow-Up Respondent Sample

Outcome	ERO Group	Non-ERO Group	Impact	Impact Effect Size	P-Value for the Difference
All schools (number of sessions)					
School-based literacy class	75.0	16.5	58.6 *	1.42 *	0.000
School-based adult tutor	29.5	11.8	17.7 *	0.50 *	0.000
Outside-school literacy class	7.1	5.1	2.0	0.09	0.077
Outside-school adult tutor	13.2	10.9	2.3	0.07	0.154
Sample size	1,260	901			
Reading Apprenticeship schools (number of sessions)					
School-based literacy class	79.3	16.8	62.5 *	1.51 *	0.000
School-based adult tutor	29.8	12.2	17.6 *	0.50 *	0.000
Outside-school literacy class	7.2	5.9	1.3	0.06	0.424
Outside-school adult tutor	13.9	9.4	4.5 *	0.13 *	0.038
Sample size	642	466			
Xtreme Reading schools (number of sessions)					
School-based literacy class	70.6	16.3	54.3 *	1.31 *	0.000
School-based adult tutor	29.2	11.3	17.9 *	0.51 *	0.000
Outside-school literacy class	6.9	4.1	2.8	0.12	0.074
Outside-school adult tutor	12.5	12.4	0.1	0.00	0.978
Sample size	618	435			

SOURCE: MDRC calculations from the Enhanced Reading Opportunities follow-up student survey.

NOTES: The student follow-up survey was administered in spring 2007 at the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (school-based class standard deviation = 41.377; school-based tutor standard deviation = 35.253; outside-school class standard deviation = 23.033; outside-school tutor standard deviation = 34.852).

A two-tailed t-test was applied to the impact estimate. The statistical significance level is indicated (*) when the p-value is less than or equal to 5 percent.

For each of the above measures, data are missing for no more than 5.2 percent of the respondents.

Rounding may cause slight discrepancies in calculating sums and differences.

Comparison of Year 1 and Year 2

An important difference between the first and second years of the ERO programs is in the timing of student recruitment and the effect that this timing had on the amount of the programs to which the ERO group students were exposed. During the first year, most schools recruited and randomly assigned students during the first months of the school year, delaying the start of ERO courses. During the second year, many schools were able to recruit students before the school year began, allowing many ERO programs to start on time. The earlier recruitment of students during the second year resulted in both the benefits and the drawbacks expected from this strategy. That is, overall, the ERO programs were able to run for almost the entire year, but 9 percent of the ERO students in the follow-up respondent sample never attended an ERO class. As shown in Table 4.5, ERO classes met for an average of 7.7 months in the first year of operation, compared with 9.1 months in the second year.[93] However, since most of the recruitment in the second year of operation happened before the start of the school year and more students are likely to leave the district in the summer between middle and high school than while classes are in session, more students left the district between random assignment and the start of the program. Among the students in the first cohort, 96 percent ever attended an ERO class, while 91 percent ever attended class in the second year. The larger number of students who never attended an ERO class in the second year of operation negatively affected the average daily and monthly attendance of ERO students. For instance, there is a more sizable gap between the average number of months that the ERO program was in operation and the average number of months that students attended ERO classes in the second year of operation (1.1 months) than in the first year of operation (0.6 month). Still, on average, students in the second cohort received more hours of ERO instruction, attending 13.2 more hours of ERO classes during the year than students in the first cohort.

Levels of student participation in supplemental literacy activities outside the ERO program were similar during the first and second years. In both the first and the second year, ERO students were typically enrolled in the same core courses as non-ERO students and in a similar variety of elective courses, with the exception of the ERO group students' substituting the ERO program for one of these electives. Figure 4.1 compares student participation in literacy support activities for the first and second cohorts of students. In both years, a similar pattern of differences is observed between ERO students' and non-ERO students' reported participation in supplemental literacy support activities. Both cohorts show large and significant impacts on

[93]Table 4.5 is based on data for every student. Statistical tests of difference were not conducted (see footnote 79).

The Enhanced Reading Opportunities Study

Table 4.5

Attendance in ERO Classes,
All Cohorts Follow-Up Respondent Sample in the ERO Group

Characteristic	All Schools		Reading Apprenticeship Schools		Xtreme Reading Schools	
	Cohort 1	Cohort 2	Cohort 1	Cohort 2	Cohort 1	Cohort 2
Ever attended an ERO class during the year (%)	95.5	91.3	94.9	92.1	96.0	90.5
Attending ERO classes at the end of the year (%)	91.2	86.9	91.0	88.2	91.4	85.6
Average daily attendance rate in ERO classes per month (%)	82.7	79.3	81.7	80.0	83.6	78.6
Average number of months ERO program was in operation	7.7	9.1	7.8	9.1	7.7	9.0
Average number of months attending ERO classes	7.1	8.0	7.1	8.1	7.1	7.8
Average number of hours ERO class met per month	13.6	13.6	13.5	13.8	13.7	13.5
Average number of hours student attended ERO class per month	11.3	10.8	11.2	11.1	11.5	10.6
Average number of hours student attended ERO during the school year[a]	87.0	98.2	87.4	100.8	88.6	95.5
Sample size	1,408	1,264	686	645	722	619

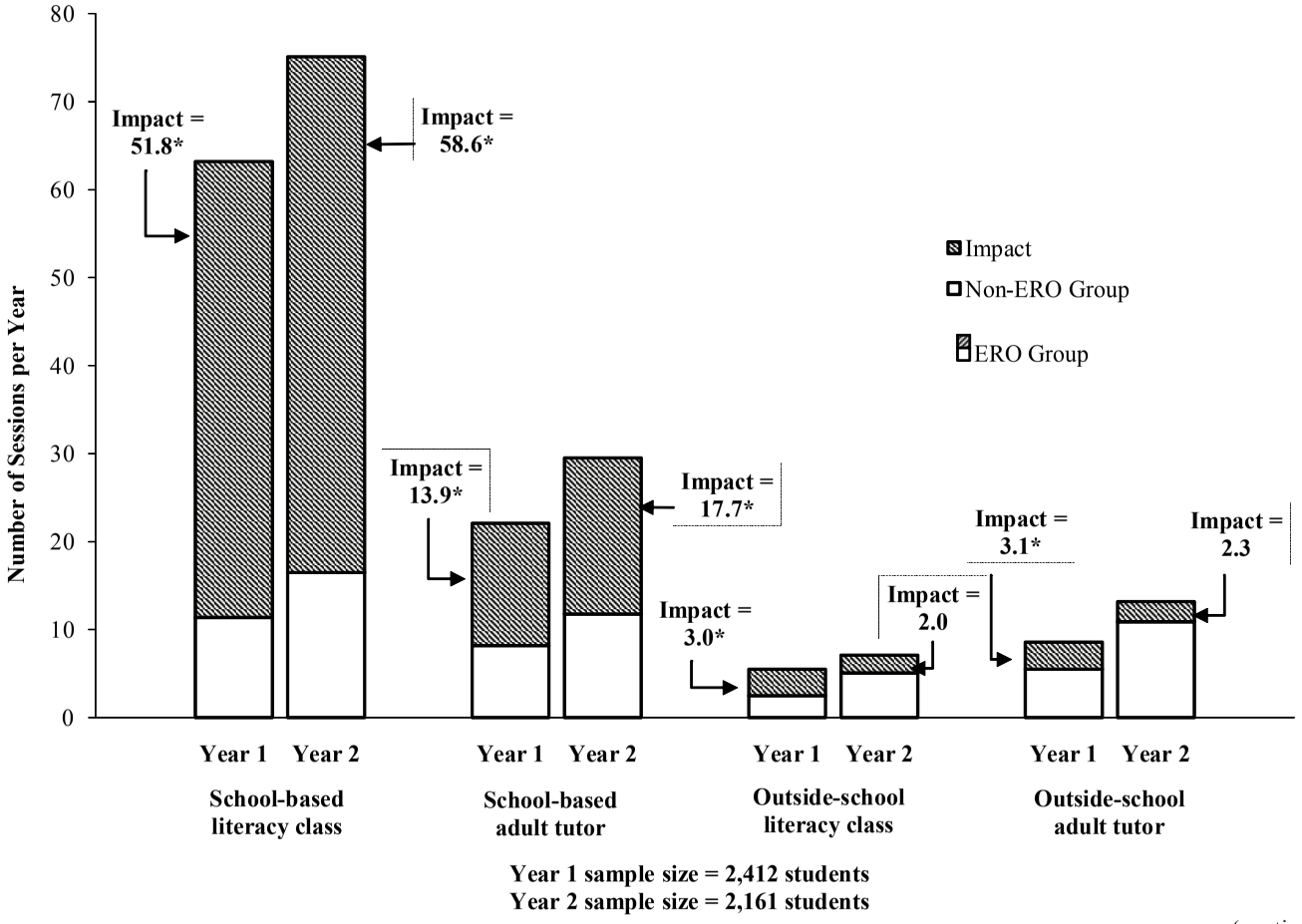
SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study monthly attendance data.

NOTE:
[a]The average number of hours student attended ERO during the school year is the average number of months ERO program was in operation multiplied by the average number of hours student attended class per month.

the frequency and duration of time spent in a school-based literacy class and with a school-based literacy tutor.[94] Both cohorts also show smaller impacts on student participation in literacy support activities outside school for ERO students, but these impacts in the second year are not statistically significant.

[94]As discussed above in this chapter, students in the ERO group in the second year of the study participated in a school-based literacy class four and a half times more frequently than students in the non-ERO group. In the first year of the study, students in the ERO group participated in a school-based literacy class five and a half times more frequently than students in the non-ERO group.

Figure 4.1
Participation in Supplemental Literacy Support Activities, Comparison of Year 1 and Year 2



Year 1 sample size = 2,412 students
Year 2 sample size = 2,161 students

(continued)

Figure 4.1 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities follow-up student survey.

NOTES: The student follow-up survey was administered in spring 2006 and 2007 at the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment.

A two-tailed t-test was applied to the impact estimate. The statistical significance level is indicated (*) when the p-value is less than or equal to 5 percent.

For each of the above measures, data are missing for no more than 5.2 percent of the respondents.

Rounding may cause slight discrepancies in calculating sums and differences.

Chapter 5

Impacts on Student Reading Achievement and Reading Behaviors

The primary focus of the Enhanced Reading Opportunities (ERO) evaluation is to assess the impact of supplemental literacy interventions on adolescent students' reading comprehension skills and behaviors and on their overall academic performance during high school. The impact analysis presented in this report addresses two questions that pertain to the second year in which the ERO programs were being implemented and to their effects, together and separately, for ninth-grade students at the end of the year in which they were enrolled in the programs:

- What is the impact of supplemental literacy programs on ninth-grade students' reading comprehension as measured by standardized test scores for reading comprehension and reading vocabulary?
- What is the impact of supplemental literacy programs on ninth-grade students' vocabulary and on their reading behaviors as measured by self-reported information about how much students read and whether they use specific reflective reading strategies?

Because the study's two supplemental literacy programs — Reading Apprenticeship Academic Literacy (RAAL) and Xtreme Reading — focus on producing immediate improvements in students' reading comprehension ability, the impact analysis presented in this report places a higher priority on the first question above. Each of the programs also endeavors to enhance students' vocabulary and their interest in reading both in and outside school and to increase their use of strategies that are characteristic of proficient readers. For this reason, the analysis also examines impacts on vocabulary test scores and on three measures of students' reading behaviors. As discussed in Chapter 2, measures of students' reading comprehension and vocabulary skills are drawn from their performance on the Group Reading Assessment and Diagnostic Examination (GRADE) administered at the end of their ninth-grade year. The measures of reading behavior were developed from the follow-up survey that was administered to students in the study sample at the end of their ninth-grade year.

This chapter first presents impact findings for all 34 of the high schools in the evaluation. The results that are pooled across the two programs from the second year of implementation provide evidence about the effectiveness of the two supplemental literacy interventions selected by the expert panel for this project as a class of interventions. The chapter then presents findings for each of the two ERO programs separately. The chapter also summarizes findings for subgroups of students defined by pre-random assignment background characteristics, in-

cluding their baseline reading test scores, whether they had repeated an earlier grade, and whether a language other than English is spoken at home.

The chapter continues with a discussion of two exploratory analyses of variation in impacts across subgroups of schools in the study. The first analysis compares the impacts for the schools where the ERO teachers in the second year of implementation had returned after having taught the entire first year of the program and the schools where the ERO teachers were new to teaching the ERO program or had only taught a portion of the first year. The second analysis compares the impacts for subgroups of schools based on two implementation characteristics: the alignment of their implementation with the program models and when program start-up occurred (that is, how early in the year the programs started).

The impact findings presented in this chapter are a follow-up to those in the prior report, about the first year of implementation. Thus, a brief review of key impact findings from the first year provides additional background for the discussion of the second-year impact findings. For the first cohort of students in the study, the ERO programs had a statistically significant impact on reading comprehension test scores, on average, across all 34 high schools. Neither program had a statistically significant impact on reading comprehension test scores, but the magnitude of the individual program impacts was the same as that of the overall impact, an effect size of 0.09 standard deviation. Although overall there was not a statistically significant relationship between implementation fidelity and reading comprehension impacts, those impacts on reading comprehension were larger in the 15 schools where (1) the ERO programs began within six weeks of the start of the school year and (2) implementation was classified as moderately or well aligned with the program models (effect size of 0.17 standard deviation; p-value = 0.002), compared with the impacts for the 19 schools where at least one of these conditions was not met (effect size of 0.01 standard deviation; p-value = 0.811).[95] The chapter ends by comparing the first-year and the second-year impact results. These comparisons are made for reading achievement outcomes, overall and by program, and also include discussions of subgroup impacts as well as the relationship between impacts and implementation.

This chapter discusses the following key findings from analysis of the second year of program implementation and comparison of the first and second years of implementation:

- Overall, the ERO programs produced a positive and statistically significant impact on reading comprehension test scores, with an effect size of 0.08 standard deviation (p-value = 0.042). This impact corresponds to an improvement from the 23rd percentile nationally, as represented by the average

[95]The 0.16 standard deviation difference in effect sizes between these two sets of schools is statistically significant (p-value = 0.035). See Kemple et al. (2008).

scores for students in the non-ERO group, to the 25th percentile nationally, as represented by the average scores for students in the ERO group.

- Seventy-seven percent of the students who enrolled in the ERO classes in the second year of the study were still reading at two or more years below grade level at the end of ninth grade, relative to the expected reading achievement of a nationally representative sample of ninth-grade students.[96]

- The RAAL program produced a statistically significant, positive impact on reading comprehension test scores, with an effect size of 0.14 standard deviation (p-value = 0.015). Although it is not statistically significant, a positive impact on the same measure, with an effect size of 0.02 standard deviation, was produced by the Xtreme Reading program. The difference in impacts between the two programs is not statistically significant, *which means that it cannot be concluded that RAAL had a different effect on reading comprehension than Xtreme Reading.* The ERO programs did not produce statistically significant impacts on vocabulary test scores.

- Overall, the ERO programs produced a positive and statistically significant impact on student-reported use of reflective reading strategies, with an effect size of 0.09 standard deviation (p-value = 0.033).

- In the 25 schools where teachers taught their second full year of the ERO classes, the ERO programs produced a statistically significant impact on reading comprehension test scores, with an effect size of 0.09 standard deviation (p-value = 0.050). Although it is not statistically significant, an impact on reading comprehension test scores, with an effect size of 0.06 standard deviation, was produced by the programs in the other nine schools where teachers were new to the ERO programs or had only taught ERO for a portion of the prior year. The difference in impacts between the two groups of schools is not statistically significant, however, *which means that it cannot be concluded that the effect of the ERO programs was different in schools with a more experienced ERO teacher than in schools whose ERO teacher had less experience with the program.*

- The ERO programs had a positive impact on reading comprehension test scores at the 13 schools rated as having implemented programs that were very well aligned with the ERO program models (0.13 standard deviation;

[96]Among ninth-graders nationally, 40 percent would be expected to score at two or more years below grade level on the same assessment.

p-value = 0.047). However, the difference in impacts between this group of schools and those 21 schools with less well-aligned program implementation is not statistically significant. *Thus, it cannot be concluded that the effect of the programs was different in schools with very well-aligned implementation fidelity than in schools whose implementation fidelity was less faithful to program specifications.*

- The ERO programs also had a positive impact on reading comprehension test scores at the 23 schools where the ERO programs started within the first two weeks of the school year (0.10 standard deviation; p-value = 0.048). However, the difference in impacts between this group of schools and those 11 schools where the programs started later than two weeks into the school year is not statistically significant. *Thus, it cannot be concluded that the impact of the ERO programs was different in schools that started their programs earlier than in schools that started their programs later.*

- The average impact of the ERO programs on reading comprehension test scores across the 34 high schools in the second year of implementation (0.08 standard deviation; p-value = 0.042) is not statistically significantly different from the impact in the first year of implementation (0.09 standard deviation; p-value = 0.019). Neither of the second-year program-specific impacts on reading comprehension test scores is statistically significantly different from those in the first year.

Impacts on Reading Achievement

The ERO study assesses the impact of supplemental literacy interventions of the type represented by RAAL and Xtreme Reading. As such, the analysis focuses first on impacts that are pooled across both interventions and all sites in the study sample. Thus, in pooling the sample across all schools in the study, the analysis has sufficient power to detect statistically significant impacts that are smaller than those that can be detected for each ERO program separately.

- Overall, the ERO programs produced a positive and statistically significant impact on reading comprehension (0.80 standard score point, which corresponds to an effect size of 0.08 standard deviation; p-value = 0.042).

The first row of data in Table 5.1 shows that, averaged across all 34 participating high schools, the ERO programs improved reading comprehension test scores by 0.8 standard score point and that this impact is statistically significant (p-value = 0.042). Expressed as a proportion

The Enhanced Reading Opportunities Study

Table 5.1

Impacts on Reading Achievement, Cohort 2 Follow-Up Respondent Sample

Outcome	ERO Group	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
All schools					
Reading comprehension					
Average standard score	90.1	89.3	0.8 *	0.08 *	0.042
Corresponding grade equivalent	*6.1*	*6.0*			
Corresponding percentile	*25*	*23*			
Reading vocabulary					
Average standard score	93.5	93.5	0.0	0.00	0.986
Corresponding grade equivalent	*7.8*	*7.8*			
Corresponding percentile	*32*	*32*			
Sample size	1,264	907			

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study follow-up GRADE assessment.

NOTES: The follow-up GRADE assessment was administered in the spring of 2007 near the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form B). No statistical tests or arithmetic operations were performed on these reference points.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (reading comprehension = 10.035; reading vocabulary = 9.827).

A two-tailed t-test was applied to the impact estimate. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

of the overall variability of test scores for students in the non-ERO group, this represents an effect size of 0.08 (or 8 percent of the standard deviation of the non-ERO group's test scores). Table 5.1 also shows that this impact corresponds to an improvement from the 23rd percentile nationally, as represented by the average scores for students in the non-ERO group, to the 25th percentile nationally, as represented by the average scores for students in the ERO group.

The first bar in Figure 5.1 places this impact estimate in the context of the actual and expected change in the ERO students' reading comprehension test scores from the end of eighth grade to the end of ninth grade. The bottom section of the bar shows the average reading comprehension test score for students in the ERO group from all 34 schools at the end of their eighth-grade year. This average of 84.6 standard score points corresponds, approximately, to a grade equivalent of 4.9 (the last month of fourth grade) and indicates an average reading level at the 14th percentile for ninth-grade students nationally. This marks the starting point for measuring both the observed growth in their reading achievement through the end of their ninth-grade year and their expected growth to be estimated through the test scores of the non-ERO group at the end of ninth grade.

The middle section of the bar represents the average growth in test scores experienced by the non-ERO group. At the end of the ninth-grade year, the non-ERO group was estimated to have achieved an average standard score of 89.3, which corresponds to a grade equivalent of 6.0 and an average reading level at the 23rd percentile for ninth-grade students nationally. This growth of 4.7 points for the non-ERO group provides the best indication of what the ERO group would have achieved by the end of their ninth-grade year had they not had the opportunity to attend the ERO classes.

The top section of the bar shows the ERO impact on reading comprehension test scores. At the end of their ninth-grade year, the ERO group achieved an average standard score of 90.1, which corresponds to a grade equivalent of 6.1 and an average reading level at the 25th percentile for ninth-grade students nationally. This means that the ERO group experienced a growth of 5.5 points in their reading comprehension skills over the course of ninth grade, which is 0.8 point higher than the growth achieved by the non-ERO group. Hence, the overall impact of the ERO programs (0.8 standard score point) represents a 17 percent improvement over and above the growth that the ERO group would have experienced had they not had the opportunity to attend the ERO classes (4.7 points),[97] and it accounts for 15 percent of the average score improvement of the ERO group itself (5.5 points).[98] Thus, the ERO programs produced more

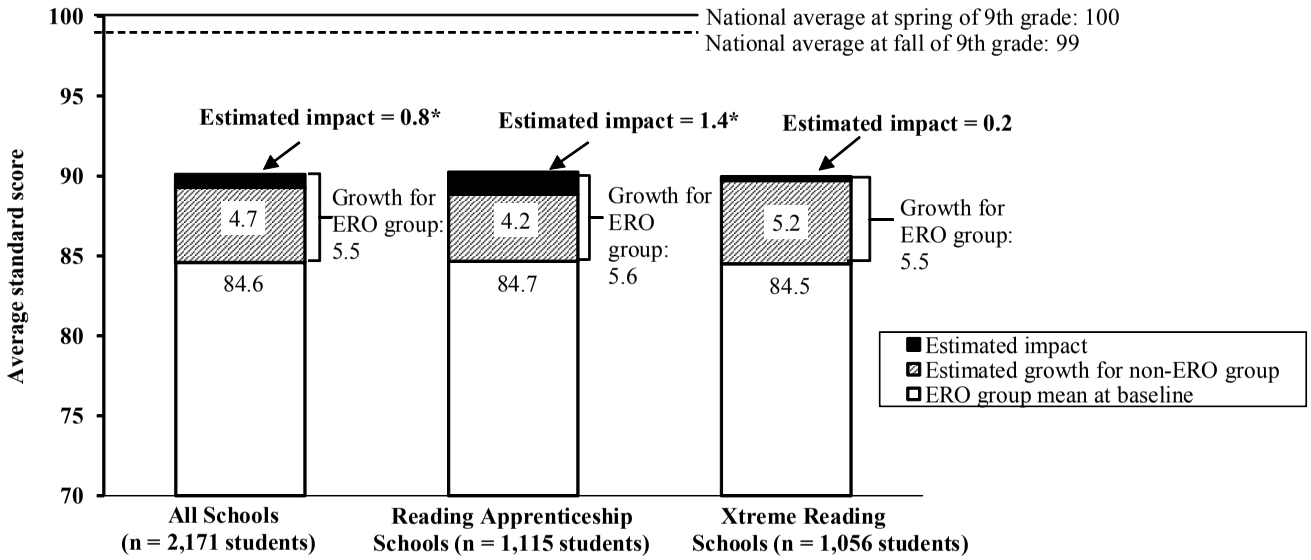
[97]The value of 17 percent was calculated by dividing the impact (0.8 standard score point) by the average improvement of the non-ERO group (4.7 standard score points).

[98]The value of 15 percent was calculated by dividing the impact (0.8 standard score point) by the average improvement of the ERO group (5.5 standard score points).

The Enhanced Reading Opportunities Study

Figure 5.1

Impacts on Reading Comprehension, Cohort 2 Follow-Up Respondent Sample



SOURCES: MDRC calculations from the Enhanced Reading Opportunities Study baseline and follow-up GRADE assessments.

NOTES: The baseline GRADE assessment was administered in the fall of 2006 at the start of students' ninth-grade year and prior to their random assignment to the ERO and non-ERO groups. The follow-up GRADE assessment was administered in the spring of 2007 near the end of students' ninth-grade year.

The ERO group growth at follow-up is calculated as the difference between the unadjusted ERO group mean at baseline and the unadjusted ERO group mean at follow-up. The impact was estimated using ordinary least squares and adjusted to account for the blocking of random assignment by school and to control for random differences between the ERO and non-ERO groups in baseline reading comprehension test scores and age at random assignment. The expected ERO group growth at follow-up is the difference between the actual ERO group growth and the impact.

A two-tailed t-test was applied to the impact estimate. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

The national average for standard score values is 100, and its standard deviation is 15.

Rounding may cause slight discrepancies in calculating sums and differences.

progress on reading comprehension than the gains expected for this sample of students had they not been selected for the programs.

The solid line at the top of Figure 5.1 shows the national average (100 standard scale points) for students at the end of ninth grade. Students scoring at this level are considered to be reading at grade level. Despite the program impact, therefore, students' reading comprehension scores still lagged nearly 10 points below the national average for performance on GRADE reading comprehension for students at the end of their ninth-grade year. In fact, 87 percent of the students in the ERO group had reading comprehension scores that were below grade level, and 77 percent had scores that were two or more years below grade level.[99] Figure 5.1 also shows that the expected fall-to-spring growth on the GRADE for a nationally representative sample of ninth-grade students is 1.0 standard score point; hence, students in both the ERO and the non-ERO group are growing by a larger amount than would be expected for the "average" ninth-grade student. However, caution should be exercised when comparing the test-score growth of the national sample to that of the study sample, since the larger growth among the latter group may be partially attributable to regression to the mean.[100]

- The RAAL program produced a statistically significant, positive impact on reading comprehension test scores, with an effect size of 0.14 standard deviation (p-value = 0.015). Although it is not statistically significant, a positive impact on the same measure, with an effect size of 0.02 standard deviation, was produced by the Xtreme Reading program. The difference in impacts between the two programs is not statistically significant, and thus it cannot be concluded that RAAL produced a different impact than Xtreme Reading.[101]

Although RAAL and Xtreme Reading share overarching goals for adolescent literacy development and share many instructional principles, these results address the issue of whether their differences in operating strategies resulted in different patterns of impacts. (For a summary of the similarities and differences between the two programs, see Chapter 3, Table 3.1.) The first row of data in the top panel of Table 5.2 shows that, averaged across the 17 high schools

[99]Differences between the ERO students and non-ERO students on these outcomes are not statistically significant. See Appendix G for the results of these impact analyses.

[100]"Regression to the mean" is a statistical artifact that makes random variation in repeated data look like true growth. Specifically, due to measurement error on the GRADE, one would expect a subset of students with lower-than-average reading comprehension scores on the GRADE (such as the students in this sample) to score closer to the national mean on the follow-up test than they did on the baseline test, even if their reading comprehension skills did not improve.

[101]As discussed more fully in Chapter 1, it is important to note that the ERO study is an evaluation of a class of reading interventions, as represented by Xtreme Reading and RAAL, as well as an evaluation of each of these two programs separately. The purpose of the study is not to test the differential impact of these two interventions.

The Enhanced Reading Opportunities Study

Table 5.2

Impacts on Reading Achievement, Cohort 2 Follow-Up Respondent Sample, by Program

Outcome	ERO Group	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
Reading Apprenticeship schools					
Reading comprehension					
Average standard score	90.2	88.9	1.4 *	0.14 *	0.015
Corresponding grade equivalent	*6.1*	*5.9*			
Corresponding percentile	*25*	*23*			
Reading vocabulary					
Average standard score	93.4	93.8	-0.4	-0.04	0.428
Corresponding grade equivalent	*7.7*	*7.8*			
Corresponding percentile	*32*	*33*			
Sample size	645	470			
Xtreme Reading schools					
Reading comprehension					
Average standard score	90.0	89.7	0.2	0.02	0.672
Corresponding grade equivalent	*6.1*	*6.0*			
Corresponding percentile	*25*	*24*			
Reading vocabulary					
Average standard score	93.5	93.1	0.4	0.04	0.468
Corresponding grade equivalent	*7.8*	*7.7*			
Corresponding percentile	*32*	*31*			
Sample size	619	437			

Difference in Impacts Between Programs	Difference in Impacts	Difference in Impact Sizes	P-Value for Difference
Reading Apprenticeship minus Xtreme Reading			
Reading comprehension standard score	1.1	0.11	0.165
Reading vocabulary standard score	-0.9	-0.09	0.285

(continued)

Table 5.2 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study follow-up GRADE assessment.

NOTES: The follow-up GRADE assessment was administered in the spring of 2007 near the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form B). No statistical tests or arithmetic operations were performed on these reference points.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (reading comprehension = 10.035; reading vocabulary = 9.827).

A two-tailed t-test was applied to the impact estimate and to the difference in impacts. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

where it was implemented, the RAAL program improved reading comprehension test scores by 1.4 standard score points and that this impact is statistically significant. This impact represents an effect size of 0.14 standard deviation (or 14 percent of the standard deviation of the non-ERO group's test scores). Table 5.2 also shows that this impact corresponds to an improvement from the 23rd percentile nationally, as represented by the average scores for students in the non-ERO group, to the 25th percentile nationally, as represented by the average scores for RAAL students in the ERO group.

As the first bar in Figure 5.1 does for the overall ERO program impact, the middle bar places the impact estimate specific to the RAAL program in the context of the actual and expected change in the RAAL students' reading comprehension test scores from the end of eighth grade to the end of ninth grade. The middle section of the bar shows a growth in test scores of 4.2 points for the non-ERO group — the best indication of what the ERO group in the 17 RAAL schools would have achieved during their ninth-grade year had they not had the opportunity to attend the RAAL classes. The top section of the bar shows the impact of 1.4 points on reading comprehension test scores, representing a 33 percent improvement over and above what the RAAL students would have achieved if they had not had the opportunity to attend the

RAAL classes.[102] Additionally, the top two sections of the middle bar in Figure 5.1 indicate that RAAL students in the ERO group improved by an average of 5.6 standard score points over the course of their ninth-grade year. Thus, the impact of the RAAL program accounts for 25 percent of the average test score improvement experienced by the RAAL students.[103]

The bottom panel of Table 5.2 shows the test score results for the Xtreme Reading program, averaged across the 17 high schools where it was implemented. The first row of data shows that the Xtreme Reading program improved reading comprehension test scores by 0.2 standard score point, but this impact is not statistically significant. It represents an effect size of 0.02 standard deviation (or 2 percent of the standard deviation of the non-ERO group's test scores). Table 5.2 also shows that this impact corresponds to an improvement from the 24th percentile nationally, as represented by the average scores for students in the non-ERO group, to the 25th percentile nationally, as represented by the average scores for Xtreme Reading students in the ERO group.

The third bar in Figure 5.1 places the impact estimate for the Xtreme Reading program in the context of the actual and expected change in the Xtreme Reading students' reading comprehension test scores from the end of eighth grade to the end of ninth grade. The middle section of the bar shows a growth in test scores of 5.2 points for the non-ERO group in the Xtreme Reading schools during their ninth-grade year. The top section of the bar shows the impact of 0.2 point on reading comprehension test scores, representing a 4 percent improvement over and above what the Xtreme Reading students would have achieved if they had not had the opportunity to attend the Xtreme Reading classes.[104] Also, the top two sections of the last bar in Figure 5.1 indicate that Xtreme Reading students in the ERO group improved by an average of 5.5 standard score points over the course of their ninth-grade year. Thus, the impact of the Xtreme Reading program accounts for 4 percent of the average test score improvement experienced by the Xtreme Reading students.[105]

The difference in the impacts produced by each program is not statistically significant. Figure 5.2 shows the estimated impacts on reading comprehension test scores, by program. The figure displays mean impact estimates for each program (represented by a circle for RAAL schools and a square for Xtreme Reading schools) and the 95 percent confidence intervals

[102]This was calculated by dividing the impact (1.4 standard score points) by the average improvement of the non-ERO group (4.2 standard score points).

[103]This was calculated by dividing the impact (1.4 standard score points) by the average improvement of the ERO group (5.6 standard score points).

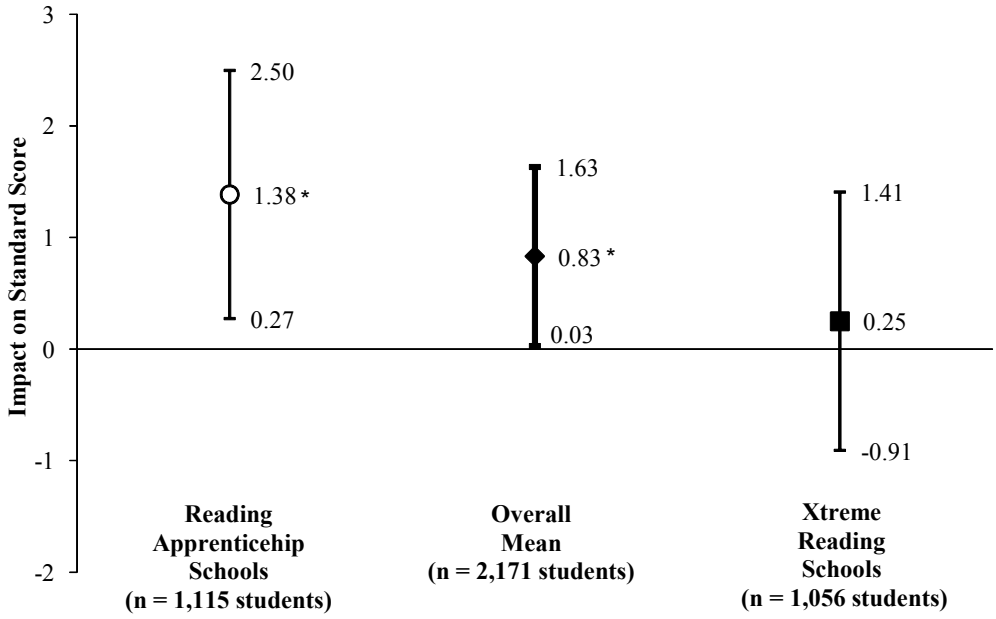
[104]This was calculated by dividing the impact (0.2 standard score point) by the average improvement of the non-ERO group (5.2 standard score points).

[105]This was calculated by dividing the impact (0.2 standard score point) by the average improvement of the ERO group (5.5 standard score points).

The Enhanced Reading Opportunities Study

Figure 5.2

Impact Estimates on Reading Comprehension, by Program



SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study follow-up GRADE assessment.

NOTES: The follow-up GRADE assessment was administered in the spring of 2007 near the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment.

A two-tailed t-test was applied to the impact estimate. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

The national average for standard score values is 100, and its standard deviation is 15.

around these estimates (represented by the lines extending above and below the circle and square). The overall mean impact across all 34 schools and its confidence interval are also presented to provide additional context. Impact estimates with confidence intervals that do not include zero are statistically significant (their p-value is less than or equal to 5 percent). The lower bound of the confidence interval of the overall impact (0.03) and that for RAAL schools (0.27) are both greater than zero. Impact estimates with confidence intervals that do not overlap are significantly different from one another. As illustrated in Figure 5.2, the confidence intervals around the impacts of the two programs overlap, showing that the difference in their impacts is not statistically significant.

Impacts are the differences between the outcomes of the ERO group and the non-ERO group (that is, the difference in the average growth of students in each group). Figure 5.1 provides additional information about growth in reading comprehension by ERO and non-ERO students at RAAL and Xtreme Reading schools. The middle and rightmost bars in the figure show a difference in growth by the non-ERO groups at each set of schools (4.2 and 5.2 standard score points at RAAL and Xtreme Reading schools, respectively). However, this difference in growth is not statistically significant. The non-ERO group at the Xtreme Reading schools grew 1.0 standard score point more than the non-ERO group at the RAAL schools (the middle sections of the middle and rightmost bars). The ERO groups in the RAAL and Xtreme Reading schools grew similar amounts: 5.6 and 5.5 standard score points, respectively (the top two sections of each of the program-specific bars).

- Overall, vocabulary test scores for students in the ERO group were estimated to be the same as those for the non-ERO group.[106] Additionally, estimated impacts on vocabulary test scores for each ERO program are not statistically significant.

Table 5.1 shows that there was no impact of the programs overall on vocabulary test scores. The vocabulary test scores averaged across all 34 participating high schools were the same for the ERO and non-ERO groups — a scale score of 93.5. Table 5.2 shows that neither ERO program produced a statistically significant impact on vocabulary test scores. In the RAAL schools, the ERO group scored 0.4 standard score point lower than the non-ERO group; in the Xtreme Reading schools, the ERO group scored 0.4 standard score point higher than the non-ERO group.

[106]The ERO study did not include a vocabulary test at baseline. As a result, it is not possible to discuss vocabulary results in the context of changes that occurred over the course of students' ninth-grade year.

Impacts on Students' Reading Behaviors

The impact analysis presented in this section of the chapter focuses on three measures of reading behavior that were derived from the student survey administered at the end of ninth grade: amount of school-related reading, amount of non-school-related reading, and use of reflective reading strategies.[107] Table 5.3 presents findings on the ERO programs' average impact on these three measures. Table 5.4 presents these results separately for each of the two ERO programs.

- Overall, the ERO program impacts on the reading behavior measures are not statistically significant.

Each of the two supplemental literacy programs seeks to motivate students to read more. They do this both by providing opportunities for students to read and discuss what they read in the ERO classes and by providing classroom libraries and assigning texts for students to read at home. The goal is to expose students to a wide range of reading opportunities, while building the strategies that proficient readers use and thereby stimulating students' interest in reading more both for school and for their own enjoyment.

The first two rows of data in Table 5.3 show the amount of reading that students in the ERO and non-ERO groups across all 34 high schools reported doing, both for school and outside school. The ERO and non-ERO groups do not differ significantly on these measures. The impact of the ERO programs on students' reports of using reflective reading strategies is statistically significant (p-value = 0.033) and represents an effect size of 0.09 standard deviation (or 9 percent of the standard deviation of the non-ERO group's test scores). As noted in Chapter 2, the analyses include qualifying statistical tests aimed at assessing the robustness of multiple impacts within the reading behavior measurement domain. The qualifying tests examine the estimated impact on a composite index of reading behaviors.[108] The composite qualifying statistical test for the multiple hypothesis tests reflected in the table indicates that the overall impact across the three reading behavior measures is not statistically significant (t-statistic = 1.15; p-value = 0.250). Therefore, the impact on the use of reflective reading strategies should be interpreted with caution. When examined separately, neither RAAL nor Xtreme Reading produced statistically significant impacts on any of the three reading behaviors. Table 5.4 shows the impacts on reading behaviors separately for each ERO program — the upper panel for RAAL schools and the lower panel for Xtreme Reading schools.

[107] A list of the survey items used to create these three measures is presented in Appendix A.

[108] See Appendix Table E.3 for the results of these qualifying tests.

The Enhanced Reading Opportunities Study

Table 5.3

Impacts on Reading Behaviors, Cohort 2 Follow-Up Respondent Sample

Outcome	ERO Group	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
All schools					
Amount of school-related reading (prior month occurrences)	46.3	45.9	0.4	0.01	0.800
Amount of non-school-related reading (prior month occurrences)	29.9	29.4	0.5	0.01	0.729
Use of reflective reading strategies (4-point scale)	2.8	2.7	0.05 *	0.09 *	0.033
Sample size	1,260	901			

SOURCE: MDRC calculations from the Enhanced Reading Opportunities follow-up student survey.

NOTES: The student follow-up survey was administered in spring 2007 at the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (school-related reading standard deviation = 38.322; non-school-related reading standard deviation = 32.976; use of reflective reading strategies standard deviation = 0.592).

A two-tailed t-test was applied to the impact estimate. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

For each of the above measures, data are missing for no more than 2.6 percent of the respondents.

Rounding may cause slight discrepancies in calculating sums and differences.

The Enhanced Reading Opportunities Study

Table 5.4

Impacts on Reading Behaviors, Cohort 2 Follow-Up Respondent Sample, by Program

Outcome	ERO Group	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
Reading Apprenticeship schools					
Amount of school-related reading (prior month occurrences)	50.5	48.8	1.6	0.04	0.483
Amount of non-school-related reading (prior month occurrences)	31.3	31.6	-0.3	-0.01	0.867
Use of reflective reading strategies (4-point scale)	2.8	2.7	0.0	0.07	0.257
Sample size	642	466			
Xtreme Reading schools					
Amount of school-related reading (prior month occurrences)	42.0	42.8	-0.8	-0.02	0.697
Amount of non-school-related reading (prior month occurrences)	28.4	26.9	1.5	0.04	0.458
Use of reflective reading strategies (4-point scale)	2.7	2.7	0.1	0.12	0.059
Sample size	618	435			

Difference in Impacts Between Programs	Difference in Impacts	Difference in Impact Effect Sizes	P-Value for Difference
Reading Apprenticeship minus Xtreme Reading			
Amount of school-related reading	2.5	0.06	0.436
Amount of non-school-related reading	-1.8	-0.05	0.522
Use of reflective reading strategies	0.0	-0.05	0.550

(continued)

Table 5.4 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities follow-up student survey.

NOTES: The student follow-up survey was administered in spring 2007 at the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (school-related reading standard deviation = 38.322; non-school-related reading standard deviation = 32.976; use of reflective reading strategies standard deviation = 0.592).

A two-tailed t-test was applied to the impact estimate and to the difference in impacts. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

For each of the above measures, data are missing for no more than 3 percent of the respondents.

Rounding may cause slight discrepancies in calculating sums and differences.

Impacts for Subgroups of Students

The ERO study sample consists of a diverse population of students who had baseline reading comprehension skills two to five years below grade level at the end of eighth grade. With this diversity in mind, the ERO evaluation was designed to allow for the estimation of impacts for key subgroups of students who face especially challenging barriers to literacy development and overall performance in high school. For example, prior research has shown that especially low literacy levels, evidence of failure in prior grades, and having English as a second language are powerful predictors of school success.[109]

This section of the chapter and Appendix H examine variation in ERO program impacts for subgroups of students defined by their baseline reading comprehension test scores, whether they were overage for the ninth grade, and whether a language other than English was spoken in their homes. As reported in Chapter 2 (see Table 2.7), 39 percent of the study sample in the second year of implementation had baseline test scores that indicate reading levels that were four to five years below grade level at the end of eighth grade, and another 29 percent were reading from three to four years below grade level. Also, over a quarter of the students in the study sample were overage for the ninth grade (that is, they were age 15 or older at the start of ninth grade), which is used to indicate that a student was retained in a prior grade.[110] Fifty per-

[109]Roderick (1993); Fine (1988).

[110]National Center for Education Statistics (1990).

cent of the students in the sample lived in households where a language other than English was spoken at least sometimes.

Table 5.5 provides a summary of impact findings for the subgroups of students defined by their baseline reading comprehension test scores, whether they were overage for the ninth grade, and whether a language other than English was spoken in their homes.[111] In general, the table indicates that the ERO programs produced positive and statistically significant impacts on reading comprehension test scores (effect size of 0.10 standard deviation; p-value = 0.028) and on the use of reflective reading strategies (effect size of 0.10; p-value = 0.034) for one of the subgroups (students not overage for grade). However, these subgroup impacts on reading comprehension and the use of reflective reading strategies are not statistically significantly different from those of the students in the counterpart subgroup (students overage for grade) Additionally, the composite qualifying statistical test for the multiple hypothesis tests reflected in the table indicates that the overall variation in impacts across the subgroups is not statistically significant (F-statistic = 1.242; p-value = 0.291).[112] As a result, although the ERO programs produced statistically significant impacts on two outcomes for a specific subgroup, the analysis does not provide adequate confidence to conclude that the programs "worked better" for students in this subgroup than it did for students in their counterpart subgroup. Thus, the two statistically significant impacts in Table 5.5 should be interpreted cautiously.

The Relationship Between Impacts and Second-Year Implementation Issues

This section explores the relationship between school-level impacts and the strength of ERO program implementation. In the first year of the study, various implementation challenges resulted in low program duration and weakly aligned implementation fidelity in several of the participating high schools. There was some concern that these schools may not have been able to conduct a fair test of their ERO program's effectiveness, and, for this reason, the first evaluation report focuses on the relationship between school-level impacts and whether a school's ERO program was especially weak in terms of its implementation. As seen in Chapters 3 and 4, however, implementation fidelity and program duration substantially improved in the second year of the study: at the spring site visit, only one school's program was deemed to be poorly aligned with the program model, and only three schools were not able to operate their program

[111]Appendix Tables H.1 through H.6 provide the outcome levels for the ERO and non-ERO groups, the estimated impacts, impact effect sizes, and p-values for the estimates presented in Table 5.5. The tables in Appendix H also show the difference in estimated impacts across subgroups and the p-values of these differences.

[112]See Appendix H. Also, as noted in Chapter 2, the impact analysis for this report does not include 762 students whose baseline reading test scores were not within the target range intended for the study. Sensitivity tests of the impact estimates indicate that the findings when including these students are similar.

The Enhanced Reading Opportunities Study

Table 5.5

Impact Effect Sizes for Student Subgroups

	Number of Students	Reading Comprehension Impact Effect Size	Reading Comprehension P-Value	Vocabulary Impact Effect Size	Vocabulary P-Value	Amount of School-Related Reading Impact Effect Size	Amount of School-Related Reading P-Value	Amount of Non-School-Related Reading Impact Effect Size	Amount of Non-School-Related Reading P-Value	Use of Reflective Reading Strategies Impact Effect Size	Use of Reflective Reading Strategies P-Value
Baseline reading comprehension											
2.0-3.0 years behind grade level	705	0.09	0.170	0.12	0.088	-0.01	0.925	0.07	0.353	0.13	0.070
3.1-4.0 years behind grade level	619	0.09	0.273	0.04	0.625	0.03	0.711	-0.04	0.595	0.05	0.552
4.1-5.0 years behind grade level	847	0.08	0.210	-0.11	0.106	0.01	0.920	0.04	0.582	0.12	0.119
Overage for grade[a]											
Students overage for grade	568	0.05	0.592	0.01	0.939	0.02	0.841	0.01	0.940	0.05	0.585
Students not overage for grade	1,603	0.10 *	0.028	0.00	0.996	0.01	0.891	0.03	0.598	0.10 *	0.034
Language spoken at home											
Students from multilingual families	1,079	0.08	0.181	-0.02	0.768	0.04	0.532	-0.02	0.778	0.07	0.224
Students from English-only families	1,092	0.09	0.107	0.03	0.550	-0.02	0.709	0.04	0.458	0.10	0.090

SOURCES: MDRC calculations from the Enhanced Reading Opportunities Study follow-up GRADE assessment and follow-up student survey.

NOTES: Appendix H provides detailed information about each of the student subgroup impact estimates including outcome levels for the ERO and non-ERO group, impact estimates, p-values, and differences in impacts between subgroups.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (reading comprehension = 10.035; reading vocabulary = 9.827; school-related reading = 38.322; non-school-related reading = 32.976; use of reading strategies = 0.592).

A two-tailed t-test was applied to the impact estimate. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

[a]A student is defined as overage for grade if he or she turned 15 before the start of ninth grade.

for more than seven and a half months. In addition, 25 of 34 ERO teachers started the second year having taught the program the entire first year of the study. Hence, unlike the first report, which focuses on schools whose implementation of the programs was problematic, the exploratory analysis in this report examines the relationship between school-level impacts and whether a school's ERO program was particularly strong in terms of its implementation, based on the timeliness of program start-up and fidelity to the program models.

The relationship between school-level impacts and strong program implementation is examined in four stages. The first stage provides an assessment of overall variation in impacts on reading comprehension test scores across the 34 participating schools. To the extent that there is variation in impacts across the sites, the overall average may be masking important differences in the effectiveness (or lack of effectiveness) of the ERO programs under some conditions. The second stage explores the policy-relevant question of whether an ERO teacher's experience with the program could be related to impacts on student achievement and reading behavior outcomes. The third stage explores two sets of relationships: (1) the relationship between impacts and implementation fidelity and (2) the relationship between impacts and the number of weeks between the start of the school year and the first day of the ERO program. The fourth stage combines these indicators and presents impacts for two groups of sites defined by whether their implementation of the ERO program was especially strong in terms of both implementation fidelity and program duration.

It is important to note that the analyses presented in this part of the chapter are exploratory. They cannot be used to establish causal links between implementation characteristics and school-level impacts on students' reading achievement. A variety of other program and school characteristics — not examined in the analyses presented here — may explain differences in impacts across the schools.

Overall Variation in ERO Impacts Across Schools

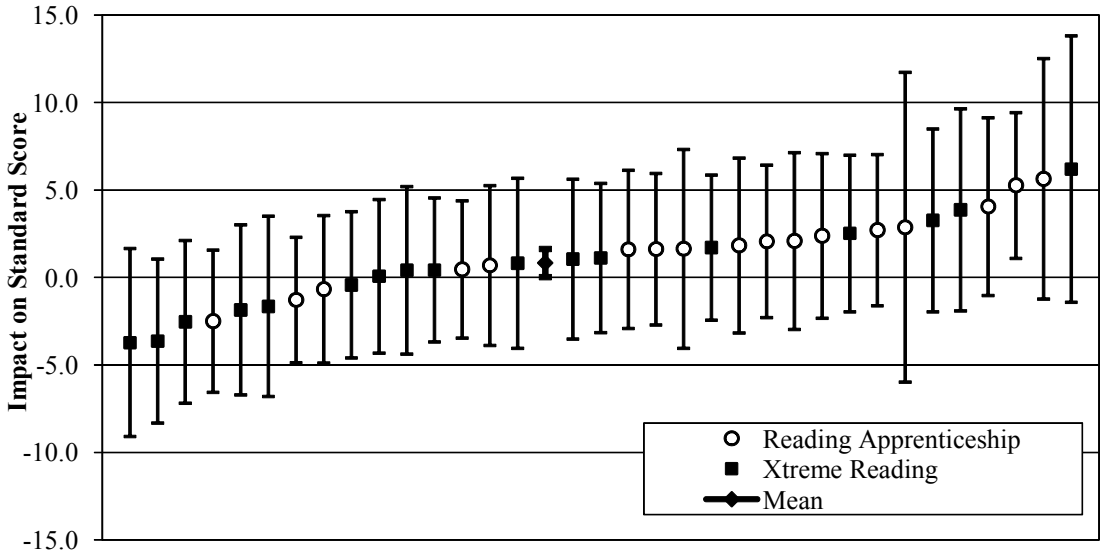
Figure 5.3 illustrates the variation in estimated program impacts on reading comprehension scores across the 34 participating high schools.[113] For each school and for the overall average, the figure displays mean impact estimates (represented by circles for RAAL schools and by squares for Xtreme Reading schools) and the 95 percent confidence intervals around the mean impact estimates (represented by the lines extending above and below the circles and squares). In this figure, the wider the confidence interval, the broader the margin of error and the greater the uncertainty about the impact estimate. Confidence intervals that do not include zero are statistically significant (the p-value is less than or equal to 5 percent). The school-by-school impact

[113]Estimated impacts are presented in numerical (ascending) order. See Appendix I for numeric values presented in Figure 5.3.

The Enhanced Reading Opportunities Study

Figure 5.3

Fixed-Effect Impact Estimates on Reading Comprehension, by School



SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study follow-up GRADE assessment.

NOTES: The follow-up GRADE assessment was administered in the spring of 2007 near the end of students' ninth-grade year.

The fixed-effect impact estimates are the regression-adjusted coefficients on the interaction between school and treatment using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment.

estimates range from an ERO program producing a decrease in reading comprehension test scores of 3.7 standard score points to an ERO program producing an increase of 6.2 standard score points. In all, 25 estimates are positive, and 9 are negative; 15 estimates are smaller than the full-sample average, and 19 estimates are about the same or larger. One of the school-level impact estimates is statistically significant.

The variation in estimated impacts displayed in Figure 5.3 overstates the variation in true impacts, however, because a large portion of the variation in estimated impacts is due to estimation error. In other words, many of the estimates in the figure appear to be highly negative or highly positive; yet, for all but one of the estimates, their confidence intervals include zero, which indicates that they cannot be distinguished reliably from zero. For example, the

second-most-negative impact is –3.6 standard score points, but its confidence interval ranges from –8.3 to 1.0 standard score points; the second-most-positive impact is 5.6 standard score points, but its confidence interval ranges from –1.2 to 12.5 standard score points.

To examine variability in impacts across schools more systematically, a composite F-test was used to assess whether the school-level impacts on reading comprehension test scores are statistically equivalent. This test accounts for estimation error in school-level impacts and provides an indication of the confidence one might have that there is variation in true impacts across the schools. The results show that the p-value for the F-test is 0.614, indicating that the school-to-school variation in impacts is not statistically significant and, thus, could have occurred by chance. This reduces the confidence that one might have that there is variation in true impacts across schools.[114]

Given the lack of statistically significant variation in true impacts, the exploratory analyses in this section are unlikely to provide any definitive answers as to whether teacher experience teaching the programs or strong program implementation explains differences in impacts across schools.[115] Nonetheless, the relationship between program implementation and school-level impacts carries practical relevance for policymakers and practitioners. Hence, the following three sections present the results from the exploratory analysis of this relationship.

Impacts Associated with Teacher Experience with the ERO Programs in the Second Year

This part of the exploratory analysis examines whether the ERO programs produce stronger impacts if they are taught by teachers who have more experience with the program. This analysis proceeds in two steps.

First, the impact of the ERO programs is examined for two subgroups of schools defined by their ERO teacher's experience with teaching the program in the second year. The first subgroup of schools consists of those whose ERO teacher had taught the entire first year of the program (that is, teachers having taught two full years of the program by the end of the study), while the second subgroup of schools consists of those whose ERO teacher was either new to the program in the second year or had only started teaching the program midway through the first year (teachers having taught less than two full years of ERO by the end of the study).

[114]See Appendix I for the results of this F-test.

[115]In particular, it is unlikely that the analysis will find statistically significant differences in impacts among subgroups of schools defined by their program implementation.

The top panel of Table 5.6 provides a summary of the impact findings for the subgroups of schools defined by these categories.[116] The second column shows the estimated impact on reading comprehension test scores, and the third column shows the estimated impact on vocabulary test scores. The three pairs of columns to the right show estimated impacts on the three reading behavior measures. All impact estimates are presented in effect-size units.

In the 25 schools whose ERO teacher returned in the second year having taught the entire first year of the program, the programs produced a statistically significant impact of 0.9 standard score point on reading comprehension test scores (an effect size of 0.09 standard deviation). In the nine schools whose ERO teacher was new to the program in the second year or had taught only a portion of the prior year, the ERO programs did not produce a statistically significant impact on reading comprehension. However, the difference in impacts between the two groups of schools is not statistically significant. Therefore, while the findings indicate that ERO programs taught by more experienced teachers had a statistically significant impact on reading comprehension scores, one cannot conclude that the impact of these programs was larger than that of programs taught by less experienced ERO teachers. The top panel of Table 5.6 indicates that ERO programs taught by more experienced teachers also produced a statistically significant impact on one of the three reading behaviors (use of reflective reading strategies). However, the statistical significance of this impact must be interpreted with caution, because composite qualifying tests indicate that these programs did not have a statistically significant impact on the composite measure of reading behaviors (see Appendix Table E.3). Finally, Table 5.6 indicates that the nine ERO programs whose teacher was less experienced did not produce a statistically significant impact on any of the outcomes.

The analysis next uses an alternate approach to examining the relationship between impacts and teacher experience with the ERO programs. Specifically, the analysis compares the impacts produced by the programs at the schools with the 25 returning teachers on the second cohort of students (Year 2) with the impacts at those same schools on the first cohort of students (Year 1). The results of this analysis are also presented in the top panel of Table 5.6. As is shown, the impact produced at these schools on reading comprehension scores is positive and statistically significant for both years. While the impact at these schools in the second year of the study is somewhat smaller than in the first year (an effect size of 0.11 standard deviation in Year 1 and 0.09 standard deviation in Year 2), the difference between the impacts is not statistically significant. Hence, one cannot infer from these results that the impact at these schools either increased or decreased with teachers in their second year of teaching the ERO program.

[116] Appendix Tables I.2 through I.11 provide the outcome levels for the ERO and non-ERO groups, the estimated impacts, impact effect sizes, and p-values for the estimates presented in Table 5.6. The tables in Appendix I also show the differences in estimated impacts across school subgroups and the p-values of these differences.

The Enhanced Reading Opportunities Study

Table 5.6

Impact Effect Sizes, by Second-Year Implementation Strength

	Number of Schools	Reading Comprehension Impact Effect Size		Reading Comprehension P-Value	Vocabulary Impact Effect Size	Vocabulary P-Value	Amount of School-Related Reading Impact Effect Size	Amount of School-Related Reading P-Value	Amount of Non-School-Related Reading Impact Effect Size	Amount of Non-School-Related Reading P-Value	Use of Reflective Reading Strategies Impact Effect Size		Use of Reflective Reading Strategies P-Value
Teacher Experience with ERO program													
Teachers who taught two full years	25	0.09	*	0.050	-0.01	0.791	-0.03	0.590	0.01	0.774	0.10	*	0.046
Teachers who taught less than two full years	9	0.06		0.487	0.03	0.694	0.11	0.150	0.02	0.843	0.07		0.441
Teachers who taught two full years Impact in Year 1[a]	25	0.11	*	0.014	0.05	0.274	0.05	0.328	0.07	0.095	-0.05		0.331
Program implementation fidelity													
Very well-aligned implementation	13	0.13	*	0.047	0.00	0.984	0.02	0.761	0.01	0.864	0.05		0.463
Well-aligned implementation	13	0.06		0.417	0.01	0.897	0.05	0.494	0.08	0.273	0.10		0.187
Moderately aligned implementation	7	0.02		0.837	-0.03	0.757	-0.05	0.526	-0.06	0.525	0.13		0.134
Poorly aligned implementation	1	–		–	–	–	–	–	–	–	–		–
Continuous fidelity measure	34	0.07		0.466	0.04	0.673	0.06	0.585	0.04	0.72793	-0.10		0.340
Weeks to program startup													
Less than 2 weeks	23	0.10	*	0.048	-0.03	0.492	0.00	0.952	0.01	0.880	0.08		0.090
2 weeks or more	11	0.04		0.546	0.08	0.324	0.03	0.683	0.03	0.679	0.10		0.217
Continuous duration measure	34	-0.01		0.532	0.00	0.787	0.01	0.355	0.01	0.436	0.01		0.605
Overall implementation strength													
Very well-aligned implementation fidelity and less than 2 weeks to program startup	12	0.13		0.062	-0.02	0.729	0.04	0.554	0.01	0.911	0.04		0.587
Weaker implementation fidelity and/or 2 weeks or more to program startup	22	0.05		0.296	0.01	0.836	-0.01	0.883	0.02	0.751	0.12	*	0.029

(continued)

Table 5.6 (continued)

SOURCES: MDRC calculations from the Enhanced Reading Opportunities Study follow-up GRADE assessment and follow-up student survey.

NOTES: Appendix I provides detailed information about each of the program implementation subgroup impact estimates including outcome levels for the ERO and non-ERO group, impact estimates, p-values, and differences in impacts between subgroups.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (reading comprehension = 10.035; reading vocabulary = 9.827; school-related reading = 38.322; non-school-related reading = 32.976; use of reading strategies = 0.592).

A two-tailed t-test was applied to the impact estimate. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

The group means and impact estimates for the "poorly aligned implementation" category are not reported in the table because only one school was deemed to have poorly aligned implementation fidelity at the second-year spring site visit. In particular, one cannot estimate the impact of the program in "poorly aligned" schools based on only one school. These results are also suppressed in order to prevent the identification of this particular school and its impact.

[a]Findings in this row represent the estimated impact in Year 1 of teachers who taught two full years of the ERO program.

The results in Table 5.6 also show that the ERO programs at schools with returning teachers produced a larger impact on students' use of reading strategies in their second year (an effect size of 0.10 standard deviation) than in their first year (an effect size of –0.05 standard deviation), and the difference between these impacts is statistically significant. However, this finding should be interpreted with caution because composite qualifying tests indicate that the difference in impacts between Year 1 and Year 2 on the composite measure across all reading behaviors is not statistically significant.

Impacts Associated with Implementation Fidelity and the Number of Weeks to Program Start-Up in the Second Year

The analysis next examines the impact of the ERO programs for subgroups of schools defined by the implementation fidelity of their ERO program at the time of the spring site visit. In particular, schools are classified as "very well aligned," "well aligned," "moderately aligned," or "poorly aligned" with their respective program models. (See Chapter 3 for a discussion of these categories.) This part of the analysis examines the hypothesis that ERO programs can produce stronger impacts if they are able to create classroom learning environments and develop instructional strategies that are very well aligned with the specifications of the program that they are using.[117]

The second panel of Table 5.6 indicates that, on average, the 13 schools whose ERO program was very well aligned to their program model on both the classroom learning environment and the comprehension instruction dimension produced positive and statistically significant impacts on reading comprehension test scores (effect size of 0.13 standard deviation; p-value = 0.047). The impact of the ERO programs in the weaker-fidelity subgroups is not statistically significant. Also, differences in impacts between the subgroups of schools defined by implementation fidelity are not statistically significant. As such, one cannot infer that schools whose ERO programs were very well aligned to developers' specifications produced larger impacts than schools whose implementation fidelity was deemed to be weaker.

The second panel of Table 5.6 also provides a test of the linear relationship between impacts and a continuous indicator of overall implementation fidelity.[118] As shown in the table,

[117]Note that impacts are not presented in Table 5.6 for the "poorly aligned" category because only one site was deemed to have poorly aligned implementation fidelity at the spring site visit. In particular, one cannot make generalized statements about the impact produced by schools with poorly aligned programs based on only one school rated in that category. The results are also suppressed to protect the identity of this particular school and to keep site-specific impacts confidential.

[118]This indicator is the average of the fidelity rating for the classroom learning environment dimension and the fidelity rating for the comprehension instruction dimension from the spring site visit. A value ranging from 1 to 3 and rounded to the nearest tenth was calculated for each school. The interaction between this indicator (continued)

the linear relationship between impacts and this overall fidelity indicator is positive but not statistically significant. The second panel of Table 5.6 also indicates that the impact of the ERO programs on outcomes other than reading comprehension is not statistically significant for any of these four subgroups of schools.

The analysis next examines the impact of the ERO programs for subgroups of schools defined by the efficiency with which they were able to start operating their ERO program in the second year of the study. "Start-up efficiency" is defined as the number of weeks between the start of the school year and the ERO program's start-up, and a lower number represents greater efficiency. A more efficient start-up is more desirable, because the earlier a school can begin operating its program, the greater the amount of time available to its ERO teacher to cover the course curriculum, and the greater the exposure of ERO students to the activities and materials planned by the developers.

The third panel of Table 5.6 shows the estimated impact of the ERO programs for two groups of sites: those where the ERO program began operating within two weeks of the start of the school year and those where program start-up was delayed by two weeks or more. Schools in the former group provided the greater opportunity for students to participate in the ERO classes: these schools were able to offer their students a full school year or almost a full school year of the ERO program.[119] In addition, a punctual program start-up ensured that students who were selected for the ERO classes did not have to withdraw from an elective course in which they had already invested considerable time.[120]

The third panel of Table 5.6 shows that in the 23 schools whose program start-up was more efficient (that is, less than two weeks), the ERO programs produced an estimated impact on reading comprehension scores that is positive and statistically significant (effect size of 0.10 standard deviation; p-value = 0.048). In the 11 schools whose program start-up was more delayed, the estimated impact of the ERO programs is smaller and not statistically significant (effect size of 0.04 standard deviation). However, because the difference in impacts between these two subgroups of schools is not statistically significant, one cannot infer than impacts were larger in schools with more efficient start-up.

---

and the treatment indicator was added to the impact estimation model. The parameter estimate for this interaction term indicates whether the ERO program impact increased or decreased as a linear function of the fidelity indicator.

[119]ERO programs are designed for a nine-month period. If an ERO program begins operating two weeks into the school year, then program duration will be about 8.5 months (that is, nine months of school minus two weeks), or 94 percent of the maximum potential program duration (8.5 divided by 9).

[120]The two-week cutoff approximates the operational reality of schools — that changes in students' course schedules often occur in the first couple of weeks of school. In addition, a two-week cutoff represented a natural break point in the distribution of weeks until program start-up in the ERO high schools.

Table 5.6 also provides a test of the linear relationship between impacts and a continuous indicator of the number of weeks to ERO program start-up.[121] The estimated linear relationship between impacts and weeks to program start-up is not statistically significant, though the estimate itself is negative, as one would expect (effect size = –0.01; p-value = 0.532). Finally, the third panel of Table 5.6 indicates that impacts on outcomes other than reading comprehension for the two subgroups of sites defined by start-up efficiency are not statistically significant.

Impacts Associated with Strong Overall Implementation in the Second Year

The analysis presented in this section of the chapter attempts to shed light on the degree to which impacts may have been stronger in schools whose overall implementation of the ERO programs was especially strong in terms of both implementation fidelity and the efficiency of program start-up. Specifically, schools are categorized as having strong overall implementation if they meet the two following thresholds: (1) implementation fidelity is "very well aligned" to the program models and (2) the ERO program started operating within two weeks of the start of the school year.

As noted above, 23 schools were able to start operating their ERO programs within two weeks of the start of the school year. In 12 of these 23 schools, the implementation of the ERO programs was also classified as being very well aligned with the program models. In all, therefore, the second-year implementation experiences of 12 of the 34 participating high schools can be considered especially strong in terms of both fidelity and program duration. The bottom panel of Table 5.6 provides a summary of impacts for these 12 schools as well as for the 22 schools categorized as having less strong overall implementation. As shown in the table, neither subgroup of schools produced impacts that are statistically significant, although the impact on reading comprehension in the 12 schools with strong overall implementation is similar in magnitude to the impact produced in the 13 schools rated as having implementation very well aligned to the program models (shown in the second panel of the table).

The bottom panel of Table 5.6 also shows that ERO students at schools with more weakly implemented ERO programs reported significantly more use of reflective reading strategies than non-ERO students did. However, the difference in impacts between the two subgroups of schools on this outcome are not statistically significant; in addition, composite qualifying tests indicate that schools characterized by weaker program implementation did not produce a statisti-

[121] A value ranging from zero to 11.6 weeks was calculated for each school. The interaction between this indicator and the treatment indicator was added to the impact estimation model. The parameter estimate for this interaction term indicates whether the ERO program impact increased or decreased as a linear function of the number of weeks to program start-up.

cally significant impact on the composite measure of reading behaviors (see Appendix Table E.3). Thus, one cannot infer that the ERO programs were more effective in increasing students' use of reflective reading strategies in schools with weaker program implementation.

Comparison of Year 1 and Year 2

Overall Impacts and Impacts by Program

One rationale for the Department of Education to include a second year of program implementation in this study was to allow for the evaluation of more mature versions of the ERO programs. The expectation was that, in the second year of implementation, problems experienced in implementation of the programs in the first year could be resolved. Both Chapter 3 and Chapter 4 present information that indicates that program implementation in the second year was stronger than it was in the first in terms of alignment of the programs as implemented with the program models and in terms of start-up and program duration. Thus, this section of the chapter compares program impacts from both years of program implementation.[122]

- The average impact of the ERO programs on reading comprehension test scores across the 34 high schools in the second year of implementation (0.08 standard deviation; p-value = 0.042) is not statistically significantly different from the impact in the first year of implementation (0.09 standard deviation; p-value = 0.019). Neither of the second-year program-specific impacts on reading comprehension test scores is statistically significantly different from those in the first year.

Figure 5.4 presents information about the impact estimates from both cohorts of students. The bottom sections of the bars show the average reading comprehension test score for students in the ERO group from all 34 schools at baseline. As discussed in Chapter 2, the difference between the first cohort's average of 85.9 standard score points and the second cohort's average of 84.6 standard score points stems from the difference in when the two cohorts took the test (the beginning of ninth grade for the first cohort and the end of eighth grade for the second cohort). This marks the starting point for measuring both the observed growth in their reading achievement through the end of their ninth-grade year and their expected growth to be estimated through the test scores of the non-ERO group at the end of ninth grade.

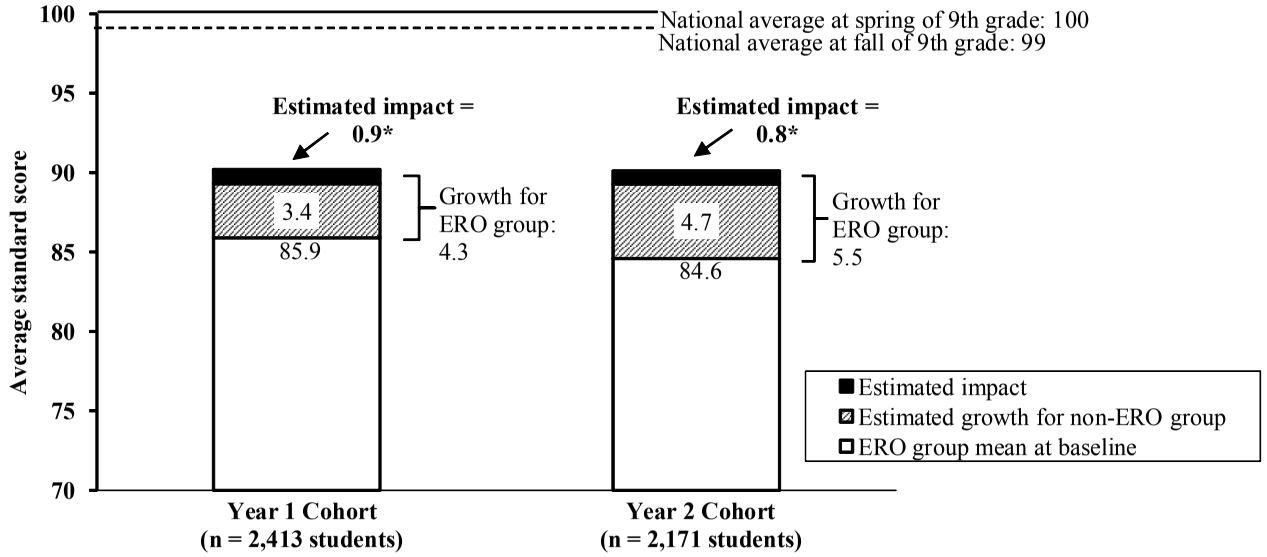
Together, the bottom two sections of the bars in Figure 5.4 show the estimated reading comprehension test scores of students in the non-ERO group from all 34 schools at the end of

[122]A two-tailed t-test was used to test the differences between Year 1 and Year 2 program impacts.

The Enhanced Reading Opportunities Study

Figure 5.4

Impacts on Reading Comprehension, Cohort 1 and Cohort 2 Follow-Up Respondent Sample



SOURCES: MDRC calculations from the Enhanced Reading Opportunities Study baseline and follow-up GRADE assessments.

NOTES: The baseline GRADE assessment was administered in the fall of 2006 at the start of students' ninth-grade year and prior to their random assignment to the ERO and non-ERO groups. The follow-up GRADE assessment was administered in the spring of 2007 near the end of students' ninth-grade year.

The ERO group growth at follow-up is calculated as the difference between the unadjusted ERO group mean at baseline and the unadjusted ERO group mean at follow-up. The impact was estimated using ordinary least squares and adjusted to account for the blocking of random assignment by school and to control for random differences between the ERO and non-ERO groups in baseline reading comprehension test scores and age at random assignment. The expected ERO group growth at follow-up is the difference between the actual ERO group growth and the impact.

A two-tailed t-test was applied to the impact estimate. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

The national average for standard score values is 100, and its standard deviation is 15.

Rounding may cause slight discrepancies in calculating sums and differences.

their ninth-grade year. The middle sections of the bars, therefore, represent the growth in test scores experienced by the non-ERO group (that is, 3.4 and 4.7 standard score points in Year 1 and Year 2, respectively). The top two sections of the bars in Figure 5.4 show the growth by the students in the ERO group in each cohort (that is, 4.3 and 5.5 standard score points in Year 1 and Year 2, respectively). The growth for both groups — ERO and non-ERO students — is greater for the students in the second cohort than for those in the first. The greater growth by the second cohort is associated with their lower baseline test scores, which, in turn, are attributable to the difference in the timing of baseline testing. The top sections of the bars show the ERO impact on reading comprehension test scores for each cohort — that is, the difference between the growth of the ERO and non-ERO groups in each cohort. The difference in the impact for the first cohort of 0.9 standard score point (effect size of 0.09 standard deviation) and the impact for the second cohort of 0.8 standard score point (effect size of 0.08 standard deviation) is not statistically significant. Thus, the average impact of the ERO programs in each year is essentially the same.

Comparing the first-cohort and second-cohort impacts on reading comprehension test scores independently for RAAL and Xtreme Reading schools yields similar results. The difference between the second-cohort impact of 1.4 standard score points and the first-cohort impact of 0.9 standard score point in RAAL schools is not statistically significant. Nor is the difference between the second-cohort impact of 0.2 standard score point and the first-cohort impact of 0.9 standard score point in Xtreme Reading schools statistically significant. Again, this suggests statistical similarity between the program-specific impacts from the first year to the second year.

Subgroup Impacts

Subgroup analyses were conducted for the first and second cohorts of students in the study. None of the subgroup impacts in the first year were replicated in the second year.[123] In terms of reading achievement outcomes, in the first year, there were statistically significant impacts on reading comprehension test scores for students overage for grade (effect size of 0.19 standard deviation; p-value = 0.007) and students from homes where a language other than English was spoken (0.12 standard deviation; p-value = 0.027). In the second year, there was a statistically significant impact on the same outcome for students *not* overage for grade (0.10 standard deviation; p-value = 0.028). In the first year, there was a statistically significant impact on the amount of non-school-related reading for students from multilingual homes (0.12 standard deviation; p-value = 0.031). In the second year, there were no statistically significant impacts on this measure and, instead, impacts on the use of reflective reading strategies for students not overage for grade (effect size of 0.10 standard deviation; p-value = 0.034). This variation in the

[123]For Year 1 subgroup impacts, see Chapter 5 and Appendix H in Kemple et al. (2008). Subgroup impacts for Year 2 are reported above in this chapter as well as in Appendix H of this report.

pattern of subgroup impacts across the two cohorts provides no evidence that, compared with other programs, the ERO programs are relatively more effective for any of the individual subgroups in the analyses.

Impacts Associated with Implementation Fidelity and Duration

In the first report from this study, findings are presented from analyses investigating the relationship between impacts and two aspects of implementation — fidelity of the programs as implemented to the RAAL and Xtreme Reading program models and duration of the programs (that is, how long they were in operation). Specifically, the analyses examine impacts for subgroups of the participating high schools that were defined by the degree to which they were able to achieve two implementation milestones during the first year of the study: whether they reached at least a moderate level of implementation fidelity (as defined in Chapter 3) and whether they were able to operate for more than seven and a half months (the average for the 34 schools in the first year). The 19 schools that were unable to reach these thresholds were deemed to have had a first-year start-up experience that was problematic. Given the challenges to program implementation that schools faced in the first year, these analyses sought to understand differences between schools where the start-up experience was problematic and schools that had an experience more in line with the expectations of the developers.

In the first year, the ERO programs produced positive and statistically significant impacts on reading comprehension test scores in the 15 schools where the ERO programs were classified as at least moderately aligned with the program model and began operation within six weeks of the start of the school year. The difference between the impacts on reading comprehension for these schools and for the remaining 19 schools is an effect size of 0.16 standard deviation.[124] This statistically significant difference in impacts is consistent with the hypothesis that a combination of higher-fidelity implementation and a more timely start-up (longer duration) may contribute positively to stronger impacts on reading comprehension. However, it is important to note that these analyses are exploratory and do not establish causal links between these early implementation milestones and variation in estimated impacts on student's reading achievement across the sites. A variety of other program and school characteristics — not examined in those analyses — may also be associated with differences in impacts across the schools.

In the second year of implementation, there was little variation across schools in fidelity to the program models (as discussed in Chapter 3) and the duration of the ERO program models (as discussed in Chapter 4). Since almost all of the high schools reached the milestones of pro-

[124]The 0.16 effect size is the difference between the 0.17 effect size for the 15 schools with better-aligned, longer-duration programs and the 0.01 effect size for the 19 schools with poorly aligned, shorter-duration programs.

gram implementation at least moderately aligned with the program models (32 schools) and program duration of seven and a half months or more (32 schools) in the second year, there is much less variation on these two aspects of implementation than in the first year. Thus, the same exploratory analyses as conducted in the first year of the relationship between impacts and these two aspects of implementation were not conducted, being less likely to provide meaningful information about differences in impacts.

The results of the first-year exploratory analyses suggest that a combination of higher-fidelity implementation and a more timely start-up (longer duration) *might* contribute positively to stronger impacts on reading comprehension. Thus, given that these implementation milestones were reached by almost all schools in the second year, one might expect that the ERO programs *should* have achieved larger impacts in the second year than in the first year. However, the comparison of impacts on reading comprehension for both cohorts indicates that they are similar. This suggests that factors other than implementation fidelity and program duration may help explain observed program impacts.

Conclusion

The impact findings indicate that, overall, the literacy programs in the ERO study produced a statistically significant improvement in students' reading comprehension skills during the second year of implementation — as they did in the first year of implementation. Although the ERO programs produced some improvement in reading comprehension test scores, students in the ERO group continued to lag behind the average ninth-grade student nationally. The 90.1 average standard score achieved by students in the ERO group at the end of their ninth-grade year corresponds, approximately, to the 6.1 grade equivalent and the 25th percentile nationally. In fact, almost 90 percent of the students in the ERO group were still reading below grade level at the end of their ninth-grade year, and 77 percent of the students in the ERO group were two or more years below grade level and, thus, would still be eligible for the ERO programs, as specified by the criteria used for this project. This finding is the same as that from the first year of implementation, despite program implementation in the second year that reflected improved alignment to the program models and longer program duration in many of the 34 participating high schools.

The ultimate goal of the two ERO programs is to improve students' academic performance during high school and to keep them on course toward graduation. With this in mind, a third and final report from the evaluation will examine the impact of the programs on the achievement and attainment outcomes of both cohorts of students as they progress through high school. The outcomes examined in the next report will include students' performance in core academic classes, their performance on the high-stakes tests required by their states, their grade-to-grade promotion rates, and whether they are on track to graduate from high school.

Appendix A

ERO Student Follow-Up Survey Measures

Two surveys were administered during the second year of the Enhanced Reading Opportunities (ERO) study. The Student Background Questionnaire, completed by all the student participants early in the 2006-2007 school year, included questions to ensure that random assignment was effective in dividing students evenly between the ERO and non-ERO groups.

This appendix describes the development of measures created from the second student-level survey, the ERO Student Follow-Up Questionnaire. This survey was administered to students in the second cohort of the study near the end of their ninth-grade year during the spring of 2007. The questions on this survey were intended to assess whether students participated in literacy support activities during the school year and to measure student attitudes and behaviors related to reading activities. A variety of measures were constructed by combining conceptually and empirically linked items from the survey. The ERO study team used a three-step process for defining and constructing the measures discussed in this appendix:

1. Identify groups of conceptually linked survey items
2. Conduct empirical tests of the correlation among the conceptually linked survey items
3. Construct multi-item outcome variables that combine the most highly correlated items

A copy of the ERO Student Follow-up Questionnaire is included at the end of this appendix.

Measures of Self-Reported Participation in Supplemental Literacy Support Activities

This section of the appendix describes four measures which assess the duration and frequency of student participation in supplemental literacy support activities: (1) attending a reading or writing class that took place in school, (2) working with a reading or writing tutor in school, (3) attending a reading or writing class that took place outside school, and (4) working with a reading or writing tutor outside school. Questions about the first of these activities were intended to determine whether students identified themselves as being enrolled in the ERO classes or similar types of classes that may have been offered in their high schools. Student reports about their participation in the other three activities were intended to provide an indication of the extent to which they utilized supplemental literacy support activities outside the ERO classes or similar classes that may have been offered in the participating high schools. The overall contrast between the ERO and non-ERO groups on these measures provides an indication of whether the ERO programs added literacy support activities to the landscape of what

would have been available to students without the programs, at least as reported by the students in the study sample.

Each of the four measures was created based on three survey items. The first item (questions 5, 8, 11, and 14) asks whether or not a student received any of these variations of extra help. (The response choices were "Yes" or "No.") The second item (questions 6, 9, 12, and 15) asks about the duration of this support. The response choices were on the following scale for the duration item:

1 = "One month"
2 = "A couple of months"
3 = "One semester or term"
4 = "Most of the year"
5 = "All year"

The third item (questions 7, 10, 13, and 16) asks about the frequency of this support. The response choices for this item were on the following scale for the frequency item:

1 = "Less than once a month"
2 = "Once a month"
3 = "Every other week"
4 = "Once a week"
5 = "Twice a week"
6 = "3-4 times a week"
7 = "Every day"

Combining responses to these three items, a measure was constructed of the total number of times during the school year that a student participated in each of the four activities. If a student answered "No" to questions 5, 8, 11, or 14, the participation measure for the activity was coded to zero (0). For students who answered "Yes" to questions 5, 8, 11, or 14, Appendix Table A.1 lists the participation values calculated for every combination of answers to the questions about duration and frequency. The columns represent duration, "how long" a student received extra help (questions 6, 9, 12, and 15). The rows represent frequency, "how often" a student received that help (questions 7, 10, 13, and 16). Duration and frequency were multiplied to create a measure of total participation throughout the school year for each student. The calculations are based on the assumption that there are 36 weeks of classes per school year and five days of classes per week.

The Enhanced Reading Opportunities Study

Appendix Table A.1

Intensity Values for Supplemental Literacy Support Measures

	One Month (4 weeks)	A Couple of Months (8 weeks)	One Semester or Term (18 weeks)	Most of the Year (27 weeks)	All Year (36 weeks)
Less than once a month (*0.1)	0.4	0.8	1.8	2.7	3.6
Once a month (*0.25)	1	2	4.5	6.75	9
Every other week (*0.5)	2	4	9	13.5	18
Once a week (*1)	4	8	18	27	36
Twice a week (*2)	8	16	36	54	72
Three to four times a week (*3.5)	14	28	63	94.5	126
Every day (*5)	20	40	90	135	180

Measures of Self-Reported Reading Behaviors

The ERO Student Follow-Up Questionnaire included 18 items aimed at measuring the frequency with which students read various texts. The ERO study team developed separate measures for reading that was related to school and for reading that was not related to school. In selecting items for these two measures, the team focused on the questions about written text that were likely to include extended passages. It also focused on groups of items for which student responses were highly correlated (that is, groups of items that were correlated with Cronbach's alpha $\geq$ .70). The seven items used to construct a measure of in-school reading frequency were correlated with Cronbach's alpha = .71, and the seven items used to construct a measure of out-of-school reading were correlated with Cronbach's alpha = .75.

The study team also developed a measure of the frequency with which a student used reading strategies in reading for other courses. The six strategies included in the measure are

often used by proficient readers and are strategies that are incorporated into the instruction of the two supplemental literacy programs for this study.[1]

Frequency of In-School Reading (7 items, Cronbach's Alpha = .71)

This construct is designed to measure the frequency with which students read extended texts for school, both during the school day and for homework. It combines student responses to questions about how often they read seven types of text during the previous month. Each possible answer is converted into a value based on the approximate number of sessions that the student reported reading these materials during the past month. The values for each of the seven types of texts were summed. If a student did not respond to an item, the value for that item is imputed using the mean of the values for the other items. If more than three of the items were missing, the entire construct is coded as missing for a given student.

Question 3. Please indicate about how OFTEN, during the <u>past month</u>, you READ each of the following in class or for homework.

a. History textbook
b. Science textbook
c. Math textbook
d. Novels, short stories, plays, poetry or essays
e. Research papers, reports, graphs, charts or tables
f. Newspaper or magazine articles
g. Workbook

Scale:

1 = "Never" = 0 sessions counted for the category
2 = "At least once" = 1 session
3 = "Every other week" = 2 sessions
4 = "Once a week" = 4 sessions
5 = "Twice a week" = 8 sessions
6 = "3-4 times a week" = 15 sessions
7 = "Every day" = 30 sessions

Frequency of Out-of-School Reading (7 items, Cronbach's Alpha = .75)

This construct is designed to measure the frequency with which students read extended texts outside school. It combines student responses to questions about how often they read seven types of text during the previous month. Each possible answer is converted into a value based on the approximate number of sessions the student reported reading a given type of ma-

[1]Biancarosa and Snow (2004).

terial during the past month. The values for each of the seven types of texts were summed. If a student did not respond to an item, the value for that item is imputed using the mean of the values for the other items. If more than four of the items were missing, the entire construct was coded as missing.

Question 4. During the past month, about how OFTEN did you READ each of the following when you were not in school and not doing homework?

b. Fiction books or stories
c. Poetry
d. Biographies or autobiographies
e. Books about science
f. Books about history
g. Newspaper or magazine articles
h. Religious books

Scale:

1 = "Never" = 0 sessions counted for the category
2 = "At least once" = 1 session
3 = "Every other week" = 2 sessions
4 = "Once a week" = 4 sessions
5 = "Twice a week" = 8 sessions
6 = "3-4 times a week" = 15 sessions
7 = "Every day" = 30 sessions

Use of Reflective Reading Strategies (4 items, Cronbach's Alpha = .77)

This construct attempts to measure the degree to which students use reading strategies in which they reflect on what they are reading and ask questions of the text to better understand what they read. These measures both are consistent with the strategies taught by the ERO programs and are seen as antecedents to reading proficiency. The two questions that make up this measure were asked in the context of the reading that students do for their English class and for the reading they do for one other core-content-area class (history, science, or math), for a total of four items.

Question 17. Please indicate how much you DISAGREE or AGREE with the following statements about your English class.

a. I ask myself questions to make sure I know the material that I have been studying for English class.
e. When I'm reading for English class I stop once in a while and go over what I have read.

Scale: 1 = "Strongly Disagree" to 4 = "Strongly Agree"

Question 18. For which one of the following classes did you do the most reading during the past school year?

4. History (or Social Studies)
5. Science
6. Math

Question 19. Please indicate how much you DISAGREE or AGREE with the following statements about the class you chose in Question 18.

a. I ask myself questions to make sure I know the material that I have been studying for class.
e. When I'm reading for class I stop once in a while and go over what I have read.

Scale: 1 = "Strongly Disagree" to 4 = "Strongly Agree"

Other Measures on Student Perceptions About Reading

The study team developed two other measures to assess the impact of the ERO program on students' perceptions of reading. The creation of each of these measures is described below.

Reading to Learn (5 items, Cronbach's Alpha = .80)

This construct was designed to measure how strongly a student connects reading with learning new things. It was created by averaging student responses to the items below. If a student did not respond to at least three items, the measure was coded as missing.

Question 2. Please indicate how much you DISAGREE or AGREE with the statements below about reading and writing.

a. When I read books, I learn a lot.
f. I read because it helps me do better in my classes.
g. I read to see what is going on in the world, the country, and/or my community.
i. I read in order to learn new things.
j. I read to learn how other people see things.

Scale: 1 = "Strongly Disagree" to 4 = "Strongly Agree"

Reading to Enjoy (2 items, Cronbach's Alpha = .82)

This construct was designed to measure whether or not a student enjoys reading. It was created by averaging student responses to the items below. If a student did not respond to at least one of the items, the measure was coded as missing.

Question 2. Please indicate how much you DISAGREE or AGREE with the statements below about reading and writing.

b. Reading is one of my favorite activities.
h. I read because I enjoy it.

ENHANCED READING OPPORTUNITIES

STUDENT FOLLOW-UP QUESTIONNAIRE
SPRING 2007
GRADE 9

First Name: «First_Name» **Last Name: «Last_Name»**

School: «School»

Student ID #: «Student_ID_Number» **Date of Birth: «Month»/ «Day»/«Year»**
Month *Day* *Year*

Today's Date: _____/_____/_____
Month *Day* *Year*

PURPOSE

We are asking you these questions to get information about your school experiences and your experiences with reading. You're the best person to help us learn about these things. We are interested in your own responses to these questions. You do not need to ask your parents, teachers, or friends for help on the answers.

This is not a test – there are no right or wrong answers. Your answers will be used for research only, so please be as honest as you can.

You do not have to answer any individual questions you don't like. We hope that you answer all the questions because we need your answers to make our research complete.

DIRECTIONS

Read each question carefully. Try to answer all questions. If no answer fits exactly, pick the one that comes closest. It is important that you follow the directions for responding to each question. Mark (✔) each answer clearly.

YOUR ANSWERS WILL BE USED FOR RESEARCH ONLY.

MDRC, New York, NY, www.mdrc.org
For questions, contact Jim Kemple at: James.Kemple@mdrc.org, Phone: (866)519-1884

FOR SURVEY ADMINISTRATOR USE ONLY
❑ **Non-ERO School Administration**

The first question asks you about your future education.

(1) How far do you think you will go in school?

Mark (✔) one answer.

- 1❍ graduate from high school
- 2❍ vocational or technical training (e.g. electrician, hairdresser, chef, pre-school teacher)
- 3❍ some college
- 4❍ graduate from a business or two-year college
- 5❍ graduate from a four-year college
- 6❍ get a master's degree
- 7❍ get a law degree, a Ph.D., or a medical doctor's degree

This section is about reading and writing.
Please mark (✔) one answer on each line.

(2) Please indicate how much you DISAGREE or AGREE with the statements below about reading and writing.

		Strongly Disagree	Disagree	Agree	Strongly Agree
a.	When I read books, I learn a lot.	1❍	2❍	3❍	4❍
b.	Reading is one of my favorite activities.	1❍	2❍	3❍	4❍
c.	Writing things like stories or letters is one of my favorite activities.	1❍	2❍	3❍	4❍
d.	Writing helps me share my ideas.	1❍	2❍	3❍	4❍
e.	When I have free time, I rarely choose to read over doing other activities.	1❍	2❍	3❍	4❍
f.	I read because it helps me do better in my classes.	1❍	2❍	3❍	4❍
g.	I read to see what is going on in the world, the country, and/or my community.	1❍	2❍	3❍	4❍
h.	I read because I enjoy it.	1❍	2❍	3❍	4❍
i.	I read in order to learn new things.	1❍	2❍	3❍	4❍
j.	I read to learn how other people see things.	1❍	2❍	3❍	4❍

Continue ➡

The next question asks about what you read for school.

(3) Please indicate about how OFTEN, during the past month, you READ each of the following in class or for homework. Mark (✓) the number on each line that applies to you.

	Never	At least once	Every other week	Once a week	Twice a week	3-4 times a week	Every day
a. History textbook	1❍	2❍	3❍	4❍	5❍	6❍	7❍
b. Science textbook	1❍	2❍	3❍	4❍	5❍	6❍	7❍
c. Math textbook	1❍	2❍	3❍	4❍	5❍	6❍	7❍
d. Novels, short stories, plays, poetry, or essays	1❍	2❍	3❍	4❍	5❍	6❍	7❍
e. Research papers, reports, graphs, charts, or tables	1❍	2❍	3❍	4❍	5❍	6❍	7❍
f. Newspaper or magazine articles	1❍	2❍	3❍	4❍	5❍	6❍	7❍
g. Websites on the Internet	1❍	2❍	3❍	4❍	5❍	6❍	7❍
h. Workbooks	1❍	2❍	3❍	4❍	5❍	6❍	7❍

This section is about reading you do that is not for school.

Please mark (✓) one answer on each line.

(4) During the past month, about how OFTEN did you READ each of the following, when you were not in school and not doing homework?

	Never	At least once	Every other week	Once a week	Twice a week	3-4 times a week	Every day
a. Fiction books or stories (books or stories about imagined events)	1❍	2❍	3❍	4❍	5❍	6❍	7❍
b. Plays	1❍	2❍	3❍	4❍	5❍	6❍	7❍
c. Poetry	1❍	2❍	3❍	4❍	5❍	6❍	7❍
d. Biographies or autobiographies	1❍	2❍	3❍	4❍	5❍	6❍	7❍
e. Books about science (for example, nature, animals, astronomy)	1❍	2❍	3❍	4❍	5❍	6❍	7❍
f. Books about history	1❍	2❍	3❍	4❍	5❍	6❍	7❍
g. Newspaper or magazine articles	1❍	2❍	3❍	4❍	5❍	6❍	7❍
h. Religious books (e.g., Koran, Bible, Catechism, Torah, other)	1❍	2❍	3❍	4❍	5❍	6❍	7❍
i. Websites on the Internet	1❍	2❍	3❍	4❍	5❍	6❍	7❍
j. Research papers, reports, graphs, charts, or tables	1❍	2❍	3❍	4❍	5❍	6❍	7❍

Continue ➡

(5) Other than your regular English class, have you taken a class, <u>in school</u> this year intended to help you with your reading and writing?	Yes	No
	1❍ If **YES**, please continue to question 6	2❍ If **NO**, please continue to question 8

(6) For how LONG did you get this help with reading and writing?	One month or less	A couple of months	One semester or term	Most of the year	All year
	1❍	2❍	3❍	4❍	5❍

(7) How OFTEN did you get this help with reading and writing?	Less than once a month	Once a month	Every other week	Once a week	Twice a week	3-4 times a week	Every day
	1❍	2❍	3❍	4❍	5❍	6❍	7❍

(8) Did an adult <u>in your school</u> help you individually with your reading and writing this year, like a tutor?	Yes	No
	1❍ If **YES**, please continue to question 9	2❍ If **NO**, please continue to question 11

(9) For how LONG did you get this help with reading and writing?	One month or less	A couple of months	One semester or term	Most of the year	All year
	1❍	2❍	3❍	4❍	5❍

(10) How OFTEN did you get this help with reading and writing?	Less than once a month	Once a month	Every other week	Once a week	Twice a week	3-4 times a week	Every day
	1❍	2❍	3❍	4❍	5❍	6❍	7❍

Continue ➡

RSRCH ID # _____

(11) Have you taken a class or participated in a program <u>outside of school</u> intended to help you with your reading and writing?	Yes	No
	1❍ If **YES**, please continue to question 12	2❍ If **NO**, please continue to question 14

(12) For how LONG did you get this help with reading and writing?	One month or less	A couple of months	One semester or term	Most of the year	All year
	1❍	2❍	3❍	4❍	5❍

(13) How OFTEN did you get this help with reading and writing?	Less than once a month	Once a month	Every other week	Once a week	Twice a week	3-4 times a week	Every day
	1❍	2❍	3❍	4❍	5❍	6❍	7❍

(14) Did an adult <u>outside of school</u> help you individually with your reading and writing this year, like a tutor or someone at an after-school program?	Yes	No
	1❍ If **YES**, please continue to question 15	2❍ If **NO**, please continue to question 17

(15) For how LONG did you get this help with reading and writing?	One month or less	A couple of months	One semester or term	Most of the year	All year
	1❍	2❍	3❍	4❍	5❍

(16) How OFTEN did you get this help with reading and writing?	Less than once a month	Once a month	Every other week	Once a week	Twice a week	3-4 times a week	Every day
	1❍	2❍	3❍	4❍	5❍	6❍	7❍

Continue ➡

This section is about your classes in school this year.

(17) Please indicate how much you DISAGREE or AGREE with the following statements about your English class. Mark (✓) the number on each line that applies to you.

		Strongly Disagree	Disagree	Agree	Strongly Agree
a.	I ask myself questions to make sure I know the material that I have been studying for English class.	1❍	2❍	3❍	4❍
b.	When I don't understand a word while reading for English class, I try to break the word down into smaller pieces.	1❍	2❍	3❍	4❍
c.	To help me understand what I'm reading for English class, I try to connect the things that are unfamiliar to me with things I already know.	1❍	2❍	3❍	4❍
d.	While reading for English class, I rarely make predictions about what will come next in a passage.	1❍	2❍	3❍	4❍
e.	When I'm reading for English class I stop once in a while and go over what I have read.	1❍	2❍	3❍	4❍
f.	When I don't know the meaning of a word while reading for English class, I often look at other words in the sentence or paragraph to help me understand.	1❍	2❍	3❍	4❍
g.	I try to identify whether what I'm reading for English class is fact or opinion.	1❍	2❍	3❍	4❍
h.	I rarely stop to summarize a passage or paragraph while reading for English class.	1❍	2❍	3❍	4❍

Continue ➡

(18) For which <u>one</u> of the following classes did you do the <u>most</u> reading during the past school year? Mark (✔) <u>one</u> answer.

1❍ History (or Social Studies) 2❍ Science 3❍ Math

(19) Please indicate how much you DISAGREE or AGREE with the following statements <u>about the class you chose in Question 18</u>. Mark (✔) the number on each line that applies to you.

	Strongly Disagree	Disagree	Agree	Strongly Agree
a. I ask myself questions to make sure I know the material that I have been studying for class.	1❍	2❍	3❍	4❍
b. When I don't understand a word while reading for class, I try to break the word down into smaller pieces.	1❍	2❍	3❍	4❍
c. To help me understand what I'm reading for class, I try to connect the things that are unfamiliar to me with things I already know.	1❍	2❍	3❍	4❍
d. While reading for class, I rarely make predictions about what will come next in a passage.	1❍	2❍	3❍	4❍
e. When I'm reading for class I stop once in a while and go over what I have read.	1❍	2❍	3❍	4❍
f. When I don't know the meaning of a word while reading for class, I often look at other words in the sentence or paragraph to help me understand.	1❍	2❍	3❍	4❍
g. I try to identify whether what I'm reading for class is fact or opinion.	1❍	2❍	3❍	4❍
h. I rarely stop to summarize a passage or paragraph while reading for class.	1❍	2❍	3❍	4❍

This final section is about your Enhanced Reading Opportunity (ERO) class (Xtreme Reading or Reading Apprenticeship For Academic Literacy). There are 3 questions.

(20) Please indicate how much you DISAGREE or AGREE with the following statements about your ERO class. Mark (✔) the number on each line that applies to you.

	Strongly Disagree	Disagree	Agree	Strongly Agree
a. I like my ERO class.	1❍	2❍	3❍	4❍
b. Compared to work I do for other subjects at school, I find the work I do for ERO to be interesting.	1❍	2❍	3❍	4❍
c. Compared with what I learn in my other subjects at school, I find what I learn in ERO to be useful.	1❍	2❍	3❍	4❍

THANK YOU!!!

Appendix B

Follow-Up Test and Survey Response Analysis

The two main data sources for the second-year impact analysis of the Enhanced Reading Opportunities (ERO) study are the Group Reading Assessment and Diagnostic Examination (GRADE) assessment of student reading skills and the student follow-up survey. Both the test and the survey were administered late in the 2006-2007 school year. Overall, 81 percent of the full study sample completed the test and survey, including 83 percent of students in the ERO program group and 79 percent of students in the non-ERO group. The lack of a 100 percent response rate combined with the discrepancy between response rates for the ERO and non-ERO student groups raises two concerns: Are the respondents representative of the full study sample? Are there systematic pre-program differences between respondents in the ERO and non-ERO groups?

The first section of this appendix discusses the follow-up test and survey response rates and examines differences between respondents and nonrespondents. The second section examines the respondent sample and assesses similarities and differences between students in the ERO and non-ERO groups.

Follow-Up Test and Survey Response Rates

Efforts were made to collect both test and survey data from all 2,679 students who make up the full study sample — ninth-grade students who consented to be in the ERO program and had pretest reading comprehension scores between the fourth- and seventh-grade levels. Sections of 25 to 30 students from both the ERO and the non-ERO group were tested and surveyed together in their high schools. The test and survey administrations took place during the school day and were proctored by members of the ERO study team. The ERO study team spent up to four days at each school locating, testing, and surveying students who did not attend the originally scheduled session.

In all, 2,160 students (81 percent of the full study sample) completed both the follow-up test and the survey. An additional 11 students completed only the follow-up test, and one student completed only the survey. Due to the similarity in response rates for the follow-up test and the survey, the nonresponse analysis in this appendix focuses on the response rate for the test. Results for the survey response and the combined response are virtually the same.

Appendix Table B.1 shows the follow-up test response rates for all 34 participating high schools combined and for the groups of schools using Reading Apprenticeship Academic Literacy (RAAL) and Xtreme Reading, respectively. Overall, 83 percent of students in the ERO group took the follow-up test, compared with 79 percent of students in the non-ERO group. The 4 percentage point difference is statistically significant (the p-value is less than or equal to 5 percent). The RAAL and Xtreme Reading schools had similar response rates for their ERO and

The Enhanced Reading Opportunities Study

Appendix Table B.1

Response Rates of Students in Cohort 2
Full Study Sample

	ERO Group	Non-ERO Group	Difference	P-Value for the Difference
All schools	82.7	79.1	3.6 *	0.018
Reading Apprenticeship schools	82.4	79.4	3.0	0.163
Xtreme Reading schools	83.0	78.7	4.3 *	0.049
Overage for grade[a]	73.4	71.0	2.4	0.455
Not overage for grade	86.4	82.8	3.6 *	0.031
Language other than English spoken at home	84.3	81.5	2.8	0.190
English only spoken at home	81.2	76.8	4.4 *	0.047
Baseline reading comprehension score				
2.0-3.0 years below grade level	85.2	81.3	3.9	0.131
3.1-4.0 years below grade level	83.7	82.9	0.8	0.767
4.1-5.0 years below grade level	80.1	73.9	6.1 *	0.016
Teachers having taught two full years of ERO	83.6	80.9	2.7	0.115
Teachers having taught less than two full years of ERO	80.1	74.1	6.0	0.054
Stronger implementation schools[b]	80.4	75.8	4.6	0.070
Weaker implementation schools[c]	84.0	81.1	3.0	0.116
Sample size	1,529	1,150		

SOURCES: MDRC calculations from the Enhanced Reading Opportunities baseline data and follow-up GRADE assessment.

NOTES: This table represents the response rates for the follow-up GRADE assessment which was administered in spring 2007 at the end of students' ninth-grade year. The follow-up student questionnaire was also administered at that time. The difference in response rates between the test and survey is negligible.

The estimated differences are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed distribution of the ERO group across random assignment blocks (i.e., schools) as the basis for the adjustment.

Rounding may cause slight discrepancies in calculating sums and differences.

[a]A student is defined as overage for grade if he or she turned 15 before the start of ninth grade.

[b]The ERO programs in these schools were deemed to have reached an implementation level that was very well aligned to both the classroom learning environment and comprehension instruction dimensions of the program model by the spring site visit, and their ERO program began operating within 2 weeks of the start of the school year.

[c]The implementation fidelity of the ERO programs in these schools was deemed to be less than very well aligned to the classroom learning environment and/or comprehension instruction dimensions of the program model by the spring site visit, and/or their ERO program began operating 2 weeks or more after the start of the school year.

non-ERO group students. The difference in response rates between the ERO and non-ERO groups is statistically significant for the Xtreme Reading schools but not for the RAAL schools.

The primary reason that students did not complete the follow-up test or survey is that they were no longer enrolled in a high school participating in the ERO study.[1] In all, 20 percent of the students in the study sample were no longer enrolled in an ERO high school at the time of the follow-up test and survey administrations. These rates are similar for the ERO group (21 percent) and the non-ERO group (20 percent). Of the students who were no longer enrolled in an ERO school, only 28 percent completed the follow-up test (compared with 95 percent of those who remained enrolled in an ERO school). As in the full sample, completion rates among students who were no longer enrolled in an ERO school differ for the ERO group (30 percent) and the non-ERO group (26 percent). Also like the full sample findings, this difference in response rates is concentrated in the Xtreme Reading sites, where 30 percent of the ERO group completed the follow-up test, compared with 24 percent of the non-ERO group. Among students who remained enrolled in an ERO school, response rates also differ between the two treatment groups: 96 percent for the ERO group and 92 percent for the non-ERO group. Unlike the full sample findings, however, the difference in response rates between treatment groups among students who remained enrolled in an ERO high school is approximately the same for Xtreme Reading and RAAL schools.

One factor that may influence the interpretation of the impact findings presented in this report is whether students who completed the follow-up test and survey are representative of the full study sample. This question was addressed in two ways. First, respondents and nonrespondents were compared directly on a range of background characteristics. The results for the full study sample are shown in Appendix Table B.2. Overall, the table indicates that nonrespondents are more likely than respondents to have characteristics associated with a risk of school failure. For example, a higher percentage of nonrespondents are overage for the ninth grade (41 percent, compared with 26 percent for respondents), thus indicating that nonrespondents are more likely to have been retained in a prior grade. In addition, a smaller percentage of nonrespondents have a mother who completed some postsecondary education (27 percent, compared with 32 percent for respondents) or a father who completed some postsecondary education (17 percent, compared with 22 percent for respondents). On average, nonrespondents also had lower reading comprehension test scores at baseline than students who completed the follow-up test (84 points for nonrespondents and 85 points for respondents). All of these differences are statistically significant at the 5 percent level. Appendix Tables B.3 and B.4 compare the respondents and nonrespondents in RAAL schools and Xtreme Reading schools, respectively.

[1]The tracking information on reasons that students did not complete the follow-up test or survey is based on data collected during the administration period and is available only in aggregate form. As a result, it does not permit breakdowns by student background characteristics.

The Enhanced Reading Opportunities Study

Appendix Table B.2

Characteristics of Students in Cohort 2: Differences Between Respondents and Nonrespondents

Characteristic	Respondents	Non-Respondents	Difference	P-Value for the Difference
Race/ethnicity (%)				
Hispanic	31.6	27.2	4.4 *	0.006
Black, non-Hispanic	46.6	48.7	-2.1	0.256
White, non-Hispanic	15.2	16.8	-1.5	0.325
Other	6.6	7.3	-0.7	0.545
Gender (%)				
Male	51.3	49.2	2.1	0.401
Female	48.7	50.8	-2.1	0.401
Average age (years)	14.7	14.9	-0.2 *	0.000
Overage for grade[a] (%)	26.2	40.6	-14.4 *	0.000
Language other than English spoken at home (%)	49.7	48.1	1.6	0.479
Language spoken at home missing (%)	1.2	5.3	-4.2 *	0.000
Mother's education level (%)				
Did not finish high school	17.4	20.9	-3.5	0.066
High school diploma or GED certificate	24.6	23.3	1.3	0.535
Completed some postsecondary education	32.3	27.3	5.0 *	0.028
Don't know	23.9	22.8	1.1	0.597
Missing	1.8	5.8	-4.0 *	0.000
Father's education level (%)				
Did not finish high school	16.2	17.5	-1.3	0.460
High school diploma or GED certificate	22.2	20.6	1.5	0.459
Completed some postsecondary education	21.6	16.6	5.1 *	0.011
Don't know	37.5	39.3	-1.7	0.477
Missing	2.5	6.0	-3.5 *	0.000
GRADE reading comprehension[b]				
Average standard score	84.7	83.6	1.1 *	0.000
Corresponding grade equivalent	*4.9*	*4.7*		
Corresponding percentile	*14*	*13*		
2.0 - 3.0 years below grade level (%)	32.5	28.2	4.2	0.065
3.1 - 4.0 years below grade level (%)	28.5	24.9	3.6	0.105
4.1 - 5.0 years below grade level (%)	39.0	46.9	-7.8 *	0.001
Sample size	2,171	508		

(continued)

Appendix Table B.2 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study baseline data.

NOTES: Baseline data for students in 30 of 34 high schools were collected in spring 2006 (during students' eighth-grade year); baseline data for students in the remaining four schools were collected in fall 2006 (at the start of students' ninth-grade year).

The estimated differences are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school. The values in the column labeled "Respondents"are the observed means for students in the study sample with follow-up data on the GRADE assessment (respondents). The "Non-Respondents" values in the next column are the regression-adjusted means for students in the study sample without follow-up data on the GRADE assessment (non-respondents), using the observed distribution of respondents across random assignment blocks (i.e., schools) as the basis for the adjustment.

A two-tailed t-test was used to test differences between the respondents and non-respondents. The statistical significance level is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

[a]A student is defined as overage for grade if he or she turned 15 before the start of ninth grade.

[b]The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form A). No statistical tests or arithmetic operations were performed on these reference points.

A second and more comprehensive strategy for assessing differences between respondents and nonrespondents is to use multiple regression to determine the extent to which the average characteristics of students who completed the follow-up test differ systematically from those of students who did not. This analysis was carried out for the full group of schools in the study and separately for the schools using RAAL and Xtreme Reading, respectively. The results are presented in Appendix Table B.5. It indicates that response rates differ by some background characteristics, including age, race/ethnicity, and baseline test scores. More important, the overall F-test for each regression indicates that there are systematic differences between the respondents and nonrespondents.

In summary, the response analysis indicates that students who completed the follow-up test and survey are not fully representative of the full study sample of 2,679 students. Thus, some caution should be exercised when attempting to generalize the findings beyond those sample members who are included in the impact analysis. Nevertheless, the overall response rates show that follow-up data are available for 81 percent of the students in the study sample, making the results reflective of the behavior of most of the targeted students.

Appendix F presents an assessment of the sensitivity of the impact findings to differences between students who completed the follow-up test and those who did not. The appendix presents estimated impacts that are weighted for differential response rates by high school,

The Enhanced Reading Opportunities Study

Appendix Table B.3

Characteristics of Students in Cohort 2: Differences Between Respondents and Nonrespondents, Reading Apprenticeship Schools

Characteristic	Respondents	Non-Respondents	Difference	P-Value for the Difference
Race/ethnicity (%)				
Hispanic	31.5	26.3	5.1 *	0.023
Black, non-Hispanic	47.4	48.8	-1.3	0.618
White, non-Hispanic	14.1	19.1	-5.1 *	0.018
Other	7.0	5.8	1.2	0.474
Gender (%)				
Male	51.1	50.6	0.5	0.879
Female	48.9	49.4	-0.5	0.879
Average age (years)	14.7	15.0	-0.2 *	0.000
Overage for grade[a] (%)	27.7	42.3	-14.6 *	0.000
Language other than English spoken at home (%)	48.4	47.6	0.8	0.785
Language spoken at home missing (%)	1.1	5.9	-4.9 *	0.000
Mother's education level (%)				
Did not finish high school	18.1	25.2	-7.1 *	0.009
High school diploma or GED certificate	25.4	23.5	1.9	0.526
Completed some postsecondary education	30.4	24.8	5.6	0.074
Don't know	24.1	20.6	3.6	0.222
Missing	2.0	5.9	-4.0 *	0.001
Father's education level (%)				
Did not finish high school	15.4	20.6	-5.1 *	0.042
High school diploma or GED certificate	23.8	18.9	4.9	0.095
Completed some postsecondary education	20.5	15.1	5.4 *	0.047
Don't know	37.5	38.6	-1.1	0.741
Missing	2.8	6.8	-4.1 *	0.002
GRADE reading comprehension[b]				
Average standard score	84.8	83.4	1.4 *	0.001
Corresponding grade equivalent	*4.9*	*4.6*		
Corresponding percentile	*15*	*12*		
2.0 - 3.0 years below grade level (%)	33.8	29.1	4.7	0.146
3.1 - 4.0 years below grade level (%)	27.1	22.8	4.3	0.161
4.1 - 5.0 years below grade level (%)	39.1	48.0	-8.9 *	0.009
Sample size	1,115	262		

(continued)

Appendix Table B.3 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study baseline data.

NOTES: Baseline data for students in 30 of 34 high schools were collected in spring 2006 (during students' eighth-grade year); baseline data for students in the remaining four schools were collected in fall 2006 (at the start of students' ninth-grade year).

The estimated differences are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school. The values in the column labeled "Respondents" are the observed means for students in the study sample with follow-up data on the GRADE assessment (respondents). The "Non-Respondents" values in the next column are the regression-adjusted means for students in the study sample without follow-up data on the GRADE assessment (non-respondents), using the observed distribution of respondents across random assignment blocks (i.e., schools) as the basis for the adjustment.

A two-tailed t-test was used to test differences between the respondents and non-respondents. The statistical significance level is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

[a]A student is defined as overage for grade if he or she turned 15 before the start of ninth grade.

[b]The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form A). No statistical tests or arithmetic operations were performed on these reference points.

overage for grade, pretest scores, and research status. These analyses yield impact estimates that are similar to those presented in the text of the report.

Characteristics of Students Who Completed the Follow-Up Test and Survey

The random assignment research design ensures that there are no systematic differences in measured and unmeasured characteristics between the students in the sample who were assigned to the ERO group and those who were not. Because the two groups began the study with equivalent characteristics, any differences that emerge after random assignment can be attributed with confidence to the fact that one group had access to the ERO programs and the other did not.

When completion rates for follow-up data collection are less than 100 percent, a key question underlying the impact analyses is: Do the response rates preserve the random assignment design? In other words, does the sample of students who completed the follow-up test and survey exhibit the same lack of systematic differences between the ERO and non-ERO groups, both overall and for groups of sites using RAAL and Xtreme Reading? To address this question, multiple regression was used to assess whether there are systematic differences in background

The Enhanced Reading Opportunities Study

Appendix Table B.4

Characteristics of Students in Cohort 2: Differences Between Respondents and Nonrespondents, Xtreme Reading Schools

Characteristic	Respondents	Non-Respondents	Difference	P-Value for the Difference
Race/ethnicity (%)				
Hispanic	31.7	28.2	3.5	0.116
Black, non-Hispanic	45.6	48.6	-2.9	0.259
White, non-Hispanic	16.5	14.3	2.2	0.320
Other	6.2	9.0	-2.8	0.107
Gender (%)				
Male	51.5	47.7	3.8	0.294
Female	48.5	52.3	-3.8	0.294
Average age (years)	14.7	14.9	-0.2 *	0.000
Overage for grade[a] (%)	24.5	38.8	-14.2 *	0.000
Language other than English spoken at home (%)	51.0	48.7	2.3	0.461
Language spoken at home missing (%)	1.2	4.6	-3.4 *	0.001
Mother's education level (%)				
Did not finish high school	16.7	16.3	0.4	0.875
High school diploma or GED certificate	23.8	23.1	0.7	0.816
Completed some postsecondary education	34.3	29.9	4.4	0.184
Don't know	23.7	25.2	-1.5	0.625
Missing	1.6	5.6	-4.0 *	0.000
Father's education level (%)				
Did not finish high school	17.0	14.3	2.7	0.308
High school diploma or GED certificate	20.5	22.5	-2.0	0.493
Completed some postsecondary education	22.8	18.1	4.7	0.111
Don't know	37.6	40.0	-2.4	0.498
Missing	2.2	5.2	-3.0 *	0.012
GRADE reading comprehension[b]				
Average standard score	84.7	83.8	0.8 *	0.042
Corresponding grade equivalent	*4.9*	*4.7*		
Corresponding percentile	*14*	*13*		
2.0 - 3.0 years below grade level (%)	31.1	27.3	3.7	0.251
3.1 - 4.0 years below grade level (%)	30.0	27.1	2.9	0.371
4.1 - 5.0 years below grade level (%)	38.9	45.6	-6.7	0.057
Sample size	1,056	246		

(continued)

Appendix Table B.4 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study baseline data.

NOTES: Baseline data for students in 30 of 34 high schools were collected in spring 2006 (during students' eighth-grade year); baseline data for students in the remaining four schools were collected in fall 2006 (at the start of students' ninth-grade year).

The estimated differences are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school. The values in the column labeled "Respondents" are the observed means for students in the study sample with follow-up data on the GRADE assessment (respondents). The "Non-Respondents" values in the next column are the regression-adjusted means for students in the study sample without follow-up data on the GRADE assessment (non-respondents), using the observed distribution of respondents across random assignment blocks (i.e., schools) as the basis for the adjustment.

A two-tailed t-test was used to test differences between the respondents and non-respondents. The statistical significance level is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

[a]A student is defined as overage for grade if he or she turned 15 before the start of ninth grade.

[b]The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form A). No statistical tests or arithmetic operations were performed on these reference points.

characteristics between the ERO and non-ERO groups. The results are presented in Appendix Table B.6. An overall F-test indicates that there is no systematic difference between the two groups either overall or for the RAAL or Xtreme Reading schools.

Comparisons of students in the ERO and non-ERO groups can also be found in Chapter 2. These comparisons are displayed in Table 2.4 for all 34 high schools in the study, in Table 2.5 for the RAAL schools, and in Table 2.6 for the Xtreme Reading schools. Each of these tables indicates a high degree of similarity between students in the ERO and non-ERO groups.

In summary, the follow-up test and survey completion rates preserve the random assignment design for the ERO study in terms of the characteristics of students measured at baseline. As a result, one may have a high degree of confidence that any differences found in the follow-up data reflect the impact of the ERO programs.

The Enhanced Reading Opportunities Study

Appendix Table B.5

Regression Coefficients for the Probability of Being in the Respondent Sample, Full Study Sample

Variable	Parameter Estimates (Standard Errors) All Schools	Reading Apprenticeship Schools	Xtreme Reading Schools
Intercept	1.360 *	1.319 *	1.370 *
	(0.322)	(0.442)	(0.470)
School 1	-0.123	-0.099	
	(0.072)	(0.093)	
School 2	-0.072	-0.033	
	(0.071)	(0.093)	
School 3	-0.127		-0.096
	(0.070)		(0.072)
School 4	-0.028		-0.005
	(0.070)		(0.072)
School 5	-0.229 *		-0.193 *
	(0.067)		(0.071)
School 6	-0.137 *		-0.098
	(0.068)		(0.071)
School 7	-0.242 *	-0.229 *	
	(0.065)	(0.090)	
School 8	-0.120	-0.107	
	(0.068)	(0.092)	
School 9	-0.046		-0.049
	(0.085)		(0.086)
School 10[a]	-0.040	--	
	(0.095)	--	
School 11	-0.127	-0.111	
	(0.069)	(0.091)	
School 12[a]	--		--
	--		--
School 13	-0.138		-0.116
	(0.073)		(0.077)
School 14	-0.214 *		-0.183 *
	(0.076)		(0.080)
School 15	-0.079	-0.055	
	(0.070)	(0.093)	
School 16	-0.159 *	-0.132	
	(0.079)	(0.099)	
School 17	-0.106	-0.074	
	(0.069)	(0.092)	
School 18	-0.137		-0.095
	(0.072)		(0.075)

(continued)

Appendix Table B.5 (continued)

Variable	Parameter Estimates (Standard Errors) All Schools	Reading Apprenticeship Schools	Xtreme Reading Schools
School 19	-0.104		-0.080
	(0.069)		(0.072)
School 20	-0.011	0.005	
	(0.074)	(0.096)	
School 21	-0.111	-0.067	
	(0.067)	(0.089)	
School 22	-0.170 *		-0.150 *
	(0.068)		(0.071)
School 23	-0.098		-0.076
	(0.066)		(0.067)
School 24	-0.098	-0.062	
	(0.068)	(0.090)	
School 25	-0.075	-0.058	
	(0.076)	(0.097)	
School 26	-0.040		-0.004
	(0.068)		(0.071)
School 27	-0.112	-0.087	
	(0.071)	(0.094)	
School 28	-0.096	-0.071	
	(0.069)	(0.092)	
School 29	-0.143 *		-0.101
	(0.070)		(0.074)
School 30	-0.065		-0.042
	(0.070)		(0.074)
School 31	-0.072	-0.060	
	(0.064)	(0.088)	
School 32	-0.163 *		-0.136
	(0.069)		(0.070)
School 33	-0.190 *	-0.173	
	(0.065)	(0.088)	
School 34	-0.157 *		-0.125
	(0.070)		(0.072)
Research status			
ERO group	0.035 *	0.023	0.046 *
	(0.015)	(0.021)	(0.022)
Non-ERO group[a]	--	--	--
	--	--	--

(continued)

Appendix Table B.5 (continued)

Variable	Parameter Estimates (Standard Errors) All Schools	Reading Apprenticeship Schools	Xtreme Reading Schools
Race/ethnicity (%)			
Hispanic	0.091 *	0.151 *	0.026
	(0.031)	(0.042)	(0.044)
Black, non-Hispanic	0.006	0.059	-0.048
	(0.026)	(0.037)	(0.038)
White, non-Hispanic[a]	--	--	--
	--	--	--
Other	0.015	0.111 *	-0.088
	(0.036)	(0.050)	(0.051)
Gender (%)			
Male	0.026	0.016	0.039
	(0.015)	(0.021)	(0.022)
Female[a]	--	--	--
	--	--	--
Average age (years)	-0.059 *	-0.069 *	-0.050
	(0.020)	(0.028)	(0.030)
Overage for grade[b] (%)	-0.038	-0.021	-0.059
	(0.028)	(0.039)	(0.041)
Language other than English spoken at home (%)	-0.021	-0.035	-0.006
	(0.018)	(0.025)	(0.026)
Language spoken at home missing (%)	-0.267 *	-0.357 *	-0.131
	(0.083)	(0.108)	(0.130)
Mother's education level (%)			
Did not finish high school[a]	--	--	--
	--	--	--
High school diploma or GED certificate	0.033	0.056	0.003
	(0.024)	(0.033)	(0.036)
Completed some postsecondary education	0.035	0.063	0.000
	(0.025)	(0.034)	(0.037)
Don't know	0.041	0.090 *	-0.014
	(0.026)	(0.036)	(0.039)
Missing	-0.077	0.061	-0.258 *
	(0.087)	(0.118)	(0.130)
Father's education level (%)			
Did not finish high school[a]	--	--	--
	--	--	--
High school diploma or GED certificate	0.015	0.075 *	-0.044
	(0.026)	(0.035)	(0.037)
Completed some postsecondary education	0.038	0.068	0.008
	(0.027)	(0.038)	(0.039)
Don't know	-0.005	0.008	-0.021
	(0.025)	(0.035)	(0.035)
Missing	0.064	0.030	0.091
	(0.071)	(0.094)	(0.111)

(continued)

Appendix Table B.5 (continued)

Variable	Parameter Estimates (Standard Errors) All Schools	Reading Apprenticeship Schools	Xtreme Reading Schools
GRADE reading comprehension			
Average standard score	0.004 *	0.005 *	0.003
	(0.001)	(0.002)	(0.002)
Sample size	2,679	1,377	1,302
Degrees of freedom	51	34	34
Mean of the dependent variable	0.810	0.810	0.811
R-square	0.080	0.099	0.078
F-statistic	4.492	4.331	3.161
P-value of F-statistic	0.000	0.000	0.000

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study baseline data.

NOTES: Baseline data for students in 30 of 34 high schools were collected in spring 2006 (during students' eighth-grade year); baseline data for students in the remaining four schools were collected in fall 2006 (at the start of students' ninth-grade year).

The statistical significance level is indicated (*) when the p-value is less than or equal to 5 percent

[a]Covariates marked by "--" were not included in the regression. The site with the highest response rate was not included.

[b]A student is defined as overage for grade if he or she turned 15 before the start of the ninth grade.

The Enhanced Reading Opportunities Study

Appendix Table B.6

Regression Coefficients for the Probability of Being in the Treatment Group, Respondent Sample

Variable	Parameter Estimates (Standard Errors) All Schools	Reading Apprenticeship Schools	Xtreme Reading Schools
Intercept	0.702	1.216	0.279
	(0.480)	(0.673)	(0.690)
School 1	0.012	-0.130	
	(0.099)	(0.130)	
School 2	0.044	-0.107	
	(0.097)	(0.129)	
School 3	0.004		-0.004
	(0.096)		(0.098)
School 4	0.003		0.011
	(0.095)		(0.096)
School 5	0.022		0.006
	(0.096)		(0.101)
School 6	0.018		-0.001
	(0.095)		(0.099)
School 7	0.047	-0.086	
	(0.093)	(0.128)	
School 8	0.022	-0.105	
	(0.096)	(0.130)	
School 9	0.157		0.163
	(0.115)		(0.115)
School 10[a]	0.142	--	
	(0.129)	--	
School 11	0.034	-0.102	
	(0.096)	(0.129)	
School 12[a]	--		--
	--		--
School 13	0.152		0.135
	(0.101)		(0.105)
School 14	0.071		0.047
	(0.109)		(0.114)
School 15	-0.027	-0.156	
	(0.096)	(0.129)	
School 16	0.177	0.041	
	(0.112)	(0.141)	
School 17	0.011	-0.118	
	(0.097)	(0.129)	
School 18	0.018		0.007
	(0.102)		(0.105)

(continued)

Appendix Table B.6 (continued)

Variable	Parameter Estimates (Standard Errors) All Schools	Reading Apprenticeship Schools	Xtreme Reading Schools
School 19	0.109		0.094
	(0.097)		(0.099)
School 20	0.072	-0.060	
	(0.101)	(0.134)	
School 21	0.012	-0.126	
	(0.092)	(0.125)	
School 22	0.049		0.026
	(0.094)		(0.097)
School 23	0.007		-0.003
	(0.090)		(0.091)
School 24	0.007	-0.117	
	(0.092)	(0.126)	
School 25	0.078	-0.066	
	(0.105)	(0.136)	
School 26	0.048		0.035
	(0.093)		(0.096)
School 27	0.082	-0.051	
	(0.097)	(0.131)	
School 28	0.033	-0.095	
	(0.093)	(0.128)	
School 29	0.104		0.079
	(0.097)		(0.102)
School 30	0.014		-0.008
	(0.095)		(0.100)
School 31	-0.005	-0.138	
	(0.088)	(0.123)	
School 32	-0.011		-0.018
	(0.096)		(0.098)
School 33	-0.020	-0.169	
	(0.092)	(0.126)	
School 34	0.028		0.018
	(0.098)		(0.101)
Race/ethnicity (%)			
Hispanic	-0.026	-0.029	-0.020
	(0.044)	(0.062)	(0.063)
Black, non-Hispanic	-0.049	-0.066	-0.025
	(0.038)	(0.055)	(0.054)
White, non-Hispanic[a]	--	--	--
	--	--	--
Other	-0.048	-0.015	-0.080
	(0.052)	(0.074)	(0.075)

(continued)

Appendix Table B.6 (continued)

Variable	Parameter Estimates (Standard Errors) All Schools		Reading Apprenticeship Schools		Xtreme Reading Schools	
Gender (%)						
Male	-0.036		-0.028		-0.046	
	(0.022)		(0.031)		(0.031)	
Female[a]	--		--		--	
	--		--		--	
Average age (years)	0.011		-0.016		0.043	
	(0.031)		(0.043)		(0.044)	
Overage for grade[b] (%)	-0.036		0.008		-0.086	
	(0.042)		(0.058)		(0.060)	
Language other than English spoken at home (%)	-0.024		-0.048		0.012	
	(0.026)		(0.037)		(0.038)	
Home language missing (%)	-0.118		0.022		-0.259	
	(0.138)		(0.200)		(0.192)	
Mother's education level (%)						
Did not finish high school[a]	--		--		--	
	--		--		--	
High school diploma or GED certificate	-0.074	*	-0.056		-0.093	
	(0.036)		(0.049)		(0.053)	
Completed some postsecondary education	-0.057		-0.034		-0.081	
	(0.036)		(0.050)		(0.053)	
Don't know	0.000		-0.001		0.006	
	(0.038)		(0.053)		(0.055)	
Missing	0.134		-0.014		0.288	
	(0.132)		(0.181)		(0.196)	
Father's education level (%)						
Did not finish high school[a]	--		--		--	
	--		--		--	
High school diploma or GED certificate	0.141	*	0.091		0.191	*
	(0.037)		(0.053)		(0.053)	
Completed some postsecondary education	0.094	*	0.060		0.130	*
	(0.040)		(0.057)		(0.055)	
Don't know	0.049		-0.005		0.102	*
	(0.036)		(0.052)		(0.050)	
Missing	0.104		0.067		0.150	
	(0.103)		(0.140)		(0.156)	
GRADE reading comprehension						
Average standard score	-0.003		-0.003		-0.004	
	(0.002)		(0.003)		(0.003)	
Sample size	2,171		1,115		1,056	
Degrees of freedom	50		33		33	
Mean of the dependent variable	0.582		0.578		0.586	
R-square	0.023		0.018		0.038	
F-statistic	0.982		0.601		1.210	
P-value of F-statistic	0.511		0.964		0.194	

(continued)

Appendix Table B.6 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study baseline data.

NOTES: Baseline data for students in 30 of 34 high schools were collected in spring 2006 (during students' eighth-grade year); baseline data for students in the remaining four schools were collected in fall 2006 (at the start of students' ninth-grade year).

The statistical significance level is indicated (*) when the p-value is less than or equal to 5 percent

[a]Covariates marked by "--" were not included in the regression. The site with the highest response rate was not included.

[b]A student is defined as overage for grade if he or she turned 15 before the start of the ninth grade.

Appendix C

Statistical Power and Minimum Detectable Effect Size

This appendix reviews the statistical-power analysis that was conducted during the design phase of the study to determine an acceptable level of precision when estimating the impact of the literacy programs in the Enhanced Reading Opportunities (ERO) study. Specifically, it reviews how the sample configuration, use of regression covariates, and other analytic assumptions would affect the precision of the impact estimates. The discussion focuses on achievement test score outcomes because of their prominence in the study.

The discussion that follows reports precision as "minimum detectable effect sizes" (MDES). Intuitively, a minimum detectable effect is the smallest program impact that could be estimated with confidence, given random sampling and estimation error.[1] This metric, which is used widely for measuring the impacts of educational programs, is defined in terms of the underlying population's standard deviation of student achievement. For example, an MDES of 0.20 indicates that an impact estimator can reliably detect a program-induced increase in student achievement that is equal to or greater than 0.20 standard deviation of the existing student distribution. This is equivalent to approximately four Normal Curve Equivalent (NCE) points on a nationally norm-referenced achievement test and translates roughly into the difference between the 25th and the 31st percentiles.

Unfortunately, there is no definitive standard for a policy-relevant or cost-effective MDES. A meta-analysis of treatment effectiveness studies sheds some light on this issue.[2] This study found that, out of 102 studies, most of which were from education research, the bottom third of the distribution of impacts ranged from about 0 to 0.32 effect size; the middle third of impacts ranged from 0.33 to 0.50; and the top third of impacts ranged from 0.56 to 1.26. Under these "rules of thumb," an MDES of 0.32 would be considered small. More recent work by Bloom et al. suggests that a 0.32 MDES would be considered quite large when placed in the context of the growth in test scores expected over the course of a full year of schooling. Based on data from many of the most widely used standardized reading tests, they find that the expected growth in reading for ninth-grade students ranges from a 0.11 effect size to a 0.26 effect size for a full year of school.[3] Documentation for the Group Reading Assessment and Diagnostic Examination (GRADE) assessment that is being used for the ERO study indicates that the expected growth for ninth-grade students is equivalent to approximately a 0.07 effect size.

The ERO impact study was designed to allow an MDES of approximately 0.06 for the full sample of schools in the study and an MDES of approximately 0.10 for the groups of schools using each of the ERO program models. The MDES estimates for the ERO study de-

[1]A minimum detectable effect is defined as the smallest true program impact that would have an 80 percent chance of being detected (have 80 percent power) using a two-tail hypothesis test at the 5 percent level of statistical significance.

[2]Lipsey (1990).

[3]Bloom, Hill, Black, and Lipsey (2006).

sign accounted for both within-site and across-site variation in the outcome in question. They also accounted for random differences between the program and control groups by including pre-random assignment reading test scores. Finally, the minimum detectable effect sizes presented in the study design were assumed to be fixed-effect estimates; that is, they did not account for variation across sites in the true impact of the program.[4] This final assumption was justified by the fact that sites for the study were to be selected purposefully. Statistically, therefore, the results reflect the impact for the particular sample of schools in the study and should not be generalized to a broader population of similar schools.

Appendix Table C.1 shows the sample sizes resulting from various configurations of schools and student subgroups. The upper panel shows sample sizes in the ideal case that follow-up data would be available for all students in the sample. The lower panel shows sample sizes in cases where those follow-up data would be available for 80 percent of the students in the sample. Each row in the exhibit shows the sample sizes for various groupings of schools. Each column in the table shows sample sizes for potential subgroups of the targeted number of students that the study aimed to include.

---

[4]Minimum detectable effect sizes were estimated as follows:

$$MDES = 2.8 * \sqrt{\frac{\sigma_y^2(1-R^2)}{P(1-P)(n)(J)(\sigma_y^2+\tau_y^2)} + \frac{\omega^2}{J(\sigma_y^2+\tau_y^2)}},$$

Where:

σ_y^2 = the (within-site) variance of the outcome in question (assumed to be 1; however, by definition of effect-size metric, does not affect the MDES).

R^2 = the explanatory power of the impact regression adjusted for pre-random assignment characteristics, that is, the proportion of the variance in *y* explained by the experiment and any pre-random assignment characteristics. In order to determine an appropriate r-square, MDRC regressed ninth-grade SAT-9 achievement on eighth-grade scores for high school students in the Houston school district in 2002. The regression produced an r-square value of 0.69, which is used in this report's calculations of effect size.

P = the proportion of students randomly assigned to the treatment group (assumed to be 0.55, based on the random assignment design for this study).

n = the number of students in each site (as listed in Appendix Table C.1).

J = the number of sites in the study (as listed in Appendix Table C.1).

τ_y^2 = the cross-site variance in the mean value of the outcome measure *y* and calculated as 0.08 (based on an assumption that the intraclass correlation $\frac{\tau^2}{\tau^2+\sigma^2}$ = 0.07, an assumption based on MDRC's analysis of achievement data across all comprehensive nonexclusive high schools in the Houston school district).

ω^2 = the cross-site variance in the true impact of the program. The minimum detectable effect sizes presented here are calculated as fixed-effects estimates; that is, they do not account for cross-site variation in the true impact of the program. Thus, ω^2 is assumed to be zero.

There are 34 schools in the ERO study sample. Initially, the study aimed to identify approximately 110 students for each of two cohorts of ninth-graders who would be eligible and appropriate for the ERO program. Of these, 60 students would be randomly assigned to enroll in the ERO classes, and the remaining 50 students would constitute the control group. Under these assumptions, the target sample for the second cohort of students in the ERO study was a total of 3,740 students. As discussed in Chapter 2, the actual sample for the second cohort was 2,679 students. This is closer to the sample displayed in the second column of numbers in Appendix Table C.1, which is highlighted to reflect the fact that most of the discussion focuses on the MDES estimates for this sample.

The two remaining columns in Appendix Table C.1 show sample sizes for subgroups comprising 50 percent of the target sample and 25 percent of the sample. The 25 percent subgroup (935 students), for example, is somewhat smaller than the actual number of students in the second cohort with baseline test scores that were between the fourth- and fifth-grade levels (1,092 students.)

The second row of numbers in Appendix Table C.1 shows sample sizes for a subgroup of 17 schools reflecting the groups using each of the two supplemental literacy programs. It shows that the target sample for each ERO program was 1,870 students. In fact, the second cohort includes 1,377 students from the 17 schools using Reading Apprenticeship Academic Literacy (RAAL) and 1,302 students from the 17 schools using Xtreme Reading. These samples are closer to those shown in the second column of numbers in Appendix Table C.1. The third and fourth rows show the sample sizes for smaller subgroups of schools — for example, if the schools within each of the programs were split into two groups (approximately eight schools each) or if there were to be district-level analyses (seven of the 10 participating districts had four schools each).

The bottom panel of Appendix Table C.1 shows sample sizes that would result from follow-up data collection from 80 percent of the students in the original sample. As discussed in Chapter 2, approximately 81 percent of the students in the study sample completed the follow-up test, for a respondent analysis sample of 2,171 students. The resulting samples sizes are closest to those shown in the second column of numbers in Appendix Table C.1.

Appendix Table C.2 shows how minimum detectable effect sizes for average reading achievement scores would vary among sample sizes associated with various configurations of sites and student subgroups. Again, as noted above, the highlighted column for 75 percent of the target sample closely approximates the minimum detectable effect sizes for the second cohort of students in the study sample. The discussion now turns to the study's key impact questions.

What is the impact of supplemental literacy interventions of the type that were selected on students' reading achievement?

Analyses that address this question rely on the full sample of students across all 34 participating high schools. The second column of numbers in the bottom panel of Appendix Table C.2 indicates that the MDES for this sample would be 0.06 standard deviation if the follow-up data collection effort achieved at least an 80 percent response rate.

What is the impact of each supplemental literacy intervention on students' reading achievement?

Analyses that address this question rely on the sample of students from 17 of the 34 participating high schools. The second column of the bottom panel of Appendix Table C.2 indicates that the MDES for this sample would be 0.09 standard deviation if the follow-up data collection effort achieved at least an 80 percent response rate.

What is the impact of each supplemental literacy intervention on reading achievement for important subgroups of students or sites?

In addition to questions regarding effects for the full sample of students and for students in high schools implementing each literacy intervention, the evaluation was designed to allow for the estimation of impacts for subgroups of students defined by pre-random assignment characteristics, including baseline reading test scores, whether students had been retained in a prior grade, and English language-learning status.

The rightmost column in Appendix Table C.2 presents the estimated minimum detectable effect sizes for subgroups of students that would comprise at least one-quarter of the intended sample and approximately one-third of the actual sample. For example, students with especially low baseline test scores (between the fourth- and fifth-grade levels) comprise a little over one-third of the actual sample. The MDES for a subgroup that is one-third of the actual sample (approximately 935 students) would be 0.11 standard deviation unit for analyses that include all 34 high schools and 0.16 for analyses that focus only on the 17 schools using one or the other of the two supplemental literacy programs.

The Enhanced Reading Opportunities Study

Appendix Table C.1

Sample Sizes, by Site and Student Subgroup Configuration, for Full Sample and 80 Percent Subsample

100 Percent Response Rate Sample Size				
Number of Schools	Target Sample	75 Percent of Target Sample	50 Percent of Target Sample	25 Percent of Target Sample
34	3,740	2,805	1,870	935
17	1,870	1,403	935	468
8	880	660	440	220
4	440	330	220	110
80 Percent Response Rate Sample Size				
Number of Schools	Target Sample	75 Percent of Target Sample	50 Percent of Target Sample	25 Percent of Target Sample
34	2,992	2,244	1,496	748
17	1,496	1,122	748	374
8	704	528	352	176
4	352	264	176	88

The Enhanced Reading Opportunities Study

Appendix Table C.2

Minimum Detectable Effect Sizes, by Site and Student Subgroup Configuration, for Full Sample and 80 Percent Subsample

100 Percent Response Rate Minimum Detectable Effect Size				
Number of Schools	Target Sample	75 Percent of Target Sample	50 Percent of Target Sample	25 Percent of Target Sample
34	0.05	0.06	0.07	0.10
17	0.07	0.08	0.10	0.14
8	0.10	0.12	0.14	0.20
4	0.14	0.17	0.20	0.29
80 Percent Response Rate Minimum Detectable Effect Size				
Number of Schools	Target Sample	75 Percent of Target Sample	50 Percent of Target Sample	25 Percent of Target Sample
34	0.06	0.06	0.08	0.11
17	0.08	0.09	0.11	0.16
8	0.11	0.13	0.16	0.23
4	0.16	0.19	0.23	0.32

Appendix D

ERO Implementation Fidelity

This appendix describes the development of measures based on the classroom observation data collected during site visits to the high schools in the Enhanced Reading Opportunities (ERO) study. The analysis of ERO program implementation fidelity in the second year of the study is based on field research visits to each of the 34 high schools during the fall of 2006 and the spring of 2007. The primary data collection instrument for the site visits was a set of protocols for classroom observations and interviews with the ERO teachers. The observation protocols provided a structured process for trained classroom observers to rate characteristics of the ERO classroom learning environments and the ERO teachers' instructional strategies. All of these characteristics (referred to as "constructs") were selected for assessment because they were aligned with program elements specified by the developers and, by design, were aligned with supplemental literacy program elements that are believed to characterize high-quality interventions for struggling adolescent readers.[1] The instrument included ratings for six general instructional constructs that are common to both literacy interventions — Reading Apprenticeship Academic Literacy (RAAL) and Xtreme Reading — and ratings for seven program-specific constructs for each of them. The program-specific constructs reflect the distinctive components of the two literacy programs and are designated with program-specific terminology. (The observation protocols are included at the end of this appendix.)

Before conducting the classroom observation visits for Year 1 of the study, observers — who were research employees of the American Institutes for Research (AIR) and MDRC who had worked previously on at least one project involving site visits — had attended a two-day training to learn about the program designs and their intended implementation strategies and to learn and practice how to use the protocols. A refresher training was provided before the Year 2 site visits, to give the observers more practice using the protocols and to address any challenges that may have arisen during Year 1. The classroom observations in Year 2 were conducted by one researcher per school district (a senior staff member with at least a master's degree) and captured between 160 and 180 minutes of instruction in each of the 34 high schools during each visit. The amount of observation time in each school ranged from at least two ERO classes (in schools with 80- to 90-minute class periods) and up to four ERO classes (in schools with 45-minute class periods).

Site visits were scheduled with the intent of observing classrooms across schools after similar amounts of instructional time had passed. On average, the fall observations occurred 15 weeks after the ERO classes started, and the spring observations took place about 16 weeks after the fall observations. The fall observations occurred at a point in time when teachers had gained some experience with the curriculum and with the teaching of the ERO programs. The spring observations occurred after the teachers had received their booster trainings with the de-

[1]Biancarosa and Snow (2004).

velopers and at a point when the teachers had covered much of the curriculum. The fact that the measurement of implementation fidelity in the second year of the study is based on two sets of classroom observations — unlike in the first year of the study, when observations from only one site visit were used — also means that the fidelity measures in Year 2 capture a fuller range of teachers' experiences with the programs, which can be used to depict changes in implementation fidelity over the course of the school year.

During the visits to a given school, the observer took detailed field notes, focusing on teachers' presentation of curriculum components, the flow of instruction, students' behavior and engagement, and teacher-student interactions. The observer then gave a summative rating across all the observed classes in the school (ranging from two to four classes), for each of six common program constructs (used in the observations for both programs) and for each of the seven program-specific constructs (with different constructs used in observations of RAAL and Xtreme Reading). The rating for each construct was accompanied by a justification statement tying the observed behaviors and activities to the descriptions of the expected behaviors and activities that were used to guide the observations. The ratings from all the site visits were then reviewed centrally by at least two senior members of the study team, who checked that the justifications for the ratings were grounded in the types of evidence called for in the observation protocols.

The observers used a three-category rating format for each of the general and program-specific constructs.[2] Although each construct was rated using criteria that were specific to that construct, the following provides a general description of the principles that were embedded in each of the three rating categories.

- **Category 3.** For each construct, classes that fell into this category included teacher behaviors and classroom activities that were well developed and highly consistent in their alignment with the intended behaviors and activities specified by the developers and described in the protocol. In these classes, teachers demonstrated confidence in what they were teaching, conveyed a thorough understanding of what was being taught conceptually and procedurally, were familiar with any materials needed, and were able to interact proactively with students who asked questions or experienced difficulty. Students appeared to be engaged in the instruction and demonstrated learning behaviors that went beyond rote performance. Teachers who fell into this category took advantage of opportunities to connect instruction to a sponta-

[2]In some cases, a rating of "not applicable" was used to show that the construct was not observed at all during the site visit. Two situations may have necessitated the need for this rating. First, the lesson being taught on the day of the observation did not call for attention to the construct. Second, opportunities to address a particular construct did not arise during the course of the class. Constructs with a "not applicable" rating were treated as missing data and were not given a numeric value.

neous event or interaction in class ("a teachable moment"). If students worked independently during some of the class, they were engaged and seemed to understand the purpose of and procedures for their activity.

- **Category 2.** For each construct, classes that fell into this category included observed teacher behaviors and classroom activities that were at least moderately aligned with the behaviors and activities specified by the developers and described in the protocols. Teachers demonstrated more than a basic understanding of what they were teaching but might not have taken full advantage of opportunities to use program materials, to capitalize on "teachable moments," or to explain fully a strategy or concept. In these classes, students, while generally attending to the instruction or task at hand, did not appear intellectually engaged, and some may have been inattentive or confused.

- **Category 1.** For each construct, classes that fell into this category were not aligned with the behaviors and activities specified by the developers and described in the protocols. Teachers may have neglected opportunities to teach, may have paid only limited attention to an aspect of the program, and may not have been responsive to students' confusion or questions. In these classes, students were sporadically engaged in the lesson, and some students may have been acting in a disruptive fashion.

There are five ways in which the study team sought reliable ratings across site visits. First, all observers were trained together to promote a common understanding of the observation process. Second, site visits were conducted by senior study team members, all of whom participated in the first year of site visits and were thoroughly trained on the observation instrument over the course of the two years of the study. Third, although a given observer conducted all observations in all of the participating high schools in a school district, the observers varied across districts, thus limiting the potential for the development of particularistic understandings by a given observer of how to rate the constructs. Fourth, the summative ratings from all the site visits were reviewed centrally by senior members of the study team, who checked that the justifications for the ratings were grounded in the types of evidence called for in the observation protocols. If the reviewers questioned a rating, the observer and reviewers reached a decision on keeping or changing the rating based on review of the observation data. Last, all of the site observers met as a group during the site visits to discuss the rating process and reinforce a common understanding of the relationship between the rating scale and the constructs.

Measuring the Classroom Learning Environment

As discussed in Chapter 3, the measurement of implementation fidelity focused on two key dimensions of implementation: learning environment and comprehension instruction. Ratings for the constructs were combined to calculate composite measures for each of these two key dimensions, for each of the two site visits. This section of the appendix describes how the composite measure of the learning environment dimension was calculated. Because the reliability of these constructs (Cronbach's alpha) is similar across the fall and spring site visits, the reliability is reported for the spring site visit only.

Learning Environment Composite

(2 items, Cronbach's alpha = .86)

This measure was designed to measure the extent to which ERO classrooms represented learning environments believed to be conducive to the effective delivery of the core instructional strategies by the teacher and the facilitation of student and teacher interactions around the reading skills that were being taught and practiced. It was created by averaging a general instructional component measured at all 34 ERO high schools and a program-specific component measured at each set of 17 schools implementing each program.

General Instructional Learning Environment Component

(2 items, Cronbach's alpha = .69)

This component is the average of two observed constructs that are part of the general instructional scales: classroom climate and on-task participation.[3]

Program-Specific Learning Environment Components

Reading Apprenticeship Academic Literacy (1 item, Cronbach's alpha = NA)

The program-specific component of the learning environment composite for RAAL schools is a single construct: social reading community. Thus the calculation of a Cronbach's Alpha is not applicable (NA).

[3]In the observation protocols, "motivation and student engagement" is used to describe both a general instructional construct and an Xtreme Reading-specific construct. In this discussion and in Table 3.4, the general instructional construct has been renamed "on-task participation" to distinguish it more clearly from the program-specific construct, still referred to as "motivation and student engagement."

Xtreme Reading (2 items, Cronbach's alpha = .88)

The program-specific component of the learning environment composite for Xtreme Reading schools is the average of two constructs: classroom management and motivation and engagement.

Equations D-1 and D-2 (below) show how the constructs and components were combined to calculate the learning environment composite measures for RAAL and Xtreme Reading schools.[4]

$$LE_{RA} = ½ (½ (GIC_1 + GIC_2) + (PSC_{RA1})) \quad \text{(D-1)}$$

Where:

LE_{RA} = learning environment composite measure in a RAAL school
GIC_1 = classroom climate (general instructional construct)
GIC_2 = on-task participation (general instructional construct)
PSC_{RA1} = social reading community (RAAL construct)

$$LE_{XR} = ½ (½ (GIC_1 + GIC_2) + ½ (PSC_{XR1} + PSC_{XR2})) \quad \text{(D-2)}$$

Where:

LE_{XR} = learning environment composite measure in an Xtreme Reading school
GIC_1 = classroom climate (general instructional construct)
GIC_2 = on-task participation (general instructional construct)
PSC_{XR1} = classroom management (Xtreme Reading construct)
PSC_{XR2} = motivation and engagement (Xtreme Reading construct)

[4]In these equations, "LE" stands for learning environment; "RA" and "XR" stand for RAAL and Xtreme Reading, respectively; and "GIC" and "PSC" stand for general instructional construct and program-specific construct, respectively.

Measuring Reading Comprehension Instruction

This section of the appendix describes how the composite measure of the second key implementation dimension, comprehension instruction, was calculated. As above, the reliability is reported for the spring site visit only.

Comprehension Instruction Composite

(2 items, Cronbach's alpha = .79)

This measure was designed to measure the quality of the reading comprehension instruction in each ERO school. As with the learning environment composite measure, it was created by averaging a general instructional component measured at each of the 34 ERO high schools and a program-specific component measured at each school — the RAAL component at each of the 17 RAAL schools and the Xtreme Reading component at each of the 17 Xtreme Reading schools.

General Instructional Comprehension Instruction Component

(2 items, Cronbach's alpha = .70)

This component is the average of two observed constructs that are part of the general instructional scales: comprehension and metacognition.

Program-Specific Comprehension Instruction Components

Reading Apprenticeship Academic Literacy (5 items, Cronbach's alpha = .69)

The program-specific component of the comprehension instruction composite for RAAL schools is the average of five constructs observed at and averaged for each school: metacognitive conversations, silent sustained reading, content/theme integration, writing, and integration of curriculum strands.

Xtreme Reading (2 items, Cronbach's alpha = .63)

The program-specific component of the comprehension instruction composite for Xtreme Reading schools is the average of two constructs: curriculum-driven (or systematic) instruction and needs-driven (or responsive) instruction. The curriculum-driven instruction construct is the average of three subconstructs: structured content, research-based methodology, and connected scaffolded and informed instruction (Cronbach's alpha = .69). The needs-driven instruction construct is the average of two subconstructs: student accommodations and feedback to students (Cronbach's alpha = .51).

Equations D-3 and D-4 (below) show how the constructs and components were combined to calculate the comprehension instruction composite measures for RAAL and Xtreme Reading schools.[5]

$$CI_{RA} = ½ (½ (GIC_1 + GIC_2) + {}^1/_5 (PSC_{RA1} + PSC_{RA2} + PSC_{RA3} + PSC_{RA4} + PSC_{RA5})) \quad \text{(D-3)}$$

Where:

CI_{RA} = comprehension instruction composite measure in a RAAL school
GIC_1 = comprehension (general instructional construct)
GIC_2 = metacognition (general instructional construct)
PSC_{RA1} = metacognitive conversations (RAAL construct)
PSC_{RA2} = silent sustained reading (RAAL construct)
PSC_{RA3} = content/theme integration (RAAL construct)
PSC_{RA4} = writing (RAAL construct)
PSC_{RA5} = integration of curriculum strands (RAAL construct)

$$CI_{XR} = ½ (½ (GIC_1 + GIC_2) + ½ (PSC_{XR1} + PSC_{XR2})) \quad \text{(D-4)}$$

Where:

CI_{XR} = comprehension instruction composite measure in an Xtreme Reading school
GIC_1 = comprehension (general instructional construct)
GIC_2 = metacognition (general instructional construct)
PSC_{XR1} = systematic instruction (Xtreme Reading construct; the average of measures of structured content, research-based methodology, and connected, scaffolded, informed instruction)
PSC_{XR2} = responsive instruction (Xtreme Reading construct; the average of measures of student accommodations and feedback to students)

Categorizing Implementation Fidelity

This section of the appendix discusses briefly how schools were categorized based on the average ratings calculated for each of the 34 participating high schools on the implementation fidelity of their classroom learning environment and for the implementation fidelity of their comprehension instruction. Each average rating ranged between 1 and 3 and was rounded to the nearest tenth of a point. Based on the composite ratings for each of the two program dimensions

[5]In these equations, "CI" stands for comprehension instruction; "RA" and "XR" stand for RAAL and Xtreme Reading, respectively; and "GIC" and "PSC" stand for general instructional construct and program-specific construct, respectively.

— learning environment and comprehension instruction — the implementation fidelity for each dimension was classified as "well aligned," "moderately aligned," or "poorly aligned" to the models specified by the program developers.

The purpose of these fidelity groupings was to identify schools where the implementation of one or both of the two key program dimensions was especially problematic and where schools' programs were not an accurate representation of the program models. This was especially important in Year 1, when implementation of the programs in some of the schools was characterized by notable challenges, as discussed in Chapter 3. Although program implementation was less problematic in Year 2, based on observer fidelity ratings from the two site visits, it remains important to identify schools with weak implementation fidelity; thus, the fidelity groupings used in Year 2 are defined in the same way as in Year 1. That said, only one school was considered poorly aligned on one of the two key dimensions by the end of Year 2 (the spring 2007 site visit).

The ranges of average scores used to define each of the three fidelity groupings are described below. Because the purpose of these groupings was to identify schools whose programs were not representative of the intended programs, also presented below is the number of constructs rated in Category 1 (the lowest score that can be assigned) in the set of schools that fell into the relevant grouping.

Well Aligned

Implementation fidelity for the learning environment or comprehension instruction dimensions was characterized as *well aligned* when the average rating across the relevant general and program-specific constructs was 2.0 or higher. That is, the school's ERO program was rated as "moderately aligned" (a Category 2 rating) or "well aligned" (a Category 3 rating) with the program models on all or almost all of the constructs included in that dimension. As it turns out, the set of schools rated as well aligned in Year 2 had no more than one construct for each implementation dimension rated in Category 1 (the lowest score that can be assigned).

Moderately Aligned

The key dimensions were designated as *moderately aligned* in terms of implementation fidelity if the average rating across the general and program-specific constructs used to create the relevant composite was within the range of 1.5 to 1.9. In these cases, the school's ERO program was observed to have some problems with implementation. In terms of the learning environment, the schools rated as moderately aligned in Year 2 had one construct rated in Category 1 (out of three or four constructs used to calculate the composite for RAAL or Xtreme Reading schools, respectively). On the comprehension instruction dimension, schools had three or fewer

constructs rated in Category 1 (out of seven constructs used to calculate the composite score). These schools also met with some implementation success, with half or more of the constructs that make up the dimension being rated as moderately or well aligned with the program models.

Poorly Aligned

The implementation fidelity of key program dimensions in a school was rated as *poorly aligned* when the average composite rating across the general and program-specific constructs fell below 1.5. In schools rated as poorly aligned in Year 2, half or more of the general or program-specific constructs that make up the dimension were rated in Category 1. These programs were the least representative of the activities and practices intended by the respective program developers.

The top two panels of Appendix Tables D.1 and D.2 provide a summary of the number of schools whose composite rating on the classroom learning environment and comprehension instruction dimensions fell into the well-aligned, moderately aligned, and poorly aligned categories of fidelity during the fall and the spring site visit, respectively. These panels are the same as the top two panels of Table 3.5a and 3.5b in Chapter 3. The bottom panel of these two tables clusters schools based on their level of implementation fidelity across both dimensions. This panel clusters the schools into more categories of combined implementation fidelity than the same panel in Tables 3.5a and 3.5b. Appendix Table D.3 presents the distribution of schools across these same categories of implementation fidelity, but based on the average of the fall and spring ratings for the learning environment and comprehension instruction dimensions. Appendix Tables D.4 and D.5 present average implementation composite scores by teachers' experience with the ERO program (that is, for the 25 schools where teachers taught two full years of the ERO program versus the nine schools where there were replacement teachers who taught less than two full years of the program), at the fall and spring site visits in Year 2, respectively. Appendix Tables D.6 and D.7 present the distribution across fidelity categories of the 25 schools where the ERO teacher taught two full years of the program, during the spring site visits in Year 1 and Year 2 of the study, respectively.

The Enhanced Reading Opportunities Study

Appendix Table D.1

Number of ERO Classrooms with Well-, Moderately, or Poorly Aligned Implementation to Program Models on Each Implementation Dimension, by ERO Program — Year 2 Fall Site Visit

Implementation Dimension	All Schools	Reading Apprenticeship Schools	Xtreme Reading Schools
Learning environment			
School average	2.46	2.47	2.44
Well-aligned implementation (composite rating is 2.0 or higher)	30	14	16
Moderately aligned implementation (composite rating is 1.5-1.9)	3	3	0
Poorly aligned implementation (composite rating is less than 1.5)	1	0	1
Comprehension instruction			
School average	2.10	2.10	2.10
Well-aligned implementation (composite rating is 2.0 or higher)	20	11	9
Moderately aligned implementation (composite rating is 1.5-1.9)	11	5	6
Poorly aligned implementation (composite rating is less than 1.5)	3	1	2
Combined dimensions			
School average	2.28	2.28	2.27
Well-aligned implementation (composite rating is 2.0 or	20	11	9
Well-aligned implementation on learning environment instruction only[a]	10	3	7
Well-aligned implementation on comprehension instruction only	0	0	0
Moderately aligned implementation on both dimensions	2	2	0
Poorly aligned implementation on learning environment only	0	0	0
Poorly aligned implementation on comprehension instruction only[a]	2	1	1
Poorly aligned implementation on both dimensions	1	0	1
Sample size	34	17	17

(continued)

Appendix Table D.1 (continued)

SOURCES: MDRC and AIR calculations from classroom observation data.

NOTES: Implementation with a composite score of less than 1.5 for a given dimension was deemed to be at the beginning stages of development. The implementation for these dimensions was designated as poorly aligned with the program models.

Implementation with composite scores between 1.5 and 1.9 for a given dimension exhibited at least moderate development in some areas while being at the begninning stages of development in other areas. The implementation for these dimensions was designated as moderately aligned.

Implementation with scores of 2.0 or higher for a given dimension exhibited well-developed fidelity on several areas and at least moderate development in most other areas. The implementation for these dimensions was designated as well aligned.

[a]One XR school was designated as being well aligned in terms of learning environment and poorly aligned in terms of comprehension instruction. Thus, this school is counted in two rows in the bottom panel of the table.

The Enhanced Reading Opportunities Study

Appendix Table D.2

Number of ERO Classrooms with Well-, Moderately, or Poorly Aligned Implementation to Program Models on Each Implementation Dimension, by ERO Program — Year 2 Spring Site Visit

Implementation Dimension	All Schools	Reading Apprenticeship Schools	Xtreme Reading Schools
Learning environment			
School average	2.46	2.63	2.28
Well-aligned implementation (composite rating is 2.0 or higher)	31	17	14
Moderately aligned implementation (composite rating is 1.5-1.9)	2	0	2
Poorly aligned implementation (composite rating is less than 1.5)	1	0	1
Comprehension instruction			
School average	2.33	2.27	2.38
Well-aligned implementation (composite rating is 2.0 or higher)	28	13	15
Moderately aligned implementation (composite rating is 1.5-1.9)	6	4	2
Poorly aligned implementation (composite rating is less than 1.5)	0	0	0
Combined dimensions			
School average	2.39	2.45	2.33
Well-aligned implementation on both dimensions	26	13	13
Well-aligned implementation on learning environment instruction only	5	4	1
Well-aligned implementation on comprehension instruction only[a]	2	0	2
Moderately aligned implementation on both dimensions	1	0	1
Poorly aligned implementation on learning environment only[a]	1	0	1
Poorly aligned implementation on comprehension instruction only	0	0	0
Poorly aligned implementation on both dimensions	0	0	0
Sample size	34	17	17

(continued)

Appendix Table D.2 (continued)

SOURCES: MDRC and AIR calculations from classroom observation data.

NOTES: Implementation with a composite score of less than 1.5 for a given dimension was deemed to be at the beginning stages of development. The implementation for these dimensions was designated as poorly aligned with the program models.

Implementation with composite scores between 1.5 and 1.9 for a given dimension exhibited at least moderate development in some areas while being at the begninning stages of development in other areas. The implementation for these dimensions was designated as moderately aligned.

Implementation with scores of 2.0 or higher for a given dimension exhibited well-developed fidelity on several areas and at least moderate development in most other areas. The implementation for these dimensions was designated as well aligned.

[a] One XR school was designated as being well aligned in terms of comprehension instruction and poorly aligned in terms of learning environment. Thus, this school is counted in two rows in the bottom panel of the table.

The Enhanced Reading Opportunities Study

Appendix Table D.3

Number of ERO Classrooms with Well-, Moderately, or Poorly Aligned Implementation to Program Models on Each Implementation Dimension, by ERO Program — Year 2 Spring Site and Fall Visits

Implementation Dimension	All Schools	Reading Apprenticeship Schools	Xtreme Reading Schools
Learning environment			
School average	2.46	2.55	2.36
Well-aligned implementation (composite rating is 2.0 or higher)	31	15	16
Moderately aligned implementation (composite rating is 1.5-1.9)	2	2	0
Poorly aligned implementation (composite rating is less than 1.5)	1	0	1
Comprehension instruction			
School average	2.21	2.18	2.24
Well-aligned implementation (composite rating is 2.0 or higher)	23	11	12
Moderately aligned implementation (composite rating is 1.5-1.9)	10	5	5
Poorly aligned implementation (composite rating is less than 1.5)	1	1	0
Combined dimensions			
School average	2.33	2.37	2.30
Well-aligned implementation on both dimensions	23	11	12
Well-aligned implementation on learning environment instruction only	8	4	4
Well-aligned implementation on comprehension instruction only[a]	0	0	0
Moderately aligned implementation on both dimensions	1	1	0
Poorly aligned implementation on learning environment only	1	0	1
Poorly aligned implementation on comprehension instruction only[a]	1	1	0
Poorly aligned implementation on both dimensions	0	0	0
Sample size	34	17	17

(continued)

Appendix Table D.3 (continued)

SOURCES: MDRC and AIR calculations from classroom observation data.

NOTES: Implementation with a composite score of less than 1.5 for a given dimension was deemed to be at the beginning stages of development. The implementation for these dimensions was designated as poorly aligned with the program models.

Implementation with composite scores between 1.5 and 1.9 for a given dimension exhibited at least moderate development in some areas while being at the begninning stages of development in other areas. The implementation for these dimensions was designated as moderately aligned.

Implementation with scores of 2.0 or higher for a given dimension exhibited well-developed fidelity on several areas and at least moderate development in most other areas. The implementation for these dimensions was designated as well aligned.

The Enhanced Reading Opportunities Study

Appendix Table D.4

Average Implementation Composite Scores, by ERO Program — Year 2 Fall

Characteristic	All Schools	Reading Apprenticeship Schools	Xtreme Reading Schools
Learning environment			
Schools with replacement teachers	2.5	2.6	2.5
Schools with teachers who taught two full years	2.4	2.4	2.4
Comprehension instruction			
Schools with replacement teachers	2.2	2.2	2.1
Schools with teachers who taught two full years	2.1	2.1	2.1
Combined dimensions			
Schools with replacement teachers	2.4	2.4	2.3
Schools with teachers who taught two full years	2.2	2.3	2.2

SOURCES: MDRC and AIR calculations from classroom observation data.

NOTES: There were 25 teachers who taught all of Year 1, 13 at Reading Apprenticeship schools and 12 at Xtreme Reading schools. There were 9 replacement teachers, 4 at Reading Apprenticeship schools and 5 at Xtreme Reading schools.

Implementation with a composite score of less than 1.5 for a given dimension was deemed to be at the beginning stages of development. The implementation for these dimensions was designated as poorly aligned with the program models.

Implementation with composite scores between 1.5 and 1.9 for a given dimension exhibited at least moderate development in some areas while being at the begninning stages of development in other areas. The implementation for these dimensions was designated as moderately aligned.

Implementation with scores of 2.0 or higher for a given dimension exhibited well-developed fidelity on several areas and at least moderate development in most other areas. The implementation for these dimensions was designated as well aligned.

The Enhanced Reading Opportunities Study

Appendix Table D.5

Average Implementation Composite Scores, by ERO Program — Year 2 Spring

Characteristic	All Schools	Reading Apprenticeship Schools	Xtreme Reading Schools
Learning environment			
Schools with replacement teachers	2.4	2.6	2.3
Schools with teachers who taught two full years	2.5	2.6	2.3
Comprehension instruction			
Schools with replacement teachers	2.3	2.3	2.2
Schools with teachers who taught two full years	2.3	2.3	2.4
Combined dimensions			
Schools with replacement teachers	2.3	2.5	2.2
Schools with teachers who taught two full years	2.4	2.4	2.4

SOURCES: MDRC and AIR calculations from classroom observation data.

NOTES: There were 25 teachers who taught all of Year 1, 13 at Reading Apprenticeship schools and 12 at Xtreme Reading schools. There were 9 replacement teachers, 4 at Reading Apprenticeship schools and 5 at Xtreme Reading schools.

Implementation with a composite score of less than 1.5 for a given dimension was deemed to be at the beginning stages of development. The implementation for these dimensions was designated as poorly aligned with the program models.

Implementation with composite scores between 1.5 and 1.9 for a given dimension exhibited at least moderate development in some areas while being at the begninning stages of development in other areas. The implementation for these dimensions was designated as moderately aligned.

Implementation with scores of 2.0 or higher for a given dimension exhibited well-developed fidelity on several areas and at least moderate development in most other areas. The implementation for these dimensions was designated as well aligned.

The Enhanced Reading Opportunities Study

Appendix Table D.6

Number of ERO Classrooms Taught by Teachers Who Taught Two Full Years with Well-, Moderately, or Poorly Aligned Implementation to Program Models on Each Implementation Dimension, by ERO Program — Year 1 Spring

Implementation Dimension	All Schools	Reading Apprenticeship Schools	Xtreme Reading Schools
Learning environment			
Well-aligned implementation (composite rating is 2.0 or higher)	22	11	11
Moderately aligned implementation (composite rating is 1.5-1.9)	3	2	1
Poorly aligned implementation (composite rating is less than 1.5)	0	0	0
Comprehension instruction			
Well-aligned implementation (composite rating is 2.0 or higher)	13	5	8
Moderately aligned implementation (composite rating is 1.5-1.9)	8	4	4
Poorly aligned implementation (composite rating is less than 1.5)	4	4	0
Combined dimensions			
Well-aligned implementation on both dimensions	13	5	8
Moderately aligned implementation on at least one dimension and moderately or well-aligned implementation on the other dimension	8	4	4
Poorly aligned implementation on at least one dimension	4	4	0
Sample size	25	13	12

SOURCES: MDRC and AIR calculations from classroom observation data.

NOTES: Implementation with a composite score of less than 1.5 for a given dimension was deemed to be at the beginning stages of development. The implementation for these dimensions was designated as poorly aligned with the program models.

Implementation with composite scores between 1.5 and 1.9 for a given dimension exhibited at least moderate development in some areas while being at the begninning stages of development in other areas. The implementation for these dimensions was designated as moderately aligned.

Implementation with scores of 2.0 or higher for a given dimension exhibited well-developed fidelity on several areas and at least moderate development in most other areas. The implementation for these dimensions was designated as well aligned.

The Enhanced Reading Opportunities Study

Appendix Table D.7

Number of ERO Classrooms Taught by Teachers Who Taught Two Full Years with Well-, Moderately, or Poorly Aligned Implementation to Program Models on Each Implementation Dimension, by ERO Program — Year 2 Spring

Implementation Dimension	All Schools	Reading Apprenticeship Schools	Xtreme Reading Schools
Learning environment			
Well-aligned implementation (composite rating is 2.0 or higher)	23	13	10
Moderately aligned implementation (composite rating is 1.5-1.9)	1	0	1
Poorly aligned implementation (composite rating is less than 1.5)	1	0	1
Comprehension instruction			
Well-aligned implementation (composite rating is 2.0 or higher)	22	10	12
Moderately aligned implementation (composite rating is 1.5-1.9)	3	3	0
Poorly aligned implementation (composite rating is less than 1.5)	0	0	0
Combined dimensions			
Well-aligned implementation on both dimensions	20	10	10
Moderately aligned implementation on at least one dimension and moderately or well-aligned implementation on the other dimension	4	3	1
Poorly aligned implementation on at least one dimension	1	0	1
Sample size	25	13	12

SOURCES: MDRC and AIR calculations from classroom observation data.

NOTES: Implementation with a composite score of less than 1.5 for a given dimension was deemed to be at the beginning stages of development. The implementation for these dimensions was designated as poorly aligned with the program models.

Implementation with composite scores between 1.5 and 1.9 for a given dimension exhibited at least moderate development in some areas while being at the begninning stages of development in other areas. The implementation for these dimensions was designated as moderately aligned.

Implementation with scores of 2.0 or higher for a given dimension exhibited well-developed fidelity on several areas and at least moderate development in most other areas. The implementation for these dimensions was designated as well aligned.

Classroom Observation Scales

Enhanced Reading Opportunities Study

American Institutes for Research

November 2006

Table of Contents

Enhanced Reading Opportunities Program

General Instruction Scales

Area of interest	**Basic Literacy Skills (Advanced phonics and decoding, fluency)**

Description

0. Not applicable. During the observed class period(s), students do not demonstrate a need for instruction in basic literacy skills.*
1. During the observed class period(s), instruction does not reflect teacher recognition of a *demonstrated* student need for increased understanding of basic literacy skills. The teacher may not recognize or acknowledge this need for practice of basic literacy skills OR these skills are addressed but in a very cursory manner (e.g., students are told to "sound out" words they don't know).
2. During the observed class period(s), instruction reflects teacher recognition of student difficulty with basic literacy skills; however, instruction is not really well developed. For example, fluency and decoding skills may be practiced in a "skill and drill" manner and never applied to authentic texts. As other examples, instruction may not be differentiated to meet individual student needs, OR the teacher may provide insufficient practice opportunities.
3. During the observed class period(s), instruction reflects teacher recognition of student difficulty with basic literacy skills *and* the instruction is provided in a manner that meets student needs. Such instruction could take several forms. For example, instruction could be differentiated for individual students, OR ample practice opportunities could be provided for those who need it, in order to facilitate increased decoding and fluency abilities, as well as the ability to apply these skills to make meaning of text. This could be evidenced by students learning or applying a systematic approach for decoding unknown words as they read a piece of literature).

*A demonstrated need could be manifested in the form of student difficulties with decoding words, or students reading haltingly or without expression.

Area of interest	**Vocabulary**

Description

0. Not applicable. There was no opportunity for vocabulary instruction to occur during the observed class period(s).

1. Students are engaged in a few vocabulary development activities, but these activities are largely superficial in nature. Vocabulary is not connected to student texts or writing. Such instruction could take the form of rote vocabulary learning methods, OR vocabulary instruction that occurs out of textual context. For example, students may be asked to look up the definitions of words in the dictionary to discover meanings.

2. Students are engaged in some vocabulary activities, but these activities are not fully developed. For example, the teacher may employing definitional and contextual information for presenting words but gives little attention to linking words to prior experiences OR to teaching strategies to help students figure out the meaning of words on their own (e.g. identifying root word, using context clues, etc).

3. Students are engaged in vocabulary instruction that is integrated throughout instruction, and multiple vocabulary strategies are used. Instruction provides students with strategies that help them to independently derive the meaning of unfamiliar words. For example, instruction may focus on using strategies to identify new words and building context for new words and concepts. Repetition and both direct and indirect techniques for teaching vocabulary may be utilized.

Area of interest	**Comprehension**

Description

1. Few opportunities are provided for students to obtain meaning from text, and comprehension strategies are addressed in a basic or superficial manner. For example, the teacher or the students may expend little effort to understand the substance of what is being read. Instruction may not be focused on reading text and meaning-making, or the teacher may do very little modeling and direct instruction of comprehension strategies. The teacher may make little or no efforts to monitor student comprehension of text.
2. Some opportunities are provided for students to try to obtain meaning from text, but comprehension strategies are not fully developed. For example, students may make some attempts to make sense of difficult or unfamiliar text, but they give up easily when they don't understand. As another example, the teacher may make some attempts to model critical thinking strategies, but direct instruction is limited to teaching basic comprehension strategies (e.g., making predictions, identifying main characters and setting, and summarizing, distinguishing between fact and opinion). The teacher may monitor or probe for student comprehension but does not necessarily use this information to target or enhance specific comprehension skills during the class period.
3. There are substantial opportunities and various approaches for students to try to obtain and validate meaning from text. Most students, for most of the time, are trying to derive meaning from the texts that they read and have concrete strategies for doing so. Opportunities for the development of student reading skills could be evidenced by teacher use of modeling and direct instruction to teach strategies and thought processes, and emphasis of critical thinking. The teacher may also encourage or facilitate purposeful student discussion and interaction with text. For example, the teacher may activate students' prior knowledge and encourage higher-order thinking. Instructional content may include components of text structure, both generically and with specific reference to content-area learning. Another example of substantial comprehension instruction could include teacher monitoring or probing for student comprehension, followed by teaching or reflecting on strategies to enhance student comprehension abilities.

Area of interest	**Metacognition**

Description
(Note: In a successful class, this becomes less visible towards the end of the year as students internalize these procedures.)

1. Little metacognitive work is apparent, and overall, metacognitive skills are not being *developed* through instruction or conscious practice. In some cases, students may be taught strategies to monitor their own reading, recognize faulty comprehension, and apply "fix-up" strategies; but these strategies are not explored. For example, the teacher either does not address metacognitive strategies (e.g., self-monitoring of reading may not be taught at all) or does so in a very limited or superficial, contrived manner (e.g., teacher and students are most often "going through the motions").
2. Instruction incorporates some development of metacognitive strategies and opportunities for student practice of metacognition, either through spoken or written expression, but these may not be fully developed. For example, instruction could include the use of "think alouds" to model strategies, self-correct, and make connections to prior knowledge. While some of the metacognitive activities flow naturally, others may appear to be forced (teacher or students appear to be "going through the motions").
3. Use of metacognitive strategies is pervasive and integrated throughout instruction. Instruction includes teacher modeling of strategies and multiple opportunities for student practice of thinking aloud through spoken or written expression with multiple forms of text. Throughout the majority of metacognitive activities, the teacher monitors and guides students in their thought processes. In addition, the majority of the metacognitive activities are conducted in a natural and thoughtful manner.

Area of interest	**Classroom Climate and Social Support for Learning**
Description	

1. The classroom environment seems disrespectful and chaotic. Students interrupt each other and interfere with one another's efforts to learn. For example, students may engage in or experience taunts, occasional threats, or slurs about themselves or backgrounds. The teacher does little, if anything, to counteract these problems. Students have little opportunity to work together (either in pairs or small groups) towards a common goal; limited student voluntary participation is observed.
2. The classroom environment seems somewhat respectful, but there are some instances of disruptive or disrespectful student behavior. For example, the teacher may attempt to provide a safe environment and/or provide some instruction on how to work together, but students occasionally engage in and/or experience put-downs, taunts, even occasional threats or slurs about themselves or backgrounds. The teacher rectifies the problem on a situation-by-situation basis. The teacher may or may not encourage reluctant students to participate in discussions.
3. The classroom environment appears to reflect mutual and widespread respect between teachers and students. The classroom is characterized by few, if any, taunts and primarily polite, appropriate interactions among students and between students and teacher. For the majority of instruction, both teacher and students solicit and welcome contributions from all students.

Area of interest	**Motivation and Student Engagement**
Description	

1. Disruptive or passive disengagement; most students are frequently off-task, as evidenced by either gross inattention or serious disruptions. For substantial portions of time, many students are either off-task or nominally on-task but not trying very hard. Students could appear to be lethargic and disinterested in class activities or they might be actively misbehaving.
2. Sporadic or episodic engagement; most students, some of the time, are engaged in class activities. Engagement may be uneven, mildly enthusiastic or dependent on frequent prodding from the teacher.
3. Engagement is widespread; most students are on-task most of the time pursuing the substance of the lesson. The majority of students seem to be taking the work seriously and trying hard.

Enhanced Reading Opportunities Program

Reading Apprenticeship Academic Literacy Fidelity Scales

Core Principle # 1	Social Reading Community

A *Social Reading Community* is established so that students can work collaboratively with their teacher and peers to derive meaning and pleasure from text.

- A safe and nurturing classroom environment is established.
- Well-established classroom routines foster peer interaction.
- Through teacher modeling, students are encouraged to recognize and use the diverse perspectives and resources brought by each member of the class.
- Students are encouraged to share their confusion and difficulties with texts, without fear of embarrassment or punishment.
- Teacher actively listens to and responds to students' comments in teacher-facilitated conversations; over the course of the year, students increasingly contribute to and guide whole-class conversations and activities.
- Teacher takes steps to encourage active student participation and to invite diverse responses.
- Teacher shares his or her own struggles, satisfactions and reading processes.

Fidelity Scale

1. The classroom environment does not promote an open exchange of student ideas about text. The teacher may do little or no modeling of such interaction.

   Such an environment could be characterized by little or no student sharing related to the evaluation or generation of meaning from text. Many students may appear to be reluctant to participate in discussions related to text most of the time. The teacher may have to work extremely hard to get students to interact about text meaning, or prompting by the teacher to encourage student conversations about literature is ineffective.

   Instruction in this category could also be characterized by students ridiculing their peers when they acknowledge confusion about text. The teacher may ignore student attempts to express confusion or may not model respect for the varied perspectives and ideas of all members of the classroom community.

2. In general, the classroom environment appears to be a safe place to interact and share ideas about text. The teacher occasionally models appropriate ways for sharing ideas about text.

   A moderately developed social reading community could be characterized by discussions about text that are primarily teacher-directed during the majority of the instructional period. Classroom routines for peer interaction may not be fully developed. Some students may appear to be hesitant to volunteer their own ideas or confusion about text. As another example, the teacher may actively listen to student responses and attempt to elicit a variety of responses from all members of the reading community, but he or she has trouble engaging the majority of students in discussion of literature or of text meaning.

3. A safe and nurturing environment is established for students to share ideas about text. When necessary, the teacher models a process for sharing ideas about text.

   This social reading community could be characterized by frequent student participation. The majority of students contribute to or guide whole-class or group conversations and activities related to literature and other forms of text. They may also volunteer confusion and difficulties with texts. A positive social reading community could also be evident during teacher-facilitated conversations that encourage active participation from all members of the classroom community.

Core Principle # 2	**Metacognitive Conversation**

Metacognitive Conversation is a regularly occurring routine which is evident in RAAL classroom work and interactions:

- Students are taught to use classroom inquiry to generate a repertoire of specific comprehension and problem-solving strategies.
- Through ongoing conversations rooted in text, students learn to ask critical questions about content, purpose, and perspective.
- Students are encouraged to draw on strategic skills they use in out-of-school settings to assist them in solving comprehension problems.
- Students recognize that confusion can be a starting place for collaborative problem-solving aimed at deriving meaning from difficult text.
- Students have many opportunities to practice sharing and exploring their thinking about texts *with peers*; these peer-guided metacognitive conversations become more text-based and sophisticated over the course of the academic year.
- Students monitor their own mental processes for reading and adjust as needed.[*]
- During discussions, teacher probes for deeper student responses to enrich student learning and thinking processes.
- Teacher models metacognitive process (e.g. Thinking Aloud, Talking to the Text) *and* follows through on such practices with continued modeling and appropriate scaffolding to ensure that streams of thought are fully developed.

Fidelity Scale

1. Students are not explicitly taught a variety of comprehension and problem-solving skills. Students are primarily engaged in instruction that is aimed at uniform understandings and single correct responses.

   For example, there is little evidence that reading comprehension difficulties are seen as valuable starting points for collaborative problem-solving. Students have few opportunities to practice discussing their thought processes about reading and to ask critical questions about text content. Students do not volunteer to discuss confusion about text. Students are never or rarely asked to make connections to strategic skills they use in out-of-school settings to assist them in solving comprehension problems.

   As another example, the teacher does not model metacognitive strategies, or does not provide scaffolds for students to practice and apply such strategies. Instruction that falls into this category could be characterized by teacher attempts to model the use of metacognitive strategies that are largely unsuccessful or ineffective.

2. Students are taught comprehension and problem-solving skills, and at least one major classroom activity provides students with an opportunity to discuss their cognitive processes.

   For example, some but not all students may share reading difficulties and confusions and collaborate in problem solving. Instruction could include opportunities for students to share problem solving and strategic skills from their lives outside of school.

   Instruction could also include teacher or student engagement in discussion or assessment of the effects of particular reading processes. While the teacher occasionally models metacognitive strategies or probes for deeper student responses in relation to text, only minimal attempts are made to follow through with additional modeling or appropriate scaffolds to ensure that thought streams are fully developed and transparent.

[*] While we are including this bullet in the general description of the principles, we will not include in the fidelity scales as this is a "high inference" item and is not easily observable.

3. Students are taught a variety of comprehension and problem-solving skills, and they actively contribute to or guide metacognitive conversations. Such conversations are predominantly text-based.

   For example, many students routinely make connections to strategic skills they use in out-of-school settings to assist them in solving comprehension problems. Students may also share their confusion with text as a basis for comprehending challenging text.

   As another example, the teacher frequently and authentically models metacognitive strategies (such as using confusion as a point to generate meaning) or probes for deeper student responses in relation to text. Initial modeling is followed by additional modeling and/or appropriate scaffolds aimed at ensuring that thought streams are fully developed and transparent.

Core Principle # 3	**Silent Sustained Reading**

Silent Sustained Reading is a well-established routine in which personal inquiry and peer social interaction is used to build motivation and extend students' interest to new books and genres.

- Students are encouraged to explore their own preferences and reactions to books.
- Students routinely discuss SSR books with classmates in both informal and occasionally formal activities (i.e. "book talks").
- Students set goals for their reading development and assess their own performance in meeting those goals (in terms of amount and range of books read, persistence, and fluency).
- Students practice metacognitive routines, language study, and cognitive strategies as they read SSR books.
- Teachers routinely provide support and show interest in students' SSR in both informal and formal activities, e.g., individual conferencing, written feedback in reading logs, sharing their own SSR books and reading processes.

Fidelity Scale

0. SSR did not take place during the observed class period(s).

1. Instructional time may be allocated for SSR, but this does not seem to be a developed routine. Instruction could be characterized either by little engagement in SSR or by some engagement in SSR that is not deep or broad. SSR may be a largely individual activity. For example, teachers may not help students select books and may in fact be disengaged from the class doing unrelated activities (e.g. grading papers). As another example, there may be little collaboration on comprehension problems or sharing of reading processes. Students do not have much opportunity to practice metacognitive routines, conduct language study, or do logging, goal-setting, or sharing related to SSR books.
2. The majority of students engage in independent reading during SSR. There is some exploration of SSR reading experiences but the routine is not fully developed. Instruction could be characterized by a few instances of student discussion of reading processes and sharing related to SSR books, personal goal-setting, or writing. As another example, teacher may provide some support of SSR by assisting students in selecting books that reflect their identities as readers, or by engaging in formal or informal feedback activities such as individual conferences to discuss their SSR books and written feedback in student reading logs.
3. Students are engaged in reading SSR books and in reflecting on them either in journals or metacognitive logs or through conversations with peers. In this category, SSR routinely involves the class community in metacognitive conversation, sharing reading strategies and examples for language study. Students set increasingly challenging goals for SSR and monitor their progress. Instruction could also be characterized by demonstrated teacher interest in SSR through both formal and informal activities. For example, the teacher may hold individual conferences with students to discuss their SSR books or provide written feedback in student reading logs.

Core Principle # 4	**Language Study**

Language Study is routinely integrated into varied literacy experiences in the RAAL classroom in both explicit and implicit ways:

- Language study activities engage students in and focus on finding and analyzing patterns at the word, sentence, and text levels.
- Students "nominate" challenging words, phrases, and sentences from their own SSR reading and/or from class readings for analysis by the whole class.
- Students build personal dictionaries of vocabulary words, drawing from key conceptual words taught explicitly as well as from words they encounter in their SSR reading.
- Teachers routinely take advantage of informal opportunities to support academic language development, e.g., by using interesting and playful language, gracefully reframing or elaborating student thinking using academic language. (S: You could tell that was going to happen. T: It really foreshadowed the tragic ending, didn't it?)
- In planning lessons, teachers analyze texts for potential language learning opportunities, and plan language study to take advantage of these.*

Fidelity Scale

0. Not applicable. Language Study did not take place during the observed class period(s).

1. The teacher makes minimal attempts to incorporate language study into instructional activities, but these opportunities are not well developed. For example, the teacher may identify important vocabulary in class and either define or ask students to define the new words; however, little instructional attention is given to the structural features of words, phrases, or texts.
2. The teacher draws students' attention to the structure of language in various course texts at the morphological, word, phrase, sentence, and discourse levels, but instruction in language study is not deep or pervasive. For example, the teacher may incorporate aspects of language study into instruction frequently but it does not appear to be consistent (part of formal instruction and informal opportunities). As another example, there may be evidence that students keep their own word lists in notebooks, but there may be little focus on students' learning to clarify the meaning of unknown words.
3. The teacher provides instruction in the structure of language in various course texts, paying attention to morphological, word, phrase, sentence, and discourse. The teacher takes advantage of informal opportunities to support academic language development. For example, the teacher uses interesting and playful language or attempts to reframe or elaborate student thinking using academic language. As another example, students keep word lists and routinely identify key words and work to clarify word meaning as they read and work with peers. Instruction could also be characterized by student identification of language for study or student engagement in class or small group analysis of challenging words, sentences, or text passages.

* While we are including this bullet in the general description of the principles, we will not include in the fidelity scales as this is a "high inference" item and is not easily observable.

Core Principle # 5	**Content and Theme**

The *Content and Theme* of each of the four thematic units* in the RAAL curriculum are integral to classroom activities and discussions:

- Students practice a variety of comprehension strategies in the context of the texts and genres presented in each of the four thematic units.
- Students are encouraged to draw on their interests in larger social, political, economic, and cultural issues as they read and discuss the texts in each thematic unit.
- Students explore personal motivations and identities as readers in relation to the four thematic units.
- Students practice analyzing and synthesizing information and ideas across multiple texts and conversations in relation to the overarching themes of the four units.
- The teacher provides instruction and support for reading the complex academic materials associated with each of the four units occurs in the classroom; reading is not merely assigned and reviewed.
- Students learn and practice academic discourse (e.g., providing evidence to support thinking, interrogating author bias) appropriate for each of the four thematic units.

Fidelity Scale

1. For the majority of the instruction period, the focus of instruction does not center on the content or theme of the current unit. If the content or theme is addressed, the class engages in only tangential discussion of the materials at hand. The teacher makes no attempt to redirect or reorient students to material relevant to current thematic unit.
2. Much of the instruction is focused on the theme of the current unit but some opportunities for integrating the overarching theme with instruction are lost. For example, students may practice a comprehension strategy in the context of the texts and genres presented in this unit, but they do not draw on their own interest in larger social or cultural issues related to the theme. As another example, students may explore personal motivations or identities related to the theme but the teacher may not provide support for reading the academic materials associated with the unit. In this category, some instruction may occur with no reference to the theme.
3. The majority of instruction focuses on text and materials relevant to the theme, and the teacher provides ample support for reading complex academic materials within the current thematic unit. For example, students have multiple or extended opportunities to practice comprehension strategies specific to the context of the texts and genres presented in this unit. As another example, students explore their personal motivations and identities in relationship to the unit and draw on their interests in larger social, political, economic, and cultural issues. Students may analyze or synthesize information across multiple texts, or they may practice academic discourse appropriate for the unit.

* The four thematic units of the RAAL curriculum consist of Unit 1: Reading Self and Society; Unit 2: Reading History; Unit 3: Reading Science; and Unit 4: Reading Media.

Core Principle # 6	**Writing**

Instruction provides on-going support for *writing to learn as well as learning to write* in the RAAL classroom:

- Students are explicitly taught writing processes and the structures of particular written forms through formal writing assignments that culminate each of the four thematic units.
- Instruction and support for writing and writing processes occur in the classroom; writing is not merely assigned and graded.
- Students use writing to support their learning of thematic content through a variety of tools, including dual entry journals, graphic organizers, interactive notebooks, personal dictionaries, word and sentence analysis notes, and reflective letters.
- Students use writing as a tool for increasing their comprehension of challenging texts (e.g., students write in metacognitive logs and practice the metacognitive routine of "talking to the text" in writing).

Fidelity Scale

0. Not applicable. The observed class period(s) did not include a writing component.

1. Students are not explicitly taught writing processes or about the structures of particular written forms. For example, writing assignments may be given to students, but they never receive guidance on the writing process. Instruction could alternatively be characterized by a lack of opportunities for students to use writing to support their learning of thematic content or to increase comprehension of text. Metacognitive logs may be used, but appear to be used in a very rote way (students write a simple sentence or two and these are not explored further).
2. Students engage in at least one activity where they are developing writing skills and using writing to support their learning of thematic content, but one aspect is developed in greater depth than the other. For example, instruction on learning to write may be emphasized (the writing process and the structures of particular written forms) without a lot of attention to the content of the writing. As another example, thematic content may be explored through writing tools such as dual entry journals, metacognitive logs, graphic organizers, interactive notebooks, personal dictionaries, word and sentence analysis notes, and reflective letters; <u>but</u> the writing process is not fully explored or developed.
3. Explicit instruction is provided in the writing processes and the structures of particular written forms related to the thematic unit; the two skill/strategies are developed hand in hand. Students use writing as a tool for increasing their comprehension of challenging texts. For example, students write in metacognitive logs and practice the metacognitive routine of "talking to the text" and hone their writing skills in the process. Students may also learn to write and use writing to support their learning of thematic content through other tools, including dual entry journals, graphic organizers, interactive notebooks, personal dictionaries, word and sentence analysis notes, and reflective letters.

Core Principle # 7	**Integration of the Curriculum Strands**

The teacher integrates the five RAAL Curriculum Strands* during literacy instruction

- Students are simultaneously engaged in at least two of the strands at any given time.
  - For example, while focusing on *Metacognitive Conversation* in discussing how students solved comprehension problems reading a piece in the anthology, the teacher might integrate *Language Study* by providing a mini-lesson on roots, prefixes and suffixes in helping students clarify the meaning of an unfamiliar word.
  - For another example, the teacher might integrate Writing and Content and Theme through student discussion and writing about the "essential questions" in any of the four thematic units.

Fidelity Scale

1. The teacher does not integrate curriculum strands in any of the major instructional activities.
   OR

   The teacher occasionally integrates two of the curriculum strands, but does not do so in a natural manner. For example, coherent connections between course themes, language study, metacognitive conversation and strategies, independent reading experiences, and/or writing are not evident throughout the majority of instruction.
2. For at least one major activity, the teacher integrates at least two strands smoothly; instruction in each of the strands is improved upon by instruction in the other. *For example,* while focusing on *Metacognitive Conversation* in discussing how students solved comprehension problems, the teacher might integrate *Language Study* by providing a mini-lesson on roots, prefixes and suffixes in helping students clarify the meaning of an unfamiliar word. During the remainder of instruction, the teacher may refer to one or more of the curriculum strands but only in passing, or without coherently integrating them with other strands.

   As another example, the teacher successfully focuses on two of the strands for the majority of the instruction but does not make attempts to integrate any remaining strands.
3. The teacher finds multiple opportunities to integrate several of the five strands "fluently" and appropriately. At least two different strands appear to be seamlessly integrated at any given time. For example, the teacher recognizes and makes use of opportunities to make natural and meaningful connections between and among course themes, language study, metacognitive conversation and strategies, independent reading experiences, and writing.

* The five strands of the RAAL Curriculum consist of Metacognitive Conversation, Silent Sustained Reading, Language Study, Content/Theme, and Writing

Enhanced Reading Opportunities Program

Xtreme Reading Fidelity Scales

Core Principle # 1	**Responsive Instruction**

Instruction is *responsive* to unique student needs to "personalize teaching and learning."

- *Assessment:* Ongoing, informal assessment is used to monitor students' performance to determine if instructional objectives are being met and strategies are being mastered.*
- *Accommodations (1.a):* Students begin learning reading strategies using materials at their reading level. They gradually work up through the reading levels across the school year.
- *Feedback (1.b):* Corrective and elaborative feedback is provided to help students better understand how to improve their performance of skills and strategies. Feedback helps students recognize correct practices, as well as patterns of errors, and target improvement in specific areas. Six steps for providing feedback are recommended:
  - Teacher tells students what they have done well.
  - Teacher helps students recognize and categorize errors made during practice attempts, in order to better understand their performance.
  - Teacher re-teaches one of the error types at a time (through explaining, modeling).
  - Teacher watches student practice and provides feedback.
  - Teacher asks student to paraphrase main elements of feedback.
  - Teacher prompts student to set goals for next practice attempt.

Fidelity Scale: (Core Principle 1.a: Accommodations)

0. There was no opportunity to make accommodations during the observed class period(s).
1. The teacher seems unaware of or unable to determine whether instructional objectives are being met and strategies are being mastered. For example, students are provided few instructional materials that match their reading level. Materials appear to be either too challenging or too easy for the majority of the students.
2. The teacher appears to be able to provide appropriate instruction to students making expected progress but appears unaware of or unable to determine appropriate instruction for students failing to make adequate progress or for students advancing rapidly through the curriculum. For example, while some students are being instructed in materials that match their reading level, the materials appear to be either too difficult or too easy for others.
3. The teacher appears to be aware of individual student needs and is able to differentiate instruction accordingly. For example, most students have been provided with instruction and are learning reading strategies using materials at their reading level.

* While we are including this bullet in the general description of the principles, we will not include in the fidelity scales as this is a "high inference" item and is not easily observable. Assessment is addressed in the teacher interview, and teachers will be asked to describe their use of assessments to make instructional decisions.

Fidelity Scale (Core Principle 1.b: Feedback)

0. There is no opportunity to provide feedback to students during the observed period (s).
1. There is opportunity but the teacher does not provide feedback to students or does so rarely. The teacher does not appear to monitor student work and performance. In general, students are expected to practice skills and strategies independently, without teacher input.
2. While the teacher occasionally provides corrective feedback to students on their practice attempts, feedback is not elaborative or mainly highlights the negative. In general, the teacher engages in only one or two of the feedback strategies outlined in the Xtreme Reading Program (telling students what they have done well, helping students to recognize and categorize errors made during practice attempts, reteaching one of the error types at a time through modeling and explaining, watching students practice, asking students to paraphrase main elements of feedback, and prompting students to set goals for their next practice attempt). There is little follow-up with students to ensure understanding so that they may improve on their next practice attempt and obtain mastery of the skill/strategy.
3. Corrective and elaborative feedback is provided to help students better understand how to improve their performance of skills and strategies. The teacher provides feedback using most or all of the strategies outlined in the Xtreme Reading Program (telling students what they have done well, helping students to recognize and categorize errors made during practice attempts, reteaching one of the error types at a time through modeling and explaining, watching students practice, asking students to paraphrase main elements of feedback, and prompting students to set goals for their next practice attempt). The teacher follows up with students to ensure understanding so that they may improve on their next practice attempt and move toward mastery of the skill/strategy.

Core Principle # 2	**Systematic Instruction**

Instruction is *systematic* in nature; that is, the information (skills, strategies, and content) taught, the sequence of instruction, and various activities and materials used are carefully planned in advance of delivering instruction. Systematic instruction is to be carefully structured, connected, and scaffolded; and it should be informative.

- *Structured Content (2.a):* Instructional content is comprised of instruction in reading strategies (e.g., vocabulary, word-identification, self-questioning, visual imagery, paraphrasing, and inferencing) and other instructional programs that support strategy instruction (ACHIEVE Skills, SCORE Skills, Talking Together, Possible Selves). Each reading strategy is divided into smaller steps/segments.
- *Research-based instructional methodology (2.b):* Each strategy is taught using an eight-stage methodology. On each day that a reading strategy is taught, the learning activities are associated with at least one of these stages. The stages include: Describe, Model, Verbal Practice, Guided Practice, Paired Practice, Independent Practice, Differentiated Practice, and Generalization.
- *Connected Instruction (2.c):* Teacher purposefully shows students how new information is related to skills, strategies, or content that has been previously learned, as well as to those that will be learned in the future. Course and Unit Organizers are provided to students to introduce main ideas and to demonstrate how critical information and concepts are related.
- *Scaffolded Instruction (2.c):* Instruction moves from teacher-mediated to student-mediated across the course of instruction in one strategy. When a new strategy is introduced, multiple instructional supports (modeling, prompts, direct explanations, targeted questions, relatively basic tasks) are initially provided by the teacher. These instructional supports are gradually reduced as the student becomes more confident and begins to move toward mastering the targeted objectives.
- *Informative Instruction (2.c):* Teacher informs students about how the learning process works and what is expected during instruction. Teacher ensures that students understand how they are progressing, how they can control their own learning at each step of the process, and why this is important.

Fidelity Scale (Core Principle 2.a: Structured Content)

1. There is little or no evidence that that the teacher is providing instruction in any of the reading strategies outlined in the Xtreme Reading curriculum (e.g., vocabulary, word-identification, self-questioning, visual imagery, paraphrasing, and inferencing) and other instructional programs that support strategy instruction (ACHIEVE Skills, SCORE Skills, Talking Together, Possible Selves). For example, the teacher appears to be using alternative instructional materials (materials outside of the Xtreme Reading curriculum).
2. While the teacher is providing instruction in one of the reading strategies or instructional programs that support strategy instruction, the teacher does not demonstrate a thorough understanding of the content. For example, students may not be provided with an in-depth, comprehensive understanding of the strategy and/or program and the teacher, while able to answer basic questions, might not be able to thoroughly respond to more complex questions on the instructional content. As another example, the teacher may be providing comprehensive instruction in the strategy but may not be providing instruction in small steps or segments appropriate for developing student understanding.
3. Instructional content is comprised of instruction in reading strategies (e.g., vocabulary, word-identification, self-questioning, visual imagery, paraphrasing, and inferencing) and other instructional programs that support strategy instruction (ACHIEVE Skills, SCORE Skills, Talking Together, Possible Selves). The teacher demonstrates a strong understanding and knowledge of the content and is able to thoroughly respond to student questions. Further, instruction in the strategy is divided into small steps or segments to facilitate the development of student understanding in this strategy

Fidelity Scale (Core Principle 2.b: Research-based Methodology)

1. The teacher does not use any of the eight instructional stages of the Xtreme Reading Program;* and the learning activities do not appear to be associated with the program's curriculum. Instruction appears unsystematic and unmethodical.
2. The teacher uses one of the eight instructional stages of the Xtreme Reading Program;* however, the teacher does not demonstrate a thorough understanding of the learning activities associated with the specific instructional stage. Although students are involved in learning activities associated with the specific instructional stage, at times, instruction appears unsystematic.
3. The reading strategy of focus is taught using one of the eight stages of the Xtreme Reading instructional methodology. The teacher engages students in learning activities associated with at least one of the eight instructional stages of the Xtreme Reading Program.* The teacher's implementation of the instructional stage reflects best practices, as outlined by the Xtreme Reading instructional methodology, and instruction is delivered in a systematic manner.

* The eight instructional stages are: Describe, Model, Verbal Practice, Guided Practice, Paired Practice, Independent Practice, Differentiated Practice, Generalization

Fidelity Scale (Core Principle 2.c: Connected, Scaffolded, and Informed Instruction)

1. Instruction is neither connected, scaffolded, nor informative. In almost all instances, the teacher does not show students how new information is related to skills, strategies, or content that they have previously learned or that will be learned in the future. Course and Unit Organizers are rarely used for this purpose. There is little evidence of the teacher providing multiple instructional supports (i.e. modeling, prompts, direct explanations, targeted questions, etc.) to facilitate movement from teacher-mediated to student-mediated instruction. The teacher rarely engages students in discussion regarding their own learning process, learning expectations, and why it is important for students to take control of their own learning.
2. Instruction may be connected, scaffolded, or informative, but it does not reflect all three characteristics. In some cases, the teacher provides a brief explanation of how new information is related to skills, strategies, or content that has been previously learned, as well as to those that will be learned in the future. The teacher uses Course and Unit Organizers to introduce new information but does not engage students to ensure their understanding. The teacher provides students with some instructional supports, but not in a systematic manner to promote movement from teacher-mediated to student-mediated instruction. Occasionally, the teacher engages students to ensure they understand how they are progressing, to inform students of how they can control their own learning and why this is important.
3. Instruction is connected, scaffolded, and informative. The teacher purposefully shows students how new information is related to skills, strategies, or content that has been previously learned, as well as to those that will be learned in the future. Course and Unit Organizers are provided to students to introduce main ideas and to demonstrate how critical information and concepts are related. The teacher provides students with multiple instructional supports (i.e. modeling, prompts, direct explanations, targeted questions, etc.) that promote movement from teacher-mediated to student-mediated instruction. The teacher informs students about how the learning process works and what is expected during instruction. The teacher ensures students understand how they are progressing, how they can control their own learning and why this is important.

Core Principle # 3	**Classroom Management**

Classroom management and planning techniques maximize the use of instructional time.

- Expectations for all activities and transitions between activities are explained, taught, and reinforced throughout instruction.
- Classroom routines are established early, and students demonstrate familiarity and comfort with these routines.
- Lessons are clearly structured, and all instructional time is used for instruction.
- Interactive learning experiences ensure that students practice, master, integrate, and generalize critical skills.

Fidelity Scale

1. There is little or no evidence of established classroom management techniques. Students do not seem familiar or comfortable with classroom routines. Instructional time is lost due to disorganized transitions between activities and to disciplinary matters. This could take the shape of disorganized, poorly structured instructional activities. As another example, the teacher may not articulate explicit expectations for activities and transitions.
2. Although classroom management techniques appear to be in place, they do not always serve to maximize instruction. At times, students demonstrate a familiarity and comfort with classroom routines. For example, teacher expectations may be articulated for some activities, but are not always reinforced throughout instruction. Some lessons are clearly structured and most instructional time is used for instruction. As another example, interactive learning experiences allow students to practice, master, integrate, and generalize critical skills, but at times students need to be redirected to stay on-task and on-topic.
3. Classroom management techniques maximize the use of instructional time. Students demonstrate a familiarity and comfort with classroom routines and remain focused throughout the instructional period. Instruction fitting this category could take the form of clear and explicit teacher expectations for all activities and transitions between activities that are reinforced throughout the instruction. As another example, lessons are clearly structured and all instructional time is used for instruction. Interactive learning experiences ensure that students practice, master, integrate, and generalize critical skills.

Core Principle # 4	**High Student Motivation and Engagement**

Instruction reflects high student motivation and engagement.

- *Student Engagement:* Engagement is maintained in the classroom through activities that enable students to focus attention on critical learning outcomes. Instruction demands a high degree of student attention and response, and expectations are set high for student work. Instruction is interactive and appropriately paced to maintain student attention.
- *Student Motivation:* Motivation is achieved by providing students with a real purpose for improving their literacy skills and by linking learning to their personal goals. In addition, interesting novels are used to motivate students to engage in reading activities.

Fidelity Scale

1. There is little or no evidence of student engagement in classroom activities, and there are few if any opportunities for active learning. For example, the pacing of instruction does not maintain student engagement; students demonstrate boredom and/or frustration regarding the content being taught. As another example, teacher expectations for quality student work and performance appear to be low.

   The teacher does not provide students with a real purpose for improving their literacy skills and engaging in the lesson activities. For example, there is little evidence to suggest students are provided with interesting novels to read while engaging in reading activities.

2. During some activities, student engagement is maintained through activities that require a high degree of student attention and response; however, not all students are engaged at all times. For example, the pacing of instruction appears appropriate for some students, but others demonstrate boredom and/or frustration with the content being taught.

   At times, the teacher provides students with a purpose for improving their literacy skills, but this purpose is not always clearly relevant, or clearly linked to students' personal goals. It appears that students have access to novels in the classroom, but it is unclear the extent to which these reading materials are used to engage students in reading activities.

3. Student engagement is maintained in the classroom through activities that enable students to focus attention on critical learning outcomes. Instruction demands a high degree of student attention and response, and expectations are set for high-quality student work. Instruction is interactive and appropriately paced to maintain student attention.

   The teacher facilitates student motivation by providing students with a real purpose for improving their literacy skills and by linking learning to their personal goals. Additionally, interesting novels are used to motivate students to engage in reading activities.

Appendix E

Technical Notes for Impact Findings

This appendix provides three sets of additional technical notes that accompany the impact findings presented in Chapter 5. The first section describes the statistical model used to estimate the impact of the two supplemental literacy programs used in the Enhanced Reading Opportunities (ERO) study — Reading Apprenticeship Academic Literacy (RAAL) and Xtreme Reading — on reading achievement and reading behaviors. The second section presents tables that show the sensitivity of the core impact findings to including student-level baseline characteristics in the statistical model, for the full sample of 34 schools and for the groups of schools using each of the two literacy programs. These tables also present the standard errors ("S.E." in the tables) and 95 percent confidence intervals for the adjusted and unadjusted impacts. The third section addresses the issues related to multiple hypothesis tests of impacts on multiple reading behavior measures. Specifically, it presents the findings from the qualifying tests that were performed to assess the robustness of the statistical significance of the impacts on the three reading behavior measures examined in Chapter 5.

Statistical Model for Estimating Impacts

The ERO study impact analysis uses the following statistical model to estimate impacts on both reading achievement and reading behaviors:

$$Y_i = \sum_n \gamma_{0n} S_{ni} + \gamma_1 Y_{-1i} + \sum_S \gamma_{2s} X_{si} + \beta_0 T_i + \varepsilon_i \qquad (1)$$

Where:

Y_i = reading achievement or reading behaviors outcome for student i

$\sum_n S_{ni}$ = school dummy variable, one if student i is in school n and zero otherwise

Y_{-1i} = the GRADE reading comprehension test score for student i before random assignment

$\sum_s X_{si}$ = other pre-random assignment characteristics for student i

T_i = one if student i is assigned to the ERO group and zero otherwise

ε_i = student-level random error term

In this model, β_0 represents the estimated impact of the ERO programs on the outcome of interest (Y_i). β_0 is a fixed-effect impact estimate that addresses the question: What is the impact of the ERO programs for the average student in the follow-up respondent sample? This approach is taken because this study most closely reflects an efficacy study of the effects of a

new supplemental literacy intervention under relatively controlled conditions. Also, the sites and students were not selected to be a random sample of a larger population of sites. Instead, sites were selected purposively through the OVAE special SLC grant competition discussed in Chapter 2, using specific criteria that differentiated these schools and districts from others that were not awarded a grant. Although, on average, the participating schools share characteristics of other low-performing urban high schools across the country, the impact estimates are not statistically generalizable to a larger population of districts, high schools, or students.

Equation 1 includes indicator variables for each of the participating high schools. These covariates capture a central feature of the study design in which random assignment was conducted within each of the participating high schools. These covariates are included to account for variation in the mean value of the dependent variable across the participating high schools.

Equation 1 also includes a covariate for each student's GRADE reading comprehension test score at baseline and a covariate indicating whether the student is overage for grade (and likely to have been retained in a prior grade). These covariates are included to improve the precision of the impact estimates.

Adjusted and Unadjusted Impact Estimates

As explained above, the impacts presented in Chapter 5 of this report are estimated using regression adjustments for random differences between the ERO and non-ERO groups in their pretest scores and whether a student was overage for the ninth grade. The first two tables in this appendix provide both regression-adjusted (in the "Estimated Impact" column) and unadjusted impacts (in the "Difference" column). These tables also include other information that may be useful to those who may wish to include these impacts in meta-analyses. Note that random assignment of students to the ERO and non-ERO groups occurred within each high school (that is, random assignment was "blocked" by school). Because of differences across schools (blocks) in the number of students eligible and appropriate for the ERO programs, the ratio of ERO group members to non-ERO group members in each site varies from 1.14 to 2.0. Thus, all the impact estimates presented in this report include controls for each block to account for random differences between the ERO and non-ERO groups that may be associated with differences in the random assignment ratios. The assessment of sensitivity to other regression adjustments presented in the appendix reflects potential differences in impact estimates that also controls for the blocking of random assignment by school.

Appendix Table E.1 is the counterpart to Tables 5.1 and 5.2 and shows adjusted and unadjusted impacts on reading achievement for all 34 schools in the study and for the groups of

The Enhanced Reading Opportunities Study

Appendix Table E.1

Impacts on Reading Achievement, Cohort 2 Follow-Up Respondent Sample

	Mean Differences Adjusting for Blocking Only					Regression-Based Impact Estimates				
Outcome	ERO Group (S.D.)	Non-ERO Group (S.D.)	Difference (S.E.)	95% Confidence Interval	P-Value for the Difference	Estimated Impact (S.E.)	95% Confidence Interval	P-Value of Estimated Impact	Impact Effect Size (S.E.)	95% Confidence Interval
All schools										
Reading comprehension										
Average standard score	90.10	89.49	0.61	-0.25 1.46	0.166	0.83	0.03 1.63	0.042	0.08	0.00 0.16
	(10.15)	(10.03)	(0.44)			(0.41)			(0.04)	
Reading vocabulary										
Average standard score	93.45	93.68	-0.22	-1.07 0.63	0.607	-0.01	-0.80 0.79	0.986	0.00	-0.08 0.08
	(10.34)	(9.83)	(0.43)			(0.41)			(0.04)	
Sample size	1,264	907								
Reading Apprenticeship schools										
Reading comprehension										
Average standard score	90.24	89.05	1.19	0.00 2.38	0.049	1.38	0.27 2.50	0.015	0.14	0.03 0.25
	(10.02)	(10.11)	(0.61)			(0.57)			(0.06)	
Reading vocabulary										
Average standard score	93.39	94.02	-0.62	-1.77 0.52	0.285	-0.43	-1.51 0.64	0.428	-0.04	-0.15 0.07
	(10.04)	(9.31)	(0.58)			(0.55)			(0.06)	
Sample size	645	470								
Xtreme Reading schools										
Reading comprehension										
Average standard score	89.96	89.97	-0.02	-1.26 1.22	0.980	0.25	-0.91 1.41	0.672	0.02	-0.09 0.14
	(10.30)	(9.95)	(0.63)			(0.59)			(0.06)	
Reading vocabulary										
Average standard score	93.52	93.32	0.20	-1.06 1.46	0.752	0.44	-0.75 1.62	0.468	0.04	-0.08 0.17
	(10.66)	(10.35)	(0.64)			(0.60)			(0.06)	
Sample size	619	437								

(continued)

Appendix Table E.1 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study follow-up GRADE assessment.

NOTES: The follow-up GRADE assessment was administered in the spring of 2007 near the end of students' ninth-grade year.

The estimated impacts in the column labeled "Difference" are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed distribution of the ERO group across random assignment blocks (i.e., schools) as the basis for the adjustment. The estimated impacts in the column labeled "Estimated Impact" are regression-adjusted for blocking of random assignment by school, as well as for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment.

The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form B). No statistical tests or arithmetic operations were performed on these reference points.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (reading comprehension = 10.035; reading vocabulary = 9.827).

A two-tailed t-test was applied to the impact estimate. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

schools using each of the two ERO programs. Appendix Table E.2 is the counterpart to Tables 5.3 and 5.4 and shows adjusted and unadjusted impacts on reading behavior measures.[1]

Addressing Risks Associated with Multiple Hypothesis Tests

In Chapter 5, statistical significance is indicated in the tables by an asterisk (*) when the p-value of the impact estimate is less than or equal to 0.05 (5 percent). As discussed in Chapter 2, however, when making judgments about statistical significance, it is important to recognize potential problems associated with conducting multiple hypothesis tests. Specifically, it is important to minimize the risk that conclusions from the study could be based on false positive results (also known as Type I errors) while simultaneously limiting the risk that important results may be neglected due to false negative results (also known as Type II errors). In other words, the analysis should avoid concluding that an impact estimate is statistically significant when, in fact, there is no true impact. Likewise the analysis should not be so conservative with respect to producing false positives that it unduly increases the likelihood of missing true impacts when they exist (that is, of producing false negatives).

As the number of hypothesis tests increases, the probability of finding a statistically significant impact estimate when there is no true impact may also increase. One could dramatically reduce this risk by making the standard for statistical significance much more stringent, for example, by setting the p-value to less than or equal to 0.001. Making the standard too stringent, however, will increase the likelihood that one would judge an impact estimate to be not statistically significant when, in fact, it represents a true impact. The approach adopted for this project provides a framework that aspires for an *acceptable balance* between the risks of making Type I and Type II errors.

The impact analysis conducted for this report includes two sets of safeguards aimed at attenuating the risk of drawing inappropriate conclusions about program effectiveness on the basis of multiple hypothesis tests. The first safeguard is to identify a *parsimonious* list of outcome measures and subgroups and then to prioritize among these to specify the primary and secondary hypothesis tests that would be used to make judgments about the overall effectiveness of the ERO programs. The shorter this list, the fewer the number of hypothesis tests and, thus, the less exposed the analysis will be to "spurious statistical significance" as a result of having tested multiple hypotheses.

[1]Results from the regression-adjusted impact analyses are presented in the columns under "Regression-Based Impact Estimates," and results from the unadjusted impact analyses are presented in the columns under "Mean Differences Adjusted for Blocking Only."

The Enhanced Reading Opportunities Study

Appendix Table E.2

Impacts on Reading Behaviors, Cohort 2 Follow-Up Respondent Sample

	Mean Differences Adjusting for Blocking Only						Regression-Based Impact Estimates					
Outcome	ERO Group (S.D.)	Non-ERO Group (S.D.)	Difference (S.E.)	95% Confidence Interval		P-Value for the Difference	Estimated Impact (S.E.)	95% Confidence Interval		P-Value of Estimated Impact	Impact Effect Size (S.E.)	95% Confidence Interval	
All schools													
Amount of school-related reading (prior month occurrences)	46.29 (36.93)	46.03 (38.32)	0.25 (1.60)	-2.88	3.38	0.875	0.40 (1.59)	-2.72	3.53	0.800	0.01 (0.04)	-0.07	0.09
Amount of non-school-related reading (prior month occurrences)	29.88 (32.36)	29.38 (32.98)	0.50 (1.41)	-2.26	3.26	0.722	0.49 (1.41)	-2.28	3.25	0.729	0.01 (0.04)	-0.07	0.10
Use of reflective reading strategies (4-point scale)	2.76 (0.59)	2.71 (0.59)	0.05 (0.03)	0.00	0.10	0.047	0.05 (0.03)	0.00	0.10	0.033	0.09 (0.04)	0.01	0.17
Sample size	1,260	901											
Reading Apprenticeship schools													
Amount of school-related reading (prior month occurrences)	50.46 (38.58)	48.85 (39.97)	1.61 (2.34)	-2.98	6.20	0.492	1.64 (2.34)	-2.95	6.23	0.483	0.04 (0.06)	-0.08	0.16
Amount of non-school-related reading (prior month occurrences)	31.30 (31.73)	31.52 (35.10)	-0.22 (2.00)	-4.15	3.71	0.913	-0.34 (2.00)	-4.27	3.59	0.867	-0.01 (0.06)	-0.13	0.11
Use of reflective reading strategies (4-point scale)	2.78 (0.59)	2.74 (0.57)	0.04 (0.03)	-0.03	0.10	0.297	0.04 (0.03)	-0.03	0.11	0.257	0.07 (0.06)	-0.05	0.18
Sample size	642	466											
Xtreme Reading schools													
Amount of school-related reading (prior month occurrences)	41.97 (34.65)	43.16 (36.29)	-1.20 (2.16)	-5.43	3.04	0.579	-0.84 (2.15)	-5.05	3.38	0.697	-0.02 (0.06)	-0.13	0.09
Amount of non-school-related reading (prior month occurrences)	28.41 (32.97)	27.14 (30.41)	1.27 (1.97)	-2.61	5.14	0.521	1.46 (1.97)	-2.41	5.34	0.458	0.04 (0.06)	-0.07	0.16
Use of reflective reading strategies (4-point scale)	2.74 (0.59)	2.68 (0.61)	0.06 (0.04)	-0.01	0.14	0.079	0.07 (0.04)	0.00	0.14	0.059	0.12 (0.06)	0.00	0.24
Sample size	618	435											

(continued)

Appendix Table E.2 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities follow-up student survey.

NOTES: The student follow-up survey was administered in spring 2007 at the end of students' ninth-grade year.

The estimated impacts in the column labeled "Difference" are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed distribution of the ERO group across random assignment blocks (i.e., schools) as the basis for the adjustment. The estimated impacts in the column labeled "Estimated Impact" are regression-adjusted for blocking of random assignment by school, as well as for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (school-related reading standard deviation = 38.322; non-school-related reading standard deviation = 32.976; use of reading strategies standard deviation = 0.592).

A two-tailed t-test was applied to the impact estimate. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

For each of the above measures, data are missing for no more than 3 percent of the respondents.

Rounding may cause slight discrepancies in calculating sums and differences.

The second safeguard uses composite statistical tests to "qualify" or call into question multiple hypothesis tests that are statistically significant individually but that may be due to chance. These composite tests are referred to as "qualifying tests."

Specifying Primary and Secondary Hypothesis Tests

The primary evidence of overall ERO program effectiveness for this report will be reflected by estimates of program impacts on reading comprehension test scores (expressed in standard score values) for the full study sample and for each of the two ERO programs being evaluated. Anchoring the study's early conclusions in a limited set of outcomes minimizes the risk of relying on a large number of impact estimates, some of which may be statistically significant only by chance. As noted above, student reading comprehension skills constitute the primary target of the ERO interventions and the primary outcome of interest for the first year of the study. Also, the study was designed to provide minimum detectable effect sizes for each ERO subgroup that may be considered policy relevant. Thus, the primary confirmatory hypotheses for the report focus on the overall and program-specific impacts on reading comprehension test scores.

Vocabulary knowledge and student reading behaviors, while targets of the interventions and important to students' literacy development, are considered secondary indicators of program effectiveness. Similarly, subgroups of students (for example, those with higher or lower baseline test scores) provide useful information about the relative impact of supplemental literacy programs, but they too are considered secondary indicators of effectiveness in this report.

Composite Qualifying Statistical Tests

A second set of safeguards against risks associated with multiple hypothesis tests involves the use of composite qualifying statistical tests that provide further context for interpreting the robustness of individual impact estimates and their statistical significance.[2] These statistical tests are applied in cases where impacts are estimated for more than one outcome in a given measurement domain (for example, the three survey measures that attempt to capture students' reading behaviors) or for subgroups of the full study sample. In general, these qualifying statistical tests estimate impacts on composite indices that encompass all the measures in a given domain or estimate the overall variation in impacts across subgroups. If the results of these tests are not statistically significant, this indicates that the statistical significance of the associated

[2]Measurement of overall effects has its roots in the literature on meta-analysis (see O'Brien, 1984; Logan and Tamhane, 2003; and Hedges and Olkin, 1985). For a discussion of qualifying statistical tests to account for the risk of Type I error, see Duflo, Glennerster, and Kremer (2007). Other applications of these approaches are discussed in Kling and Liebman (2004) and Kling, Liebman, and Katz (2007).

individual impact estimates *may have* occurred by chance. In these cases, the discussion of the impacts should include cautions or qualifiers about the robustness of the individual findings.[3]

To test the robustness of the statistical significance of impact estimates for multiple outcomes within a measurement domain (in this case, the three reading behavior measures), the study uses a single composite index consisting of the average of the standardized values for each outcome.[4] Then the estimated impact on this composite measure is calculated for the full study sample. If this qualifying test shows that the composite impact estimate is not statistically significant (its p-value is greater than 0.05), then one concludes that statistically significant impacts for the component outcomes could be due to chance and should be interpreted cautiously.

Specifically, the analysis took the following steps in creating a composite index and assessing impacts on reading behaviors.[5] First, z-scores were created for each reading behavior outcome by subtracting the non-ERO group mean and dividing by the non-ERO group standard deviation. Thus, each component of the index has a mean of zero and a standard deviation of one for the non-ERO group. The z-scores from each component were averaged to obtain the index which was then included in the standard impact estimation model. If the estimated impact for the composite index is not statistically significant, then the statistical significance of impact estimates for the component measures may have occurred by chance and the finding should be interpreted cautiously. In other words, the report qualifies or calls into question a statistically significant individual impact estimate by suggesting that it may have occurred by chance.

To test the robustness of the statistical significance of impact estimates for subgroups of students, a composite F-test is used to assess whether the variation in impacts across all student

[3]Alternative strategies that involve (1) adjusting significance levels (through Bonferroni methods) or (2) adjusting significance thresholds (through Benjamini and Hochberg methods) are overly conservative with respect to making Type I errors and can thereby greatly increase the likelihood of making Type II errors. There are two reasons for this. First, these methods treat all hypotheses as though they were independent of each other. Hence, each hypothesis is treated as representing an independent opportunity to make a Type I error. However, many impact estimates in an evaluation study are correlated with each other and thus do not represent independent opportunities to make Type I errors. In the extreme, for example, if all measures were perfectly correlated, there is only one opportunity to make a Type I error even though there are many outcome measures and, thus, many statistical hypothesis tests. The above methods assume, however, that the number of opportunities to make a Type I error equals the number of hypothesis tests conducted. To the degree that hypothesis tests are correlated with each other, these methods overcompensate (often by a lot) for the risks of Type I error in multiple hypothesis tests. A second source of conservatism with respect to Type I error is the fact that the above methods assume that all null hypotheses may be true. As a result, they consider the potential number of false positives to equal the total number of hypothesis tests conducted. However, the actual number of potential false positives equals the total number of true null hypotheses, not the total number of hypotheses tested. This is because only true null hypotheses can produce false positives. Hence, the methods overcompensate for the number of hypotheses tested.

[4]See Duflo, Glennerster, and Kremer (2007).

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

subgroups is statistically significant. For example, the analysis examines impacts for three sets of student subgroups: those defined by baseline reading test scores (comprising three subgroups); those defined by whether a student was overage for the start of ninth grade (comprising two subgroups); and those defined by whether a student's family spoke a language other than English at home (comprising two subgroups). The composite qualifying test for these analyses assesses whether variation in estimated impacts across these seven subgroups accounts for a statistically significant level of unexplained variance in the test score or other outcome being examined. In other words, the test assesses whether the change in the F-statistic from the core impact regression to the impact regression with the subgroup interaction terms is statistically significant (its p-value is less than or equal to 0.05). If the change in unexplained variance due to the subgroup impact interactions is not statistically significant, then the statistical significance of impact estimates for the component subgroups may have occurred by chance and the findings should be interpreted cautiously.

Finally, the analysis includes qualifying statistical tests to assess the statistical significance of the difference in impacts between the subgroups of students or schools. If these qualifying tests show that the difference in impacts across subgroups is not statistically significant (p-value is greater than 0.05), then one concludes that statistically significant impacts for individual subgroups could be due to chance and should be interpreted cautiously.[6] For example, suppose the findings indicate that impacts on reading comprehension for one group of participating high schools are positive and statistically significant while the result for a second group of schools is also positive but is not statistically significant. If the difference in impacts between the two groups of schools is not statistically significant, one should be especially cautious about concluding that the ERO programs were more effective for some schools than for others.

Appendix Table E.3 displays the results of the composite qualifying statistical tests for the three reading behavior measures discussed in Chapter 5. As discussed above, the composite index was created by averaging the standardized values of the three reading behaviors outcomes: amount of school-related reading, amount of non-school-related reading, and use of reflective reading strategies. Appendix Table E.3 shows results for the full sample of all schools, for each of the two ERO programs separately, and for the various subgroups that are discussed in Chapter 5. None of the estimated impacts on the composite index is statistically significant at the 5 percent level. Thus, readers should exercise caution in interpreting statistically significant impacts for the individual components of the composite index, since these may be due to chance. Appendix Table E.3 also includes the results of the composite qualifying statistical test of the robustness of statistical significance of the *difference in impacts* across subgroups of students or schools. It shows that the difference in impacts is also not statistically significant for the full sample or any of the subgroups. Thus, the difference in impacts should be interpreted cautiously.

[6]Note that one conducts qualifying statistical tests using the composite index when assessing the robustness of impacts for multiple measures across multiple subgroups of the study sample.

The Enhanced Reading Opportunities Study

Appendix Table E.3

Impacts on Reading Behaviors Composite Index, Cohort 2 Respondent Sample and Subgroups

Subgroup	Estimated Impact	P-Value for Estimated Impact
All Schools	0.04	0.250
Programs		
Reading Apprenticeship schools	0.03	0.493
Xtreme Reading schools	0.05	0.323
Difference in impacts	*-0.02*	*0.822*
Baseline comprehension performance		
2.0-3.0 years below grade level	0.06	0.282
3.1-4.0 years below grade level	0.00	0.887
4.1-5.0 years below grade level	0.05	0.318
Difference in impacts, 2.0-3.0 minus 3.1-4.0	*0.06*	*0.533*
Difference in impacts, 2.0-3.0 minus 4.1-5.0	*0.01*	*0.912*
Overage for grade[a]		
Student is overage for grade	0.03	0.654
Student is not overage for grade	0.04	0.262
Difference in impacts	*-0.01*	*0.893*
Language spoken at home		
Students from multilingual families	0.03	0.488
Students from English-only families	0.04	0.387
Difference in impacts	*-0.01*	*0.936*
Teacher experience with the ERO program		
Teachers having taught two full years	0.03	0.465
Teachers having taught less than two full years	0.06	0.324
Difference in impacts	*-0.03*	*0.623*
Teachers having taught two full years - Year 1	0.03	0.407
Teachers having taught two full years - Year 2	0.03	0.465
Difference in impacts	*0.00*	*0.965*
Second-year implementation strength		
Stronger implementation schools[b]	0.03	0.561
Weaker implementation schools[c]	0.04	0.332
Difference in impacts	*-0.01*	*0.896*

(continued)

Appendix Table E.3 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities follow-up student survey.

NOTES: The reading behaviors composite index is the average of the standardized values of the three reading behavior measures: amount of school-related reading, amount of non-school-related reading, and use of reflective reading strategies. The values were standardized using the non-ERO group mean and standard deviation.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment.

A two-tailed t-test was applied to the impact estimate. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating differences.

[a]A student is defined as overage for grade if he or she turned 15 before the start of ninth grade.

[b]The ERO programs in these schools were deemed to have reached an implementation level that was very well aligned to both the classroom learning environment and comprehension instruction dimensions of the program model by the spring site visit, and their ERO program began operating within 2 weeks of the start of the school year.

[c]The implementation fidelity of the ERO programs in these schools was deemed to be less than very well aligned to the classroom learning environment and/or comprehension instruction dimensions of the program model by the spring site visit, and/or their ERO program began operating 2 weeks or more after the start of the school year.

Appendix F

Impact Estimates Weighted for Nonresponse

As discussed in Appendix B, the two main data sources for this second-year impact analysis of the Enhanced Reading Opportunities (ERO) study are the student follow-up survey and the Group Reading Assessment and Diagnostic Examination (GRADE) assessment of student reading skills. The response analysis revealed several differences between students who completed the follow-up test and those who did not. Most notably, there were differences in response rates between the ERO group and the non-ERO group. In addition, nonrespondents were more likely to be overage for the ninth grade and to have lower pretest scores. As a result, students with these characteristics are underrepresented in the sample used to estimate impacts. The over- or underrepresentation of students with certain characteristics in the impact analysis sample may lead to findings that cannot be generalized to the original sample.

This appendix assesses the sensitivity of the impact estimates to the over- or underrepresentation of key baseline characteristics in the impact analysis sample. Specifically, it examines impact estimates that are weighted to account for differential response rates between the ERO and non-ERO groups and across high schools, as well as differential response rates associated with being overage for grade and baseline test scores. Sampling weights were constructed using multiple regressions in which response rates were predicted based on a student's baseline test score and an indicator of whether the student was overage for the ninth grade. Separate regressions were estimated for each high school and for the ERO students and non-ERO students within each school. The sampling weights were constructed as the inverse of the predicted response rate for each student in the full study sample.

These sampling weights ensure that each high school and the ERO and non-ERO groups within each high school can be represented in the impact analysis in the same proportion as they are in the full study sample. They also ensure that the distribution of overage-for-grade and baseline tests scores in the impact sample is equivalent to their representation in the full sample.

Appendix Table F.1 displays the weighted impact estimates for reading achievement for all 34 high schools and for the schools using each of the two supplemental reading programs. It shows that, together, the ERO programs produced a statistically significant weighted impact on reading comprehension of 0.9 standard score point (p-value = 0.035). This is slightly larger than the estimated impact for the respondent sample presented in Chapter 5 (0.8 standard score point; p-value = 0.042). The weighted impact of the Reading Apprenticeship Academic Literacy (RAAL) program on reading comprehension test scores is 1.4 standard score points and statistically significant (p-value = 0.035); this weighted estimate is of the same magnitude as the estimated impact for the respondent sample. The weighted impact of the Xtreme Reading program

The Enhanced Reading Opportunities Study

Appendix Table F.1

Impacts on Reading Achievement Weighted by School Response Rate, Cohort 2 Follow-Up Respondent Sample

Outcome	ERO	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
All schools					
Reading comprehension					
Average standard score	89.9	89.1	0.9 *	0.08 *	0.035
Corresponding grade equivalent	*6.1*	*5.9*			
Corresponding percentile	*24*	*23*			
Reading vocabulary					
Average standard score	93.3	93.4	-0.1	-0.01	0.818
Corresponding grade equivalent	*7.7*	*7.7*			
Corresponding percentile	*32*	*32*			
Sample size	1,264	907			
Reading Apprenticeship schools					
Reading comprehension					
Average standard score	90.0	88.6	1.4 *	0.13 *	0.013
Corresponding grade equivalent	*6.1*	*5.8*			
Corresponding percentile	*25*	*22*			
Reading vocabulary					
Average standard score	93.2	93.7	-0.5	-0.05	0.357
Corresponding grade equivalent	*7.7*	*7.8*			
Corresponding percentile	*31*	*32*			
Sample size	645	470			
Xtreme Reading schools					
Reading comprehension					
Average standard score	89.8	89.5	0.3	0.03	0.630
Corresponding grade equivalent	*6.1*	*6.0*			
Corresponding percentile	*24*	*24*			
Reading vocabulary					
Average standard score	93.3	93.0	0.3	0.03	0.582
Corresponding grade equivalent	*7.7*	*7.7*			
Corresponding percentile	*32*	*31*			
Sample size	619	437			

(continued)

Appendix Table F.1 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study follow-up GRADE assessment.

NOTES: The follow-up GRADE assessment was administered in the spring of 2007 near the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form B). No statistical tests or arithmetic operations were performed on these reference points.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (reading comprehension = 11.294; reading vocabulary = 11.099).

A two-tailed t-test was applied to the impact estimate. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

on reading comprehension is not statistically significant (0.3 standard score point; p-value = 0.630), although it is slightly larger in magnitude than the unweighted impact estimate (0.2 standard score point; p-value = 0.672). Appendix Table F.1 also shows that the ERO programs did not have a statistically significant weighted impact on vocabulary test scores.

Appendix Table F.2 displays the weighted impacts on the reading behavior measures. These results are similar to those estimated for the respondent sample (see Tables 5.3 and 5.4).

In summary, differences between students who completed the follow-up test and survey and those who did not do not appear to change the underlying pattern of impacts on test scores or reading behaviors.

The Enhanced Reading Opportunities Study

Appendix Table F.2

Impacts on Reading Behaviors Weighted by School Response Rate, Cohort 2 Follow-Up Respondent Sample

Outcome	ERO	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
All schools					
Amount of school-related reading (prior month occurrences)	46.56	46.39	0.17	0.00	0.919
Amount of non-school-related reading (prior month occurrences)	30.20	29.80	0.40	0.01	0.781
Use of reflective reading strategies (4-point scale)	2.77	2.71	0.06 *	0.09 *	0.023
Sample size	1,260	901			
Reading Apprenticeship schools					
Amount of school-related reading (prior month occurrences)	50.59	49.37	1.22	0.03	0.616
Amount of non-school-related reading (prior month occurrences)	31.46	32.14	-0.67	-0.02	0.746
Use of reflective reading strategies (4-point scale)	2.78	2.74	0.04	0.07	0.208
Sample size	642	466			
Xtreme Reading schools					
Amount of school-related reading (prior month occurrences)	42.35	43.21	-0.86	-0.02	0.691
Amount of non-school-related reading (prior month occurrences)	28.88	27.23	1.65	0.04	0.401
Use of reflective reading strategies (4-point scale)	2.75	2.67	0.07 *	0.11 *	0.050
Sample size	618	435			

(continued)

Appendix Table F.2 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study follow-up student survey.

NOTES: The student follow-up survey was administered in spring 2007 at the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The estimated impact effect size is calculated as a proportion of the weighted standard deviation of the non-ERO group average (school-related reading standard deviation = 43.462; non-school related reading standard deviation = 37.334; use of reflective reading strategies standard deviation = 0.666).

A two-tailed t-test was applied to the impact estimate. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

For each of the above measures, data are missing for no more than 3 percent of the respondents.

Rounding may cause slight discrepancies in calculating sums and differences.

Appendix G

Impacts on Supplementary Measures of Reading Achievement and Behaviors

In an effort to understand more about the extent and nature of the impacts on student outcomes of the two supplemental literacy programs used in the Enhanced Reading Opportunities (ERO) study — Reading Apprenticeship Academic Literacy (RAAL) and Xtreme Reading — the ERO study team performed secondary impact analyses. These analyses fall into two categories. First, the supplemental analyses explore additional measures from the ERO follow-up student survey. These measures were created to complement the reading behaviors measures discussed in the report. They contribute to a more detailed picture of how the program changed or did not change students' attitudes toward reading and their behavior in school. Second, the study team analyzed the impact of the ERO program on the percentage of students who were less than two years behind grade level in reading by the end of the school year. Given that students needed to be at least two years below grade level in reading to be eligible for the program, those students who have attained reading levels above this cutoff have succeeded in moving beyond the scope of the program during the school year.

Impacts on Students' Attitudes and Perceptions of Reading and School

As discussed in Appendix A, the ERO follow-up student survey included a variety of questions related to students' attitudes and perceptions of reading and school. Beyond the three reading behaviors measures discussed in the report, two other secondary measures were explored, including reading to learn and reading for enjoyment.

These measures are not included in the report because they were less directly related to ERO program goals or less likely to display short-term impacts. Appendix Table G.1 shows the impact findings for each of these measures. As shown in this table, Reading Apprenticeship Academic Literacy (RAAL) had a statistically significant impact on students' enjoyment of reading (effect size of 0.13 standard deviation).

Impacts on the Percentage of Students No Longer Eligible for the ERO Programs

Both RAAL and Xtreme Reading attempt to accelerate literacy learning through their instructional programs to help struggling students attain the reading skill levels needed to succeed in high school classes. One way of measuring the impact of the ERO programs is to look at whether more ERO students are bridging this gap in skills during their first year of high school students who did not participate in ERO. To answer this question, the study team analyzed the programs' impact on the percentage of students who were less than two years behind grade level in reading comprehension by the end of the school year, and, therefore, were no longer eligible for the program. The percentage of ERO students whose score on the follow-up

The Enhanced Reading Opportunities Study

Appendix Table G.1

Impacts on Perceptions of Reading, Cohort 2 Follow-Up Respondent Sample

Outcome	ERO	Non-ERO Group	Estimated Impact	Estimated Impact Effect	P-Value for Estimated Impact
All schools					
Reading to learn (4-point scale)	2.73	2.69	0.04	0.06	0.117
Reading for enjoyment (4-point scale)	2.24	2.19	0.05	0.07	0.108
Sample size	1,260	901			
Reading Apprenticeship schools					
Reading to learn (4-point scale)	2.77	2.73	0.03	0.06	0.294
Reading for enjoyment (4-point scale)	2.32	2.22	0.10 *	0.13 *	0.026
Sample size	642	466			
Xtreme Reading schools					
Reading to learn (4-point scale)	2.69	2.65	0.04	0.07	0.240
Reading for enjoyment (4-point scale)	2.17	2.16	0.00	0.00	0.960
Sample size	618	435			

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study follow-up student survey.

NOTES: The student follow-up survey was administered in spring 2007 at the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (reading to learn standard deviation = 0.591; reading for enjoyment standard deviation = 0.776).

A two-tailed t-test was applied to the impact estimate. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

For each of the above measures, data are missing for no more 1.5 percent than of the respondents.

Rounding may cause slight discrepancies in calculating sums and differences.

GRADE reading comprehension assessment was 98 or above and whose corresponding grade equivalent was greater than 7.9 were compared with the percentage of non-ERO students who scored at or above this level on the GRADE follow-up test.[1] As shown in Appendix Table G.2, the ERO program impacts for the entire sample and for each of the programs are not statistically significant at the 5 percent level. This indicates that a similar percentage of ERO and non-ERO students reached the benchmark level of reading achievement that would make them ineligible to reenroll in the ERO program.

[1]A student who is two years below grade level at the end of ninth grade — and who would therefore still be eligible for the ERO program — reads at a grade equivalent (GE) of 7.9 (= 9.9 –2.0). Hence, in order to no longer be eligible for the program, a student must read at a grade level of at least 8.0 at the end of ninth grade. A GE of 8.0 does not map perfectly onto GRADE standard scores, however. A standard score of 97 translates into a grade equivalent of 7.8, while a score of 98 maps onto a grade equivalent of 8.2. Since a score of 97 would render a student eligible for the program, a standard score of 98 and above is used as the cutoff for defining a student as no longer eligible for the ERO program.

The Enhanced Reading Opportunities Study

Appendix Table G.2

Impacts on Percentage of Students No Longer Eligible for Program, Cohort 2 Follow-Up Respondent Sample

Outcome	ERO	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
All schools					
No longer eligible for program[a] (%)	22.78	19.90	2.88	0.07	0.093
Sample size	1,264	907			
Reading Apprenticeship schools					
No longer eligible for program (%)	22.79	18.78	4.01	0.10	0.090
Sample size	645	470			
Xtreme Reading schools					
No longer eligible for program (%)	22.78	21.05	1.73	0.04	0.487
Sample size	619	437			

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study follow-up GRADE assessment.

NOTES: The follow-up GRADE assessment was administered in the spring of 2007 near the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (standard deviation = 40.317).

A two-tailed t-test was applied to the impact estimate. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

[a]Students with scores on the GRADE pretest between 2 and 5 years below grade level were eligible for the program. Students are considered no longer eligible for the program if their score on the follow-up GRADE assessment corresponded to a grade equivalent greater than 7.9 (standard score of 98), suggesting that the student is now less than two years behind grade level.

Appendix H

Impacts for Student Subgroups

While all students in the study sample for the Enhanced Reading Opportunities (ERO) evaluation had baseline reading comprehension skills that were two to five years below grade level in the spring of eighth grade, the ERO study sample includes a diverse population of students. With this diversity in mind, the ERO evaluation was designed to allow for the estimation of impacts for key subgroups of students who face especially challenging barriers to literacy development and overall school performance in high school. For example, prior research has shown that especially low literacy levels, evidence of failure in prior grades, and having English as a second language are powerful predictors of school success.[1]

This appendix examines variation in ERO program impacts for subgroups of students defined by their baseline reading comprehension test scores, whether they were overage for the ninth grade, and whether a language other than English was spoken in their homes. Among the respondent sample, 39 percent had baseline test scores that indicate reading levels that were four to five years below grade level at the times of testing, and another 29 percent were reading from three to four years below grade level. Also, 26 percent of the students in the study sample were overage for the ninth grade, which is used to indicate that a student was retained in a prior grade. Half of the students in the sample lived in households where a language other than English was spoken.

- **Differences in impacts across subgroups of students with different baseline reading comprehension test scores are not statistically significant for reading comprehension and reading behaviors, but are statistically significant for reading vocabulary.**

Appendix Tables H.1 and H.2 correspond to the top panel of Table 5.5 and present impact findings for the subgroups of students defined by their baseline reading comprehension test scores. Appendix Table H.1 indicates that the ERO programs did not produce statistically significant impacts on reading comprehension or reading vocabulary test scores for any of the three subgroups defined by baseline test scores.

That said, the bottom section of Appendix Table H.1 indicates that the impact of the programs on reading vocabulary scores differs by a statistically significant amount across some of the subgroups. Specifically, the estimated impact on reading vocabulary scores for students who were two to three years below grade level at baseline is significantly larger than the estimated impact for students who were four to five years below grade level at baseline (difference in effect size = 0.23 standard deviation; p-value = 0.019).

[1]Roderick (1993); Fine (1988).

The Enhanced Reading Opportunities Study

Appendix Table H.1

Impacts on Reading Achievement, Cohort 2 Follow-Up Respondent Sample, by Baseline Reading Comprehension Performance

Outcome	ERO	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
2.0-3.0 years below grade level					
Reading comprehension					
Average standard score	94.3	93.4	0.9	0.09	0.170
Corresponding grade equivalent	*7.2*	*7.0*			
Corresponding percentile	*34*	*33*			
Reading vocabulary					
Average standard score	97.7	96.5	1.2	0.12	0.088
Corresponding grade equivalent	*8.5*	*8.2*			
Corresponding percentile	*42*	*39*			
Sample size	404	301			
3.1-4.0 years below grade level					
Reading comprehension					
Average standard score	91.1	90.2	0.9	0.09	0.273
Corresponding grade equivalent	*6.3*	*6.1*			
Corresponding percentile	*26*	*25*			
Reading vocabulary					
Average standard score	94.3	93.9	0.4	0.04	0.625
Corresponding grade equivalent	*7.8*	*7.8*			
Corresponding percentile	*34*	*33*			
Sample size	350	269			
4.1-5.0 years below grade level					
Reading comprehension					
Average standard score	86.1	85.3	0.8	0.08	0.210
Corresponding grade equivalent	*5.1*	*5.0*			
Corresponding percentile	*17*	*15*			
Reading vocabulary					
Average standard score	89.5	90.6	-1.1	-0.11	0.106
Corresponding grade equivalent	*7.1*	*7.3*			
Corresponding percentile	*23*	*26*			
Sample size	510	337			

(continued)

Appendix Table H.1 (continued)

Difference in Impacts Between Subgroups	Difference in Impacts	Difference in Impact Effect Sizes	P-Value for Difference
2.0-3.0 minus 3.1-4.0			
Reading comprehension standard score	0.0	0.00	0.968
Reading vocabulary standard score	0.8	0.09	0.414
2.0-3.0 minus 4.1-5.0			
Reading comprehension standard score	0.1	0.01	0.897
Reading vocabulary standard score	2.3 *	0.23 *	0.019

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study follow-up GRADE assessment.

NOTES: The follow-up GRADE assessment was administered in the spring of 2007 near the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form B). No statistical tests or arithmetic operations were performed on these reference points.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (reading comprehension = 10.035; reading vocabulary = 9.827).

A two-tailed t-test was applied to the impact estimate and to the difference in impacts. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

The Enhanced Reading Opportunities Study

Appendix Table H.2

Impacts on Reading Behaviors, Cohort 2 Follow-Up Respondent Sample, by Baseline Reading Comprehension Performance

Outcome	ERO	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
2.0-3.0 years behind grade level					
Amount of school-related reading (prior month occurrences)	47.1	47.3	-0.3	-0.01	0.925
Amount of non-school-related reading (prior month occurrences)	28.7	26.4	2.3	0.07	0.353
Use of reflective reading strategies (4-point scale)	2.8	2.7	0.1	0.13	0.070
Sample size	403	299			
3.1-4.0 years behind grade level					
Amount of school-related reading (prior month occurrences)	48.0	46.8	1.1	0.03	0.711
Amount of non-school-related reading (prior month occurrences)	30.4	31.9	-1.4	-0.04	0.595
Use of reflective reading strategies (4-point scale)	2.8	2.8	0.0	0.05	0.552
Sample size	347	268			
4.1-5.0 years behind grade level					
Amount of school-related reading (prior month occurrences)	44.5	44.3	0.3	0.01	0.920
Amount of non-school-related reading (prior month occurrences)	30.4	29.2	1.3	0.04	0.582
Use of reflective reading strategies (4-point scale)	2.7	2.6	0.1	0.12	0.119
Sample size	510	334			

(continued)

Appendix Table H.2 (continued)

Difference in Impacts Between Subgroups	Difference in Impacts	Difference in Impact Effect Sizes	P-Value for Difference
2.0-3.0 minus 3.1-4.0			
Amount of school-related reading	-1.4	-0.04	0.739
Amount of non-school-related reading	3.7	0.11	0.307
Use of reflective reading strategies	0.0	0.08	0.445
2.0-3.0 minus 4.1-5.0			
Amount of school-related reading	-0.5	-0.01	0.891
Amount of non-school-related reading	1.1	0.03	0.755
Use of reflective reading strategies	0.0	0.01	0.924

SOURCE: MDRC calculations from the Enhanced Reading Opportunities follow-up student survey.

NOTES: The student follow-up survey was administered in spring 2007 at the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (school-related reading standard deviation = 38.322; non-school-related reading standard deviation = 32.976; use of reflective reading strategies standard deviation = 0.592).

A two-tailed t-test was applied to the impact estimate and to the difference in impacts. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

For each of the above measures, data are missing for no more than 3.4 percent of the respondents.

Rounding may cause slight discrepancies in calculating sums and differences.

Appendix Table H.2 shows that the ERO programs had a statistically significant impact on the use of reflective reading strategies for students whose scores were from four to five years below grade level. However, it cannot be concluded that the impacts of the ERO programs on this measure was different for students with different baseline reading comprehension test scores for two reasons. First, the qualifying tests conducted for this subgroup of students (see Appendix E) indicate that the ERO programs did not produce a statistically significant impact on the composite index of the three reading behavior measures. Second, the difference between the impact on reading strategies for this subgroup and the impact for each of the other two subgroups is not statistically significant.

- **Differences in impacts across subgroups of students who were overage for the ninth grade or not overage for the ninth grade are not statistically significant.**

Appendix Tables H.3 and H.4 correspond to the middle panel of Table 5.5 and present impact findings for the subgroups of students defined by whether they were overage for the ninth grade. Appendix Table H.3 indicates that the ERO programs produced positive and statistically significant impacts on reading comprehension test scores for students who were not overage for grade (effect size = 0.10 standard deviation; p-value = 0.028). Although the impact on reading comprehension test scores for this group is statistically significant, the difference between this impact and the impact for students who were overage for grade is not statistically significant. Appendix Table H.4 shows that the ERO programs also produced a statistically significant impact on the use of reflective reading strategies for students who were not overage for grade (effect size = 0.10 standard deviation; p-value = 0.034). However, the difference between this impact and the impact for students who were overage for grade is not statistically significant. Thus, it cannot be concluded that the impacts of the ERO programs on these measures were different for students who were not overage for grade compared to those who were. Moreover, the qualifying tests conducted for students who are not overage for grade (see Appendix E) indicate that the ERO programs did not produce a statistically significant impact for this subgroup on the composite index of the three reading behavior measures.

- **Differences in impacts across subgroups of students from multilingual families and those from English-only families are not statistically significant.**

Appendix Tables H.5 and H.6 correspond to the bottom panel of Table 5.5 and present impact findings for the subgroups of students defined by whether a language other than English was spoken in their homes. Appendix Table H.5 indicates that the ERO programs did not produce statistically significant impacts on reading comprehension test scores for either of the two

subgroups defined by language spoken at home. Similarly, Appendix Table H.6 indicates that the ERO programs did not have a statistically significant impact on reading behaviors for students from multilingual families or from English-only families.

An overall F-test was used to test whether there is systematic variation in impacts on reading comprehension across the various subgroups defined by baseline reading comprehension test scores, overage status, and language spoken at home.[2] This test indicates that the overall variation in impacts across all of these subgroups is not statistically significant (F-statistic = 1.242; p-value = 0.291), further suggesting that the statistical significance of reading comprehension impacts for specific subgroups should be interpreted with caution.[3]

[2]This test was conducted by adding a set of interactions between the treatment indicator and the subgroup indicators (two subgroup indicators for baseline reading comprehension score, an indicator for overage status, and an indicator for language spoken at home) to the impact regression model fit to the Cohort 2 follow-up respondent sample. An F-test was then used to test for the joint significance of these interaction terms.

[3]This overall F-test was also conducted for reading vocabulary, the three reading behaviors outcomes, and the reading behavior composite. These tests indicate that there is overall variation in impacts across the student subgroups for reading vocabulary (F-statistic = 3.69; p-value = 0.005), which is partially driven by the difference in impacts across subgroups defined by reading comprehension scores at baseline. There is also overall variation in impacts on the frequency of school-related reading (F-statistic = 2.70; p-value = 0.029), and the reading behaviors composite (F-statistic = 2.64; p-value = 0.032).

The Enhanced Reading Opportunities Study

Appendix Table H.3

Impacts on Reading Achievement, Cohort 2 Follow-Up Respondent Sample, by Whether Students Were Overage for Grade

Outcome	ERO Group	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
Overage for grade[a]					
Reading comprehension					
Average standard score	87.4	87.0	0.5	0.05	0.592
Corresponding grade equivalent	*5.4*	*5.3*			
Corresponding percentile	*19*	*18*			
Reading vocabulary					
Average standard score	91.3	91.2	0.1	0.01	0.939
Corresponding grade equivalent	*7.5*	*7.4*			
Corresponding percentile	*27*	*27*			
Sample size	323	245			
Not overage for grade					
Reading comprehension					
Average standard score	91.0	90.0	1.0 *	0.10 *	0.028
Corresponding grade equivalent	*6.3*	*6.1*			
Corresponding percentile	*26*	*25*			
Reading vocabulary					
Average standard score	94.2	94.2	0.0	0.00	0.996
Corresponding grade equivalent	*7.8*	*7.8*			
Corresponding percentile	*33*	*33*			
Sample size	941	662			

Difference in Impacts Between Subgroups	Difference in Impacts	Difference in Impact Effect Sizes	P-Value for Difference
Overage minus not overage			
Reading comprehension standard score	-0.5	-0.05	0.587
Reading vocabulary standard score	0.1	0.01	0.949

(continued)

Appendix Table H.3 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study follow-up GRADE assessment.

NOTES: The follow-up GRADE assessment was administered in the spring of 2007 near the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form B). No statistical tests or arithmetic operations were performed on these reference points.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (reading comprehension = 10.035; reading vocabulary = 9.827).

A two-tailed t-test was applied to the impact estimate and to the difference in impacts. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

[a]A student is defined as overage for grade if he or she turned 15 before the start of ninth grade.

The Enhanced Reading Opportunities Study

Appendix Table H.4

Impacts on Reading Behaviors, Cohort 2 Follow-Up Respondent Sample, by Whether Students Were Overage for Grade

Outcome	ERO Group	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
Overage for grade[a]					
Amount of school-related reading (prior month occurrences)	45.4	44.7	0.7	0.02	0.841
Amount of non-school-related reading (prior month occurrences)	31.7	31.5	0.2	0.01	0.940
Use of reflective reading strategies (4-point scale)	2.8	2.7	0.0	0.05	0.585
Sample size	320	240			
Not overage for grade					
Amount of school-related reading (prior month occurrences)	46.6	46.3	0.2	0.01	0.891
Amount of non-school-related reading (prior month occurrences)	29.3	28.4	0.8	0.03	0.598
Use of reflective reading strategies (4-point scale)	2.8	2.7	0.1 *	0.10 *	0.034
Sample size	940	661			

Difference in Impacts Between Subgroups	Difference in Impacts	Difference in Impact Effect Sizes	P-Value for Difference
Overage minus not overage			
Amount of school-related reading	0.4	0.01	0.910
Amount of non-school-related reading	-0.6	-0.02	0.863
Use of reflective reading strategies	0.0	-0.06	0.574

(continued)

Appendix Table H.4 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities follow-up student survey.

NOTES: The student follow-up survey was administered in spring 2007 at the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (school-related reading standard deviation = 38.322; non-school-related reading standard deviation = 32.976; use of reflective reading strategies standard deviation = 0.592).

A two-tailed t-test was applied to the impact estimate and to the difference in impacts. Statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

For each of the above measures, data are missing for no more than 4.2 percent of the respondents.

Rounding may cause slight discrepancies in calculating sums and differences.

[a]A student is defined as overage for grade if he or she turned 15 before the start of ninth grade.

The Enhanced Reading Opportunities Study

Appendix Table H.5

Impacts on Reading Achievement, Cohort 2 Follow-Up Respondent Sample, by Language Spoken at Home

Outcome	ERO Group	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
Students from multilingual families					
Reading comprehension					
Average standard score	89.6	88.8	0.8	0.08	0.181
Corresponding grade equivalent	*6.0*	*5.8*			
Corresponding percentile	*24*	*22*			
Reading vocabulary					
Average standard score	92.0	92.2	-0.2	-0.02	0.768
Corresponding grade equivalent	*7.6*	*7.6*			
Corresponding percentile	*29*	*29*			
Sample size	621	458			
Students from English-only families					
Reading comprehension					
Average standard score	90.6	89.7	0.9	0.09	0.107
Corresponding grade equivalent	*6.2*	*6.0*			
Corresponding percentile	*25*	*24*			
Reading vocabulary					
Average standard score	94.9	94.5	0.3	0.03	0.550
Corresponding grade equivalent	*7.9*	*7.9*			
Corresponding percentile	*35*	*34*			
Sample size	643	449			

Difference in Impacts Between Subgroups	Difference in Impacts	Difference in Impact Effect Sizes	P-Value for Difference
Multilingual minus English-only			
Reading comprehension standard score	-0.1	-0.01	0.868
Reading vocabulary standard score	-0.5	-0.05	0.531

(continued)

Appendix Table H.5 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study follow-up GRADE assessment.

NOTES: The follow-up GRADE assessment was administered in the spring of 2007 near the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form B). No statistical tests or arithmetic operations were performed on these reference points.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (reading comprehension = 10.035; reading vocabulary = 9.827).

A two-tailed t-test was applied to the impact estimate and to the difference in impacts. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

The Enhanced Reading Opportunities Study

Appendix Table H.6

Impacts on Reading Behaviors, Cohort 2 Follow-Up Respondent Sample, by Language Spoken at Home

Outcome	ERO	Non-ERO Group	Estimated Impact	Estimated Impact Effect	P-Value for Estimated Impact
Students from multilingual families					
Amount of school-related reading (prior month occurrences)	47.4	45.9	1.5	0.04	0.532
Amount of non-school-related reading (prior month occurrences)	30.3	30.8	-0.6	-0.02	0.778
Use of reflective reading strategies (4-point scale)	2.8	2.7	0.0	0.07	0.224
Sample size	621	455			
Students from English-only families					
Amount of school-related reading (prior month occurrences)	45.2	46.1	-0.8	-0.02	0.709
Amount of non-school-related reading (prior month occurrences)	29.5	28.0	1.4	0.04	0.458
Use of reflective reading strategies (4-point scale)	2.8	2.7	0.1	0.10	0.090
Sample size	639	446			

Difference in Impacts Between Subgroups	Difference in Impacts	Difference in Impact Effect Sizes	P-Value for Difference
Multilingual minus English-only			
Amount of school-related reading	2.3	0.06	0.476
Amount of non-school-related reading	-2.0	-0.06	0.473
Use of reflective reading strategies	0.0	-0.03	0.765

(continued)

Appendix Table H.6 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities follow-up student survey.

NOTES: The student follow-up survey was administered in spring 2007 at the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (school-related reading standard deviation = 38.322; non-school-related reading standard deviation = 32.976; use of reflective reading strategies standard deviation = 0.592).

A two-tailed t-test was applied to the impact estimate and to the difference in impacts. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

For each of the above measures, data are missing for no more than 3.4 percent of the respondents.

Rounding may cause slight discrepancies in calculating sums and differences.

Appendix I

The Relationship Between Impacts and Second-Year Implementation

This appendix presents results from exploratory analyses that examine the relationship between school-level impacts and various aspects of implementation in the second year of the Enhanced Reading Opportunities (ERO) study. As noted in Chapter 5, school-to-school variation in impacts was not statistically significant in the second year of the study; hence, these exploratory analyses are unlikely to provide conclusive information about the factors that predict differences in impacts across schools. Nonetheless, the relationship between impacts and certain policy-relevant aspects of implementation was examined and the results of these analyses are presented in this appendix.

Given that the supplemental literacy programs used in the ERO study — Reading Apprenticeship Academic Literacy (RAAL) and Xtreme Reading — operated in more favorable conditions in the second year of the study than in the first (as discussed in Chapter 3), this exploratory analysis focuses on the relationship between school-level impacts and whether a school's ERO program was particularly exemplary in terms of its implementation. This stands in contrast to the exploratory analysis conducted in the first-year report, which focuses instead on the relationship between school-level impacts and program implementation that was especially problematic.

The first set of results in this appendix examines the relationship between impacts and the experience of a school's ERO teacher with the program. Specifically, this analysis investigates whether the impact of the programs in the second year of the study differs between the 25 schools whose ERO teacher had also taught the entire first year, and the 9 schools whose ERO teacher was new to the program in the second year of the study or had been brought in as a replacement midway through the first year. In addition, the analysis compares the impacts produced by the 25 returning teachers in their second year of teaching the program (Cohort 2) and the impacts produced in their first year of teaching the program (Cohort 1).

The second set of results in this appendix examines the relationship between impacts and the strength of program implementation. Two types of relationship are explored: (1) the relationship between impacts and implementation fidelity and (2) the relationship between impacts and the number of weeks between the start of the school year and program start-up. The indicators are then combined, and impacts are then estimated for two groups of sites defined by whether the implementation of their ERO program was strong on both of these indicators.

It is important to note that the analyses presented in this appendix are exploratory in nature and cannot be used to establish causal links between these aspects of implementation and variation in program impacts across the sites.

Variation in Impacts Across Schools

Appendix Table I.1 is the counterpart to Figure 5.2. It presents the reading comprehension impact estimates of each of the 34 participating high schools in ascending order. It also includes the standard error and 95 percent confidence intervals for these impacts. One of the 34 schools had a statistically significant positive impact. A composite F-test was used to assess whether the school-level impacts on reading comprehension test scores are statistically equivalent. The F-value is 0.91, and the p-value is 0.614, indicating that the school-to-school variation in impacts may have occurred by chance.

Relationship Between Impacts and Teacher Experience with the ERO Program

Impacts by ERO Teacher's Experience in the Second Year

Appendix Tables I.2 and I.3 correspond with the upper half of the top panel of Table 5.6. These tables present impacts on reading test scores and reading behaviors, consecutively, for the 25 schools whose ERO teacher in the second year of the study had also taught the entire first year of the study (that is, teachers having taught two full years of the program by the end of the study) and for the 9 schools whose ERO teacher was either new at the start of the second year or had replaced another teacher midway through the first year of the study (that is, teachers having taught less than two full years of the program by the end of the study).

Appendix Table I.2 indicates that while the impact of the ERO programs was positive and statistically significant in sites where the ERO teacher had greater experience with the program (effect size = 0.09 standard deviation; p-value = 0.050), this impact is not statistically different from the impact for less experienced teachers (effect size = 0.06 standard deviation; p-value = 0.487). Hence, it cannot be concluded that impacts were larger in sites with more experienced teachers.

Similarly, Appendix Table I.3 shows that the impact of the ERO programs on the use of reflective reading strategies was also positive and statistically significant in schools where the ERO teacher had more experience with the program (effect size = 0.10 standard deviation; p-value = 0.046); however, this impact is not statistically different from the impact of the ERO programs on reading strategies in schools where the ERO teacher was less experienced (effect size = 0.07 standard deviation; p-value = 0.441). Here, again, it cannot be concluded that impacts were larger in sites with more experienced teachers. In addition, the statistical significance of the impact among more experienced teachers is called into question by the qualifying tests in Appendix E, which shows that the impact of the programs on the reading behavior composite index in sites with a more experienced ERO teacher is not statistically significant.

The Enhanced Reading Opportunities Study

Appendix Table I.1

Fixed-Effect Impact Estimates on Reading Comprehension, by School

Variable	Impact Estimate	Standard Error	95% Confidence Interval	
School 1[a]	-3.7	2.74	-9.11	1.65
School 2	-3.6	2.39	-8.33	1.04
School 3	-2.5	2.37	-7.19	2.11
School 4	-2.5	2.07	-6.57	1.55
School 5	-1.9	2.48	-6.73	3.00
School 6	-1.7	2.63	-6.82	3.50
School 7	-1.3	1.83	-4.88	2.29
School 8	-0.7	2.15	-4.89	3.53
School 9	-0.4	2.13	-4.61	3.74
School 10	0.1	2.24	-4.33	4.45
School 11	0.4	2.44	-4.39	5.20
School 12	0.4	2.10	-3.70	4.53
School 13	0.5	2.00	-3.46	4.37
School 14	0.7	2.33	-3.89	5.25
School 15	0.8	2.48	-4.06	5.68
School 16	1.0	2.33	-3.53	5.62
School 17	1.1	2.18	-3.17	5.37
School 18	1.6	2.31	-2.93	6.12
School 19	1.6	2.21	-2.71	5.94
School 20	1.6	2.90	-4.05	7.32
School 21	1.7	2.11	-2.44	5.85
School 22	1.8	2.55	-3.17	6.82
School 23	2.1	2.22	-2.31	6.42
School 24	2.1	2.58	-2.97	7.13
School 25	2.4	2.40	-2.33	7.07
School 26	2.5	2.28	-1.96	6.99
School 27	2.7	2.21	-1.62	7.03

(continued)

Appendix Table I.1 (continued)

Variable	Impact Estimate	Standard Error	95% Confidence Interval	
School 28	2.9	4.52	-6.00	11.72
School 29	3.3	2.66	-1.96	8.48
School 30	3.9	2.94	-1.91	9.63
School 31	4.0	2.59	-1.03	9.13
School 32	5.3 *	2.13	1.09	9.43
School 33	5.6	3.50	-1.24	12.50
School 34	6.2	3.89	-1.43	13.81

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study follow-up GRADE assessment.

NOTES: The follow-up GRADE assessment was administered in the spring of 2007 near the end of students' ninth-grade year.

The fixed-effect estimated impacts are the regression-adjusted coefficients on the interaction between school and treatment using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment.

A two-tailed t-test was applied to the impact estimate. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

A composite F-test was used to assess whether the school-level impacts on reading comprehension test scores are statistically equivalent. The F-value is 0.91, and the p-value is 0.614, indicating that the school-to-school variation in impacts is likely to have occurred by chance.

[a]The schools are listed in ascending order by their impact estimate.

The Enhanced Reading Opportunities Study

Appendix Table I.2

Impacts on Reading Achievement, Cohort 2 Follow-Up Respondent Sample, by Teacher Experience with the ERO Program

Outcome	ERO Group	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
Teachers having taught two full years					
Reading comprehension					
Average standard score	90.0	89.1	0.9 *	0.09 *	0.050
Corresponding grade equivalent	*6.1*	*5.9*			
Corresponding percentile	*25*	*23*			
Reading vocabulary					
Average standard score	93.1	93.3	-0.1	-0.01	0.791
Corresponding grade equivalent	*7.7*	*7.7*			
Corresponding percentile	*31*	*32*			
Sample size	942	680			
Teachers having taught less than two full years[a]					
Reading comprehension					
Average standard score	90.4	89.8	0.6	0.06	0.487
Corresponding grade equivalent	*6.2*	*6.1*			
Corresponding percentile	*25*	*24*			
Reading vocabulary					
Average standard score	94.4	94.1	0.3	0.03	0.694
Corresponding grade equivalent	*7.8*	*7.8*			
Corresponding percentile	*34*	*33*			
Sample size	322	227			

Difference in Impacts Between Subgroups	Difference in Impacts	Difference in Impact Effect Sizes	P-Value for Difference
Entire year minus new or less than a year			
Reading comprehension standard score	0.4	0.04	0.710
Reading vocabulary standard score	-0.5	-0.05	0.636

(continued)

Appendix Table I.2 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study follow-up GRADE assessment.

NOTES: The follow-up GRADE assessment was administered in the spring of 2007 near the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form B). No statistical tests or arithmetic operations were performed on these reference points.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (reading comprehension = 10.035; reading vocabulary = 9.827).

A two-tailed t-test was applied to the impact estimate and to the difference in impacts. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

[a]Includes both ERO teachers who started in Year 2, and replacement teachers who began mid-year in Year 1.

The Enhanced Reading Opportunities Study

Appendix Table I.3

Impacts on Reading Behaviors, Cohort 2 Follow-Up Respondent Sample, by Teacher Experience with the ERO Program

Outcome	ERO	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
Teachers having taught two full years					
Amount of school-related reading (prior month occurrences)	45.6	46.6	-1.0	-0.03	0.590
Amount of non-school-related reading (prior month occurrences)	29.2	28.8	0.5	0.01	0.774
Use of reflective reading strategies (4-point scale)	2.8	2.7	0.06 *	0.10 *	0.046
Sample size	941	678			
Teachers having taught less than two full years[a]					
Amount of school-related reading (prior month occurrences)	48.3	43.9	4.4	0.11	0.150
Amount of non-school-related reading (prior month occurrences)	31.7	31.2	0.6	0.02	0.843
Use of reflective reading strategies (4-point scale)	2.8	2.7	0.04	0.07	0.441
Sample size	319	223			

Difference in Impacts Between Subgroups	Difference in Impacts	Difference in Impact Effect Sizes	P-Value for Difference
Entire year minus new or less than a year			
Amount of school-related reading	-5.4	-0.14	0.131
Amount of non-school-related reading	-0.1	0.00	0.976
Use of reflective reading strategies	0.02	0.03	0.765

(continued)

Appendix Table I.3 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities follow-up student survey.

NOTES: The student follow-up survey was administered in spring 2007 at the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (school-related reading standard deviation = 38.322; non-school-related reading standard deviation = 32.976; use of reflective reading strategies standard deviation = 0.592).

A two-tailed t-test was applied to the impact estimate and to the difference in impacts. Statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

For each of the above measures, data are missing for no more than 3.2 percent of the respondents.

Rounding may cause slight discrepancies in calculating sums and differences.

[a]Includes both ERO teachers who started in Year 2, and replacement teachers who began mid-year in Year 1.

Impacts by Study Year for Teachers Having Taught Both Years

Appendix Tables I.4 and I.5 correspond with the lower half of the top panel of Table 5.6. These tables present impacts on reading test scores and reading behaviors, consecutively, for the 25 schools whose ERO teacher taught two full years of the program, during each of the two years that they taught the program.

Appendix Table I.4 indicates that the impact produced by returning teachers on reading comprehension scores was positive and statistically significant in both years (effect size = 0.11 standard deviation in the first year and 0.09 standard deviation in the second year). However, the difference between these two impacts is not statistically significant.

Appendix Table I.5 shows that returning ERO teachers produced a larger impact on students' use of reading strategies in their second year of teaching the program (effect size = 0.10 standard deviation; p-value = 0.046) than in their first year (effect size = –0.05; p-value = 0.331) and that the difference between these impacts is statistically significant. Given that the composite qualifying tests in Appendix E indicate that the difference in impacts between Year 1 and Year 2 on the composite measure of reading behaviors is not statistically significant, one cannot be certain that the statistical significance of this finding is not the result of chance.

Relationship Between Impacts and Program Implementation

Relationship Between Impacts and Implementation Fidelity

Appendix Tables I.6 and I.7 correspond with the second panel of Table 5.6. These tables present impacts on reading test scores and reading behaviors, consecutively, by implementation fidelity at the second-year spring site visit. Schools are categorized as having ERO programs that are either "very well aligned," "well aligned," "moderately aligned," or "poorly aligned" with their respective program models (see Chapter 3 and Appendix D for the definition of these categories). Note that impacts are not presented for the "poorly aligned" category because only one site was deemed to have poorly aligned implementation fidelity at the spring site visit. In particular, one cannot make generalized statements about the impact produced by schools with "poorly aligned" programs based on only one school rated in that category. The results are also suppressed to protect the identity of this particular school and its associated impact.[1]

[1]Given that there was only one school whose implementation fidelity was poorly aligned to the program models, one possibility would have been to combine the "moderately aligned" and "poorly aligned" schools into one category. However, these two categories were kept separate in order to preserve the same fidelity groupings that were used in the first-year report of the ERO study (see Kemple et al. 2008), as well as in Chapter 3 of the present report.

The Enhanced Reading Opportunities Study

Appendix Table I.4

Impacts on Reading Achievement in Schools Where Teacher Taught Two Full Years of the ERO Program, by Cohort Respondent Sample

Outcome	ERO Group	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
Cohort 1					
Reading comprehension					
Average standard score	90.7	89.6	1.1 *	0.11 *	0.014
Corresponding grade equivalent	*6.2*	*6.0*			
Corresponding percentile	*26*	*24*			
Reading vocabulary					
Average standard score	93.6	93.1	0.5	0.05	0.274
Corresponding grade equivalent	*7.8*	*7.7*			
Corresponding percentile	*32*	*31*			
Sample size	1040	748			
Cohort 2					
Reading comprehension					
Average standard score	90.0	89.1	0.9 *	0.09 *	0.050
Corresponding grade equivalent	*6.1*	*5.9*			
Corresponding percentile	*25*	*23*			
Reading vocabulary					
Average standard score	93.1	93.3	-0.1	-0.01	0.791
Corresponding grade equivalent	*7.7*	*7.7*			
Corresponding percentile	*31*	*32*			
Sample size	942	680			

Difference in Impacts Between Cohorts	Difference in Impacts	Difference in Impact Sizes	P-Value for Difference
Cohort 2 minus Cohort 1			
Reading comprehension standard score	-0.2	-0.02	0.782
Reading vocabulary standard score	-0.6	-0.06	0.341

(continued)

Appendix Table I.4 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study follow-up GRADE assessment.

NOTES: The follow-up GRADE assessment was administered in the spring of 2007 near the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form B). No statistical tests or arithmetic operations were performed on these reference points.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (reading comprehension = 10.035; reading vocabulary = 9.827).

A two-tailed t-test was applied to the impact estimate and to the difference in impacts. The statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

The Enhanced Reading Opportunities Study

Appendix Table I.5

Impacts on Reading Behaviors in Schools Where Teacher Taught Two Full Years of the ERO Program, by Cohort Respondent Sample

Outcome	ERO	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
Cohort 1					
Amount of school-related reading (prior month occurrences)	45.3	43.3	2.0	0.05	0.328
Amount of non-school-related reading (prior month occurrences)	27.5	25.1	2.4	0.07	0.095
Use of reflective reading strategies (4-point scale)	2.61	2.64	-0.03	-0.05	0.331
Sample size	1039	747			
Cohort 2					
Amount of school-related reading (prior month occurrences)	45.6	46.6	-1.0	-0.03	0.590
Amount of non-school-related reading (prior month occurrences)	29.2	28.8	0.5	0.01	0.774
Use of reflective reading strategies (4-point scale)	2.76	2.70	0.06 *	0.10 *	0.046
Sample size	941	678			

Difference in Impacts Between Subgroups	Difference in Impacts	Difference in Impact Effect Sizes	P-Value for Difference
Cohort 2 minus Cohort 1			
Amount of school-related reading	-3.1	-0.08	0.276
Amount of non-school-related reading	-2.0	-0.06	0.366
Use of reflective reading strategies	0.09 *	0.15 *	0.038

(continued)

Appendix Table I.5 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities follow-up student survey.

NOTES: The student follow-up survey was administered in spring 2007 at the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (school-related reading standard deviation = 38.322; non-school-related reading standard deviation = 32.976; use of reflective reading strategies standard deviation = 0.592).

A two-tailed t-test was applied to the impact estimate and to the difference in impacts. Statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

For each of the above measures, data are missing for no more than 3.6 percent of the respondents.

Rounding may cause slight discrepancies in calculating sums and differences.

The Enhanced Reading Opportunities Study

Appendix Table I.6

Impacts on Reading Achievement, Cohort 2 Follow-Up Respondent Sample, by Program Implementation Fidelity at Spring Site Visit

Outcome	ERO	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
Very well-aligned implementation[a]					
Reading comprehension					
Average standard score	90.0	88.7	1.3 *	0.13 *	0.047
Corresponding grade equivalent	*6.1*	*5.8*			
Corresponding percentile	*25*	*22*			
Reading vocabulary					
Average standard score	93.2	93.2	0.0	0.00	0.984
Corresponding grade equivalent	*7.7*	*7.7*			
Corresponding percentile	*31*	*31*			
Sample size	506	370			
Well-aligned implementation					
Reading comprehension					
Average standard score	90.1	89.6	0.6	0.06	0.417
Corresponding grade equivalent	*6.1*	*6.0*			
Corresponding percentile	*25*	*24*			
Reading vocabulary					
Average standard score	93.0	92.9	0.1	0.01	0.897
Corresponding grade equivalent	*7.7*	*7.7*			
Corresponding percentile	*31*	*31*			
Sample size	430	293			
Moderately aligned implementation					
Reading comprehension					
Average standard score	89.9	89.7	0.2	0.02	0.837
Corresponding grade equivalent	*6.1*	*6.0*			
Corresponding percentile	*24*	*24*			
Reading vocabulary					
Average standard score	94.1	94.4	-0.2	-0.03	0.757
Corresponding grade equivalent	*7.8*	*7.8*			
Corresponding percentile	*33*	*34*			
Sample size	307	236			

(continued)

Appendix Table I.6 (continued)

Outcome	ERO	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
Poorly aligned implementation					
Reading comprehension					
Average standard score	–	–	–	–	–
Corresponding grade equivalent					
Corresponding percentile					
Reading vocabulary					
Average standard score	–	–	–	–	–
Corresponding grade equivalent					
Corresponding percentile					
Sample size	–	–			

Difference in Impacts Between Subgroups	Difference in Impacts	Difference in Impact Effect Sizes	P-Value for Difference
Very well aligned minus moderately aligned			
Reading comprehension standard score	1.1	0.11	0.272
Reading vocabulary standard score	0.3	0.03	0.798
Very well aligned minus well aligned			
Reading comprehension standard score	0.7	0.07	0.467
Reading vocabulary standard score	-0.1	-0.01	0.933

(continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study follow-up GRADE assessment.

NOTES: The follow-up GRADE assessment was administered in the spring of 2007 near the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form B). No statistical tests or arithmetic operations were performed on these reference points.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (reading comprehension = 10.035; reading vocabulary = 9.827).

A two-tailed t-test was applied to the impact estimate and to the difference in impacts. Statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Appendix Table I.6 (continued)

Rounding may cause slight discrepancies in calculating sums and differences.

[a]The fidelity of program implementation is measured on two dimensions: learning environment and comprehension instruction. On each dimension, schools were measured in terms of their depth of alignment to the program model. Schools that were very well aligned on both dimensions (i.e., with average scores greater or equal to 2.5 on both dimensions) are categorized as having "very well-aligned implementation." Schools that were well aligned to both dimensions are categorized as being "well-aligned." Schools that were moderately aligned to at least one dimension and moderately or well aligned to the other dimension are categorized as being "moderately aligned." Schools that were poorly aligned to one or both dimensions are categorized as being "poorly aligned."

The group means and impact estimates for the "poorly aligned"category are not reported in the table because only one school was deemed to have poorly aligned implementation fidelity at the second-year spring site visit. In particular, one cannot estimate the impact of the program in "poorly aligned"category schools based on only one school. These results are also suppressed in order to prevent the identification of this particular school and its impact.

The Enhanced Reading Opportunities Study

Appendix Table I.7

Impacts on Reading Behaviors, Cohort 2 Follow-Up Respondent Sample, by Program Implementation Fidelity at Spring Site Visit

Outcome	ERO	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value Estimated Impact
Very well-aligned implementation[a]					
Amount of school-related reading (prior month occurrences)	48.0	47.2	0.8	0.02	0.761
Amount of non-school-related reading (prior month occurrences)	31.5	31.1	0.4	0.01	0.864
Use of reflective reading strategies (4-point scale)	2.8	2.8	0.0	0.05	0.463
Sample size	506	368			
Well-aligned implementationa					
Amount of school-related reading (prior month occurrences)	43.2	41.4	1.8	0.05	0.494
Amount of non-school-related reading (prior month occurrences)	27.0	24.5	2.5	0.08	0.273
Use of reflective reading strategies (4-point scale)	2.7	2.6	0.1	0.10	0.187
Sample size	429	293			
Moderately aligned implementation					
Amount of school-related reading (prior month occurrences)	48.4	50.5	-2.0	-0.05	0.526
Amount of non-school-related reading (prior month occurrences)	32.2	34.1	-1.9	-0.06	0.525
Use of reflective reading strategies (4-point scale)	2.8	2.7	0.1	0.13	0.134
Sample size	304	232			
Poorly aligned implementation					
Amount of school-related reading (prior month occurrences)	–	–	–	–	–
Amount of non-school-related reading (prior month occurrences)	–	–	–	–	–
Use of reflective reading strategies (4-point scale)	–	–	–	–	–
Sample size	–	–			

(continued)

Appendix Table I.7 (continued)

Difference in Impacts Between Subgroups	Difference in Impacts	Difference in Impact Effect Sizes	P-Value for Difference
Very well aligned minus moderately aligned			
Amount of school-related reading	2.8	0.07	0.493
Amount of non-school-related reading	2.3	0.07	0.542
Use of reflective reading strategies	-0.1	-0.08	0.427
Very well aligned minus well aligned			
Amount of school-related reading	-1.1	-0.03	0.776
Amount of non-school-related reading	-2.1	-0.06	0.514
Use of reflective reading strategies	0.0	-0.06	0.566

SOURCE: MDRC calculations from the Enhanced Reading Opportunities follow-up student survey.

NOTES: The student follow-up survey was administered in spring 2007 at the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (school-related reading standard deviation = 38.322; non-school-related reading standard deviation = 32.976; use of reading strategies standard deviation = 0.592).

A two-tailed t-test was applied to the impact estimate and to the difference in impacts. Statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

For each of the above measures, data are missing for no more than 2.2 percent of the respondents.

Rounding may cause slight discrepancies in calculating sums and differences.

[a]The fidelity of program implementation is measured on two dimensions: learning environment and comprehension instruction. On each dimension, schools were measured in terms of their depth of alignment to the program model. Schools that were very well aligned on both dimensions (i.e., with average scores greater or equal to 2.5 on both dimensions) are categorized as having "very well-aligned implementation." Schools that were well aligned to both dimensions are categorized as being "well aligned." Schools that were moderately aligned to at least one dimension and moderately or well aligned to the other dimension are categorized as being "moderately aligned." Schools that are poorly aligned to one or both dimensions are categorized as being "poorly aligned."

The group means and impact estimates for the "poorly aligned" category are not reported in the table because only one school was deemed to have poorly aligned implementation fidelity at the second-year spring site visit. In particular, one cannot estimate the impact of the program in "poorly aligned" schools based on only one school. These results are also suppressed in order to prevent the identification of this particular school and its impact.

Appendix Table I.6 shows that the group of schools whose ERO program implementation was very well aligned produced a statistically significant impact on reading comprehension test scores (effect size = 0.13 standard deviation; p-value = 0.047). However, since the difference in impacts across the subgroups of schools defined by implementation fidelity is not statistically significant, it cannot be concluded that the ERO programs were more effective in schools where implementation was rated as very well aligned. Appendix Table I.7 shows that the estimated impact of the programs on reading behaviors is not statistically significant for any of the subgroups of schools defined by implementation fidelity.

Relationship Between Impacts and Number of Weeks to Program Start-Up

Appendix Tables I.8 and I.9 correspond with the third panel of Table 5.6. These tables present impacts on reading test scores and reading behaviors, consecutively, for two groups of schools defined by the number of weeks to program start-up in the second year: those that began operating their ERO program within 2 weeks of the start of the school year and those whose program start-up was delayed by 2 weeks or more.

Appendix Table I.8 shows a statistically significant impact on reading comprehension scores for schools that were able to start operating their program earlier (effect size = 0.10 standard deviation; p-value = 0.048). The difference in impacts across the two subgroups of sites, however, is not statistically significant, and thus it cannot be concluded that the ERO programs were more effective in schools with earlier program start-up. Appendix Table I.9 shows that the impact of the programs on reading behaviors is not statistically significant for either of the two subgroups defined by the number of weeks to program start-up.

To further test the impact of program fidelity and weeks to program start-up on reading comprehension, a composite qualifying statistical test for multiple hypothesis testing was conducted.[2] This test indicates that the overall variation in impacts across the subgroups defined by implementation fidelity and program start-up is not statistically significant (F-statistic = 0.80; p-value = 0.526), suggesting that the statistical significance of reading comprehension impacts for some of the subgroups in Appendix Tables I.6 through I.9 should be interpreted with caution.[3]

[2]This test was conducted by adding a set of interactions between the treatment indicator and the subgroup indicators (three subgroup indicators for implementation fidelity and one indicator for program start-up) to the impact regression model fit to the Cohort 2 follow-up respondent sample. An F-test was then used to test for the joint significance of these interaction terms.

[3]This overall F-test was also conducted for reading vocabulary, the three reading behaviors outcomes, and the reading behavior composite index. These tests indicate that the overall variation in impacts on these reading outcomes is not statistically significant.

The Enhanced Reading Opportunities Study

Appendix Table I.8

Impacts on Reading Achievement, Cohort 2 Follow-Up Respondent Sample, by Number of Weeks Between School Start and ERO Program Start

Outcome	ERO Group	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
Less than 2 weeks					
Reading comprehension					
Average standard score	90.4	89.4	1.0 *	0.10 *	0.048
Corresponding grade equivalent	*6.2*	*6.0*			
Corresponding percentile	*25*	*24*			
Reading vocabulary					
Average standard score	93.4	93.7	-0.3	-0.03	0.492
Corresponding grade equivalent	*7.7*	*7.8*			
Corresponding percentile	*32*	*32*			
Sample size	881	631			
2 weeks or more					
Reading comprehension					
Average standard score	89.5	89.1	0.4	0.04	0.546
Corresponding grade equivalent	*6.0*	*5.9*			
Corresponding percentile	*24*	*23*			
Reading vocabulary					
Average standard score	93.7	92.9	0.7	0.08	0.324
Corresponding grade equivalent	*7.8*	*7.7*			
Corresponding percentile	*32*	*31*			
Sample size	383	276			

Difference in Impacts Between Subgroups	Difference in Impacts	Difference in Impact Effect Sizes	P-Value for Difference
Difference in impacts			
Reading comprehension standard score	0.5	0.05	0.534
Reading vocabulary standard score	-1.1	-0.11	0.229

(continued)

Appendix Table I.8 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study follow-up GRADE assessment.

NOTES: The follow-up GRADE assessment was administered in the spring of 2007 near the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form B). No statistical tests or arithmetic operations were performed on these reference points.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (reading comprehension = 10.035; reading vocabulary = 9.827).

A two-tailed t-test was applied to the impact estimate and to the difference in impacts. Statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

The Enhanced Reading Opportunities Study

Appendix Table I.9

Impacts on Reading Behaviors, Cohort 2 Follow-Up Respondent Sample, by Number of Weeks Between School Start and ERO Program Start

Outcome	ERO	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
Less than 2 weeks					
Amount of school-related reading (prior month occurrences)	47.2	47.1	0.1	0.00	0.952
Amount of non-school-related reading (prior month occurrences)	30.6	30.4	0.3	0.01	0.880
Use of reflective reading strategies (4-point scale)	2.8	2.8	0.0	0.08	0.090
Sample size	877	626			
2 weeks or more					
Amount of school-related reading (prior month occurrences)	44.2	43.1	1.1	0.03	0.683
Amount of non-school-related reading (prior month occurrences)	28.2	27.1	1.0	0.03	0.679
Use of reflective reading strategies (4-point scale)	2.7	2.6	0.1	0.10	0.217
Sample size	383	275			

Difference in Impacts Between Subgroups	Difference in Impacts	Difference in Impact Effect Sizes	P-Value for Difference
Difference in impacts			
Amount of school-related reading	-1.0	-0.03	0.764
Amount of non-school-related reading	-0.8	-0.02	0.800
Use of reflective reading strategies	0.0	-0.02	0.873

(continued)

Appendix Table I.9 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities follow-up student survey.

NOTES: The student follow-up survey was administered in spring 2007 at the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (school-related reading standard deviation = 38.322; non-school-related reading standard deviation = 32.976; use of reflective reading strategies standard deviation = 0.592).

A two-tailed t-test was applied to the impact estimate and to the difference in impacts. Statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

For each of the above measures, data are missing for no more than 1.4 percent of the respondents.

Rounding may cause slight discrepancies in calculating sums and differences.

Relationship Between Impacts and Strong Overall Implementation

Appendix Tables I.10 and I.11 correspond with the final panel in Table 5.6. These tables present the impact estimates by the overall implementation strength of a school's ERO program in the second year of the study. Impacts are presented for the 12 schools whose ERO program was especially strongly implemented as defined by (1) implementation fidelity that was very well aligned to developers' specifications and (2) the program starting up within two weeks of the start of the school year. Impacts are also presented for the 22 schools that did not meet these two conditions.

Table I.10 shows that while ERO programs that were strongly implemented produced a larger impact on reading comprehension than programs that were more weakly implemented (effect size = 0.13 standard deviation and 0.05 standard deviation, respectively), neither of these impacts is statistically significant, nor is the difference between them. Also, while Table I.11 indicates that programs characterized by weaker implementation produced a statistically significant impact on the use of reading strategies (effect size = 0.12 standard deviation; p-value = 0.029), the difference in impacts between the two subgroups of schools is not statistically significant. The composite qualifying test presented in Appendix E indicates that programs that were more weakly implemented did not produce a statistically significant impact on the composite measure of reading behaviors (see Appendix Table E.3). Thus, it cannot be concluded that ERO program effectiveness varied for groups of schools categorized by overall strength of implementation.

The Enhanced Reading Opportunities Study

Appendix Table I.10

Impacts on Reading Achievement, Cohort 2 Follow-Up Respondent Sample, by Second-Year Implementation Strength

Outcome	ERO	Non-ERO Group	Estimated Impact	Estimated Impact Effect Size	P-Value for Estimated Impact
Very well-aligned implementation fidelity and less than 2 weeks to program startup[a]					
Reading comprehension					
Average standard score	90.2	88.9	1.3	0.13	0.062
Corresponding grade equivalent	*6.1*	*5.9*			
Corresponding percentile	*25*	*23*			
Reading vocabulary					
Average standard score	93.2	93.4	-0.2	-0.02	0.729
Corresponding grade equivalent	*7.7*	*7.7*			
Corresponding percentile	*31*	*32*			
Sample size	464	337			
Weaker implementation fidelity and/or 2 weeks or more to program startup[b]					
Reading comprehension					
Average standard score	90.1	89.5	0.5	0.05	0.296
Corresponding grade equivalent	*6.1*	*6.0*			
Corresponding percentile	*25*	*24*			
Reading vocabulary					
Average standard score	93.6	93.5	0.1	0.01	0.836
Corresponding grade equivalent	*7.8*	*7.8*			
Corresponding percentile	*32*	*32*			
Sample size	800	570			

Difference in Impacts Between Subgroups	Difference in Impacts	Difference in Impact Effect Sizes	P-Value for Difference
Difference in impacts			
Reading comprehension standard score	0.8	0.08	0.378
Reading vocabulary standard score	-0.3	-0.03	0.689

(continued)

Appendix Table I.10 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities Study follow-up GRADE assessment.

NOTES: The follow-up GRADE assessment was administered in the spring of 2007 near the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The national average for standard score values is 100, and its standard deviation is 15. The grade equivalent and percentile are those associated with the average standard score as indicated in the GRADE *Teacher's Scoring and Interpretive Manual* (Level H, Grade 9, Spring Testing, Form B). No statistical tests or arithmetic operations were performed on these reference points.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (reading comprehension = 10.035; reading vocabulary = 9.827).

A two-tailed t-test was applied to the impact estimate and to the difference in impacts. Statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

[a]The ERO programs in these schools were deemed to have reached an implementation level that was very well aligned to both the classroom learning environment and comprehension instruction dimensions of the program model by the spring site visit, and their ERO program began operating within 2 weeks of the start of the school year.

[b]The implementation fidelity of the ERO programs in these schools was deemed to be less than very well aligned to the classroom learning environment and/or comprehension instruction dimensions of the program model by the spring site visit, and/or their ERO program began operating 2 weeks or more after the start of the school year.

The Enhanced Reading Opportunities Study

Appendix Table I.11

Impacts on Reading Behaviors, Cohort 2 Follow-Up Respondent Sample, by Second-Year Implementation Strength

Outcome	ERO	Non-ERO Group	Estimated Impact	Estimated Impact Effect	P-Value for Estimated Impact
Very well-aligned implementation fidelity and less than 2 weeks to program startup[a]					
Amount of school-related reading (prior month occurrences)	48.1	46.5	1.6	0.04	0.554
Amount of non-school-related reading (prior month occurrences)	30.7	30.5	0.3	0.01	0.911
Use of reflective reading strategies (4-point scale)	2.8	2.8	0.0	0.04	0.587
Sample size	464	335			
Weaker implementation fidelity and/or 2 weeks or more to program startup[b]					
Amount of school-related reading (prior month occurrences)	45.3	45.5	-0.3	-0.01	0.883
Amount of non-school-related reading (prior month occurrences)	29.4	28.8	0.6	0.02	0.751
Use of reflective reading strategies (4-point scale)	2.7	2.7	0.1 *	0.12 *	0.029
Sample size	796	566			

Difference in Impacts Between Subgroups	Difference in Impacts	Difference in Impact Effect Sizes	P-Value for Difference
Differences in impact			
Amount of school-related reading	1.9	0.05	0.572
Amount of non-school-related reading	-0.3	-0.01	0.919
Use of reflective reading strategies	0.0	-0.08	0.325

(continued)

Appendix Table I.11 (continued)

SOURCE: MDRC calculations from the Enhanced Reading Opportunities follow-up student survey.

NOTES: The student follow-up survey was administered in spring 2007 at the end of students' ninth-grade year.

The estimated impacts are regression-adjusted using ordinary least squares, controlling for blocking of random assignment by school and for random differences between the ERO and non-ERO groups in their baseline reading comprehension test scores and age at random assignment. The values in the column labeled "ERO Group" are the observed means for students randomly assigned to the ERO group. The "Non-ERO Group" values in the next column are the regression-adjusted means for students randomly assigned to the non-ERO group, using the observed mean covariate values for the ERO group as the basis for the adjustment.

The estimated impact effect size is calculated as a proportion of the standard deviation of the non-ERO group average (school-related reading standard deviation = 38.322; non-school-related reading standard deviation = 32.976; use of reflective reading strategies standard deviation = 0.592).

A two-tailed t-test was applied to the impact estimate and to the difference in impacts. Statistical significance is indicated (*) when the p-value is less than or equal to 5 percent.

For each of the above measures, data are missing for no more than 2.1 percent of the respondents.

Rounding may cause slight discrepancies in calculating sums and differences.

[a]The ERO programs in these schools were deemed to have reached an implementation level that was very well aligned to both the classroom learning environment and comprehension instruction dimensions of the program model by the spring site visit, and their ERO program began operating within 2 weeks of the start of the school year.

[b]The implementation fidelity of the ERO programs in these schools was deemed to be less than very well aligned to the classroom learning environment and/or comprehension instruction dimensions of the program model by the spring site visit, and/or their ERO program began operating 2 weeks or more after the start of the school year.

References

American Guidance Service. 2001a. *Group Reading Assessment and Diagnostic Evaluation: Teacher's Scoring and Interpretive Manual, Level H.* Circle Pines, MN: American Guidance Service.

American Guidance Service. 2001b. *Group Reading Assessment and Diagnostic Evaluation: Technical Manual*. Circle Pines, MN: American Guidance Service.

American Institutes for Research. 2004. "Request for Proposals from Vendors of High Quality Supplemental Literacy Programs for Striving Readers in Ninth Grade." Available from the author upon request at ERO@air.org.

Balfanz, Robert, and Nettie Legters 2004. *Locating the Dropout Crisis.* Baltimore, MD: Center for Social Organization of Schools.
Web site: http://www.csos.jhu.edu/tdhs/rsch/Locating_Dropouts.pdf.

Biancarosa, Gina, and Catherine E. Snow. 2004. *Reading Next — A Vision for Action and Research in Middle and High School Literacy: A Report to Carnegie Corporation of New York.* Washington, DC: Alliance for Excellent Education.
Web site: http://www.all4ed.org/publications/ReadingNext/ReadingNext.pdf

Bloom, Harold, Carolyn Hill, Alison Rebeck Black, and Mark Lipsey. 2006. "Effect Sizes in Education Research: What They Are, What They Mean, and Why They Are Important." Presentation for the Institute of Education Sciences 2006 Research Conference. New York: MDRC.

Cronbach, Lee J. 1951. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika* 16: 297-334.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. *Using Randomization in Development Economics: A Toolkit.* London: Centre for Economic Policy Research.
Web site: http://www.cepr.org/pubs/dps/DP6059.asp.

Fine, Michelle. 1988. "Deinstitutionalizing Educational Inequity." Pages 88-119 in Council of Chief State School Officers (eds.), *School Success for Students at Risk.* New York: Harcourt.

Hedges, Larry V., and Ingram Olkin. 1985. *Statistical Methods for Meta-Analysis.* San Diego, CA: Academic Press.

Diana Oxley (ed.). *Critical Issues in Development and Implementation of High School Small Learning Communities.* Washington, DC: U.S. Department of Education.

Kemple, James J., William Corrin, Elizabeth Nelson, Terry Salinger, Suzannah Herrmann, and Kathryn Drummond. 2008. *The Enhanced Reading Opportunities Study: Early Impacts and Implementation Findings* (NCEE 2008-4015). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Kling, Jeffrey R., and Jeffrey B. Liebman, 2004. "Experimental Analysis of Neighborhood Effects on Youth." KSG Working Paper RWP04-034. Cambridge, MA: Harvard University, John F. Kennedy School of Government.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica,* 75, 1: 83-119.

Lipsey, Mark. 1990. *Design Sensitivity: Statistical Power for Experimental Research.* Newbury Park, CA: Sage.

Logan, Brent R., and Ajit C. Tamhane. 2003. "On O'Brien's OLS and GLS Tests for Multiple Endpoints." Department of Industrial Engineering and Management Sciences Working Paper 03-0004. Evanston, IL: Northwestern University.

National Center for Education Statistics. 1990. *National Education Longitudinal Study of 1998: A Profile of the American Eighth Grader: NELS: -88 Student Descriptive Summary.* Washington, DC: U.S. Department of Education. Web site: http://nces.ed.gov/pubs90/90458.pdf.

O'Brien, Peter C. 1984. "Procedures for Comparing Samples with Multiple Endpoints." *Biometrics* 40: 1079-1087.

Roderick, Melissa. 1993. *The Path to Dropping Out: Evidence for Intervention.* Westport, CT: Auburn House.

Schoenbach, Ruth, Cynthia Greenleaf, Christine Cziko, and Lori Hurwitz. 1999. *Reading for Understanding: A Guide to Improving Reading in Middle and High School Classrooms.* San Francisco: Jossey-Bass.

Schumaker, Jean B., and Donald D. Deshler. 2003. "Designs for Applied Educational Research." Pages 283-500 in H. Lee Swanson, Karen. R. Harris, and Steve Graham (eds.), *Handbook of Learning Disabilities.* New York: Guilford.

Schumaker, Jean B., and Donald D. Deshler. 2004. "Teaching Adolescents to be Strategic Learners." In Donald D. Deshler and Jean B. Schumaker (eds.), *High School Students with Disabilities: Strategies for Accessing the Curriculum.* New York: Corwin.

U.S. Department of Education. 2005. *Smaller Learning Communities: Special Competition for Supplemental Reading Program Research Evaluation.* Washington DC: U.S. Department of Education.

WestEd. 2004. "High Quality Supplemental Literacy Programs for Striving Readers in Ninth Grade." Unpublished proposal. San Francisco: WestEd.

Wigfield, Allen. 2004. "Motivation for Reading During the Early Adolescent and Adolescent Years." Pages 56-69 in Dorothy Strickland and Donna Alvermann (eds.), *Bridging the Literacy Achievement Gap, Grades 4-12.* New York: Teachers College Press.