U. S. DEPARTMENT OF EDUCATION

National Assessment of Title I:
Interim Report

Volume II: Closing the Reading Gap

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

National Assessment of Title I Interim Report

Volume II: Closing the Reading Gap: First Year Findings from a Randomized Trial of Four Reading Interventions for Striving Readers

A Report Prepared for IES
by the Corporation for the Advancement of Policy Evaluation

Joseph Torgesen, Florida Center for Reading Research

David Myers, Allen Schirm, Elizabeth Stuart, Sonya Vartivarian, & Wendy Mansfield
Mathematica Policy Research

Fran Stancavage, American Institutes for Research

Donna Durno and Rosanne Javorsky
Allegheny Intermediate Unit

Cinthia Haan
Haan Foundation

Institute of Education Sciences
National Center for Education Evaluation and Regional Assistance NCEE 2006-4002
U.S. Department of Education February 2006

U. S. Department of Education
Margaret Spellings
Secretary

Institute of Education Sciences
Grover J. Whitehurst
Director

National Center for Education Evaluation and Regional Assistance
Phoebe Cottingham
Commissioner

February 2006

To order copies of this report,

- Write to ED Pubs, Education Publications Center, U.S. Department of Education, P.O. Box 1398, Jessup, MD 20794-1398.

- Call in your request toll free to 1-877-4ED-Pubs. If 877 service is not yet available in your area, call 800-872-5327 (800-USA-LEARN). Those who use a telecommunications device for the deaf (TDD) or a teletypewriter (TTY) should call 800-437-0833.

- Fax your request to 301-470-1244.

- Order online at www.edpubs.org.

This report is also available on the Department's Web site at http://www.ed.gov/ies.

Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-9895 or 202-205-8113.

ACKNOWLEDGMENTS

This report reflects the contributions of many institutions and individuals. We would like to first thank the study funders. The Institute of Education Sciences of the U.S. Department of Education and the Smith Richardson Foundation funded the evaluation component of the study. Funders of the interventions included the Heinz Endowments, the W.K. Kellogg Foundation, the Grable Foundation, the Institute of Education Sciences, the Ambrose Monell Foundation, Barksdale Reading Institute, the Haan Foundation for Children, the Richard King Mellon Foundation, the Raymond Foundation, and the Rockefeller Foundation. We also thank the Rockefeller Brothers Fund for the opportunity to hold a meeting of the Scientific Advisory Panel and research team at their facilities in 2004.

We gratefully acknowledge Audrey Pendleton of the Institute of Education Sciences for her support and encouragement throughout the study. Many individuals at Mathematica Policy Research contributed to the writing of this report. In particular, Mark Dynarski provided critical comments and review of the report. Micki Morris and Daryl Hall were instrumental in editing and producing the document, with assistance from Donna Dorsey and Alfreda Holmes.

Important contributions to the study were received from several others. At Mathematica, Nancy Carey, Valerie Williams, Jessica Taylor, Season Bedell-Boyle, and Shelby Pollack assisted with data collection, and Mahesh Sundaram managed the programming effort. At the Allegheny Intermediate Unit (AIU), Jessica Lapinski served as the liaison between the evaluators and AIU school staff. At AIR, Marian Eaton and Mary Holte made major contributions to the design and execution of the implementation study, while Terry Salinger, Sousan Arafeh, and Sarah Shain made additional contributions to the video analysis. Paul William and Charles Blankenship were responsible for the programming effort, while Freya Makris and Sandra Smith helped to manage and compile the data. We also thank Anne Stretch, a reading specialist and independent consultant, for leading the training on test administration.

Finally, we would particularly like to acknowledge the assistance and cooperation of the teachers and principals in the Allegheny Intermediate Unit, without whom this study would not have been possible.

CONTENTS

CONTENTS *(continued)*

Chapter **Page**

EXECUTIVE SUMMARY

EVALUATION CONTEXT

According to the National Assessment of Educational Progress (U.S. Department of Education 2003), nearly 4 in 10 fourth graders read below the basic level. Unfortunately, these literacy problems get worse as students advance through school and are exposed to progressively more complex concepts and courses. Historically, nearly three-quarters of these students never attain average levels of reading skill. While schools are often able to provide some literacy intervention, many lack the resources—teachers skilled in literacy development and appropriate learning materials—to help older students in elementary school reach grade level standards in reading.

The consequences of this problem are life changing. Young people entering high school in the bottom quartile of achievement are substantially more likely than students in the top quartile to drop out of school, setting in motion a host of negative social and economic outcomes for students and their families.

For their part, the nation's 16,000 school districts are spending hundreds of millions of dollars on often untested educational products and services developed by textbook publishers, commercial providers, and nonprofit organizations. Yet we know little about the effectiveness of these interventions. Which ones work best, and for whom? Under what conditions are they most effective? Do these programs have the potential to close the reading gap?

To help answer these questions, we initiated an evaluation of either parts or all of four widely used programs for elementary school students with reading problems. The programs are Corrective Reading, Failure Free Reading, Spell Read P.A.T., and Wilson Reading, all of which are expected to be more intensive and skillfully delivered than the programs typically provided in public schools.[1] The programs incorporate explicit and systematic instruction in the basic reading skills in which struggling readers are frequently deficient. Corrective Reading, Spell Read P.A.T., and Wilson Reading were implemented to provide word-level instruction, whereas Failure Free Reading focused on building reading comprehension and vocabulary in addition to word-level skills. Recent reports from small-scale research and clinical studies provide some evidence that the reading skills of students with severe reading difficulties in late elementary school can be substantially improved by providing, for a sustained period of time, the kinds of skillful, systematic, and explicit instruction that these programs offer (Torgesen 2005).

EVALUATION PURPOSE AND DESIGN

Conducted just outside Pittsburgh, Pennsylvania, in the Allegheny Intermediate Unit (AIU), the evaluation is intended to explore the extent to which the four reading programs can affect both the word-level reading skills (phonemic decoding, fluency, accuracy) and reading comprehension of students in grades three and five who were identified as struggling readers by their teachers and by low test scores. Ultimately, it will provide educators with rigorous evidence of what could happen in terms of reading

[1] These four interventions were selected from more than a dozen potential program providers by members of the Scientific Advisory Board of the Haan Foundation for Children. See Appendix Q for a list of the Scientific Advisory Board members.

improvement if intensive, small-group reading programs like the ones in this study were introduced in many schools.

This study is a large-scale, longitudinal evaluation comprising two main elements. The first element of the evaluation is an impact study of the four interventions. This evaluation report is addressing three broad types of questions related to intervention impacts:

- What is the impact of being in any of the four remedial reading interventions, considered as a group, relative to the instruction provided by the schools? What is the impact of being in one of the remedial reading programs that focuses primarily on developing word-level skills, considered as a group, relative to the instruction provided by the schools? What is the impact of being in each of the four particular remedial reading interventions, considered individually, relative to the instruction provided by the schools?

- Do the impacts of programs vary across students with different baseline characteristics?

- To what extent can the instruction provided in this study close the reading gap and bring struggling readers within the normal range, relative to the instruction provided by their schools?

To answer these questions, the impact study was based on a scientifically rigorous design—an experimental design that uses random assignment at two levels: (1) 50 schools from 27 school districts were randomly assigned to one of the four interventions, and (2) within each school, eligible children in grades 3 and 5 were randomly assigned to a treatment group or to a control group. Students assigned to the intervention group (treatment group) were placed by the program providers and local coordinators into instructional groups of three students. Students in the control groups received the same instruction in reading that they would have ordinarily received. Children were defined as eligible if they were identified by their teachers as struggling readers and if they scored at or below the 30th percentile on a word-level reading test and at or above the 5th percentile on a vocabulary test. From an original pool of 1,576 3rd and 5th grade students identified as struggling readers, 1,042 also met the test-score criteria. Of these eligible students, 772 were given permission by their parents to participate in the evaluation.

The second element of the evaluation is an implementation study that has two components: (1) an exploration of the similarities and differences in reading instruction offered in the four interventions and (2) a description of the regular instruction that students in the control group received in the absence of the interventions and the regular instruction received by the treatment group beyond the interventions.

Test data and other information on students, parents, teachers, classrooms, and schools is being collected several times over a three-year period. Key data collection points pertinent to this summary report include the period just before the interventions began, when baseline information was collected, and the period immediately after the interventions ended, when follow-up data were collected. Additional follow-up data for students and teachers are being collected in 2005 and again in 2006.

THE INTERVENTIONS

We did not design new instructional programs for this evaluation. Rather, we employed either parts or all of four existing and widely used remedial reading instructional programs: Spell Read P.A.T., Corrective Reading, Wilson Reading, and Failure Free Reading.

As the evaluation was originally conceived, the four interventions would fall into two instructional classifications with two interventions in each. The interventions in one classification would focus only on word-level skills, and the interventions in the other classification would focus equally on word-level skills and reading comprehension/vocabulary.

Corrective Reading and Wilson Reading were modified to fit within the first of these classifications. The decision to modify these two intact programs was justified both because it created two treatment classes that were aligned with the different types of reading deficits observed in struggling readers and because it gave us sufficient statistical power to contrast the relative effectiveness of the two classes. Because Corrective Reading and Wilson Reading were modified, results from this study do not provide complete evaluations of these interventions; instead, the results suggest how interventions using primarily the word-level components of these programs will affect reading achievement.

With Corrective Reading and Wilson Reading focusing on word-level skills, it was expected that Spell Read P.A.T. and Failure Free Reading would focus on both word-level skills and reading comprehension/vocabulary. In a time-by-activity analysis of the instruction that was actually delivered, however, it was determined that three of the programs—Spell Read P.A.T., Corrective Reading, and Wilson Reading—focused primarily on the development of word-level skills), and one—Failure Free Reading—provided instruction in both word-level skills and the development of comprehension skills and vocabulary.

- ***Spell Read Phonological Auditory Training (P.A.T.)*** provides systematic and explicit fluency-oriented instruction in phonemic awareness and phonics along with every-day experiences in reading and writing for meaning. The phonemic activities include a wide variety of specific tasks focused on specific skill mastery and include, for example, building syllables from single sounds, blending consonant and vowel sounds, and analyzing or breaking syllables into their individual sounds. Each lesson also includes reading and writing activities intended to help students apply their phonically based reading skills to authentic reading and writing tasks. The Spell Read intervention had originally been one of the two "word-level plus comprehension" interventions, but after the time x activity analysis, we determined that it was more appropriately grouped as a "word-level" intervention.

- ***Corrective Reading*** uses scripted lessons that are designed to improve the efficiency of instruction and to maximize opportunities for students to respond and receive feedback. The lessons involve very explicit and systematic instructional sequences, including a series of quick tasks that are intended to focus students' attention on critical elements for successful word identification as well as exercises intended to build rate and fluency through oral reading of stories that have been constructed to counter word-guessing habits. Although the Corrective Reading program does have instructional procedures that focus on comprehension, they were originally designated as a "word-level intervention," and the developer was asked not to include these elements in this study.

- ***Wilson Reading*** uses direct, multi-sensory, structured teaching based on the Orton-Gillingham methodology. The program is based on 10 principles of instruction, some of which involve teaching fluent identification of letter sounds; presenting the structure of language in a systematic, cumulative manner; presenting concepts in the context of controlled as well as non-controlled text; and teaching and reinforcing concepts with visual-auditory-kinesthetic-tactile methods. Similar to Corrective Reading, the Wilson Program has instructional procedures that focus on comprehension and vocabulary, but since they were originally designated as a "word-level" intervention, they were asked not to include these in this study.

- ***Failure Free Reading*** uses a combination of computer-based lessons, workbook exercises, and teacher-led instruction to teach sight vocabulary, fluency, and comprehension. The program is designed to have students spend approximately one-third of each instructional session working within each of these formats, so that they are not taught simultaneously as a group. Unlike the other three interventions in this study, Failure Free does not emphasize phonemic decoding strategies. Rather, the intervention depends upon building the student's vocabulary of "sight words" through a program involving multiple exposures and text that is engineered to support learning of new words. Students read material that is designed to be of interest to their age level while also challenging their current independent and instructional reading level. Lessons are based on story text that is controlled for syntax and semantic content.

MEASURES OF READING ABILITY

Seven measures of reading skill were administered at the beginning and end of the school year to assess student progress in learning to read. As outlined below, these measures of reading skills assessed phonemic decoding, word reading accuracy, text reading fluency, and reading comprehension.

Phonemic Decoding

- Word Attack (WA) subtest from the Woodcock Reading Mastery Test-Revised (WRMT-R)
- Phonemic Decoding Efficiency (PDE) subtest from the Test of Word Reading Efficiency (TOWRE)

Word Reading Accuracy and Fluency

- Word Identification (WI) subtest from the WRMT-R
- Sight Word Efficiency (SWE) subtest from the TOWRE
- Oral Reading Fluency subtest from Edformation, Inc. The text of this report refers to the reading passages as "Aimsweb" passages, which is the term used broadly in the reading practice community.

Reading Comprehension

- Passage Comprehension (PC) subtest from the WRMT-R
- Passage Comprehension from the Group Reading Assessment and Diagnostic Evaluation (GRADE)

For all tests except the Aimsweb passages, the analysis uses grade-normalized standard scores, which indicate where a student falls within the overall distribution of reading ability among students in the same grade. Scores above 100 indicate above-average performance; scores below 100 indicate below-average performance. In the population of students across the country at all levels of reading ability, standard scores are constructed to have a mean of 100 and a standard deviation of 15, implying that approximately 70 percent of all students' scores will fall between 85 and 115 and that approximately 95

percent of all students' scores will fall between 70 and 130. For the Aimsweb passages, the score used in this analysis is the median correct words per minute from three grade-level passages.

IMPLEMENTING THE INTERVENTIONS

The interventions were implemented from the first week of November 2003 through the first weeks in May 2004. During this time students received, on average, about 90 hours of instruction, which was delivered five days a week to groups of three students in sessions that were approximately 50 minutes long. A small part of the instruction was delivered in groups of two, or 1:1, because of absences and make-up sessions. Since many of the sessions took place during the student's regular classroom reading instruction, teachers reported that students in the treatment groups received less reading instruction in the classroom than did students in the control group (1.2 hours per week versus 4.4 hours per week.). Students in the treatment group received more small-group instruction than did students in the control group (6.8 hours per week versus 3.7 hours per week). Both groups received a very small amount of 1:1 tutoring in reading from their schools during the week.

Teachers were recruited from participating schools on the basis of experience and the personal characteristics relevant to teaching struggling readers. They received, on average, nearly 70 hours of professional development and support during the implementation year as follows:

- About 30 hours during an initial week of intensive introduction to each program
- About 24 hours during a seven-week period at the beginning of the year when the teachers practiced their assigned methods with 4th-grade struggling readers in their schools
- About 14 hours of supervision during the intervention phase

According to an examination of videotaped teaching sessions by the research team, the training and supervision produced instruction that was judged to be faithful to each intervention model. The program providers themselves also rated the teachers as generally above average in both their teaching skill and fidelity to program requirements relative to other teachers with the same level of training and experience.

CHARACTERISTICS OF STUDENTS IN THE EVALUATION

The characteristics of the students in the evaluation sample are shown in Table 1 (see the end of this summary for all tables). About 45 percent of the students qualified for free or reduced-price lunches. In addition, about 27 percent were African American, and 73 percent were white. Fewer than two percent were Hispanic. Roughly 33 percent of the students had a learning disability or other disability.

On average, the students in our evaluation sample scored about one-half to one standard deviation below national norms (mean 100 and standard deviation 15) on measures used to assess their ability to decode words. For example, on the Word Attack subtest of the Woodcock Reading Mastery Test-Revised (WRMT-R), the average standard score was 93. This translates into a percentile ranking of 32. On the TOWRE test for phonemic decoding efficiency (PDE), the average standard score was 83, at approximately the 13th percentile. On the measure of word reading accuracy (Word Identification subtest for the WRMT-R), the average score placed these students at the 23rd percentile. For word reading fluency, the average score placed them at the 16th percentile for word reading efficiency

(TOWRE SWE), and third- and fifth-grade students, respectively, read 41 and 77 words per minute on the oral reading fluency passages (Aimsweb). In terms of reading comprehension, the average score for the WRMT-R test of passage comprehension placed students at the 30th percentile, and for the Group Reading and Diagnostic Assessment (GRADE), they scored, on average, at the 23rd percentile.

This sample, as a whole, was substantially less impaired in basic reading skills than most samples used in previous research with older reading disabled students. These earlier studies typically examined samples in which the phonemic decoding and word reading accuracy skills of the average student were below the tenth percentile and, in some studies, at only about the first or second percentile. Students in such samples are much more impaired and more homogeneous in their reading abilities than the students in this evaluation and in the population of all struggling readers in the United States. Thus, it is not known whether the findings from these previous studies pertain to broader groups of struggling readers in which the average student's reading abilities fall between, say, the 20th and 30th percentiles. This evaluation can help to address this issue. It obtained a broad sample of struggling readers, and is evaluating in regular school settings the kinds of intensive reading interventions that have been widely marketed by providers and widely sought by school districts to improve such students' reading skills.

DISCUSSION OF IMPACTS

This first year report assesses the impact of the four interventions on the treatment groups in comparison with the control groups immediately after the end of the reading interventions. In particular, we provide detailed estimates of the impacts, including the impact of being randomly assigned to receive any of the interventions, being randomly assigned to receive a word-level intervention, and being randomly assigned to receive each of the individual interventions. For purposes of this summary, we focus on the impact of being randomly assigned to receive any intervention compared to receiving the instruction that would normally be provided. These findings are the most robust because of the larger sample sizes. The full report also estimates impacts for various subgroups, including students with weak and strong initial word attack skills, students with low or high beginning vocabulary scores, and students who either qualified or did not qualify for free or reduced price school lunches. [2]

The impact of each of the four interventions is the difference between average treatment and control group outcomes. Because students were randomly assigned to the two groups, we would expect the groups to be statistically equivalent; thus, with a high probability, any differences in outcomes can be attributed to the interventions. Also because of random assignment, the outcomes themselves can be defined either as test scores at the end of the school year, or as the change in test scores between the beginning and end of the school year (the "gain"). In the tables of impacts (Tables 2-4), we show three types of numbers. The baseline score shows the average standard score for students at the beginning of the school year. The control gain indicates the improvement that students would have made in the absence of the interventions. Finally, the impact shows the value added by the interventions. In other words, the impact is the amount that the interventions increased students' test scores relative to the

[2] The impacts described here represent the impact of being selected to participate in one of the interventions. A small number of students selected for the interventions did not participate, and about 7.5 percent received less than a full dose (80 hours) of instruction. Estimation of the effect of an intervention on participants and those who participated for 80 or more hours requires that stronger assumptions be made than when estimating impacts for those offered the opportunity to participate, and we cannot have the same confidence in the findings as we do with the results discussed in this summary. Our full report presents estimates of the effects for participants and those who participated for at least 80 hours. These findings are similar to those reported here.

control group. The gain in the intervention group students' average test scores between the beginning and end of the school year can be calculated by adding the control group gain and the impact.

In practice, impacts were estimated using a hierarchical linear model that included a student-level model and a school-level model. In the student-level model, we include indicators for treatment status and grade level as well as the baseline test score. The baseline test score was included to increase the precision with which we measured the impact, that is, to reduce the standard error of the estimated impact. The school-level model included indicators that show the intervention to which each school was randomly assigned and indicators for the blocking strata used in the random assignment of schools to interventions. Below, we describe some of the key interim findings:

- ***For third graders, we found that the four interventions combined had impacts on phonemic decoding, word reading accuracy and fluency, and reading comprehension. There are fewer significant impacts for fifth graders than for third graders (see Table 2). The impacts of the three word-level interventions combined were similar to those for all four interventions combined.*** Although many of the impacts shown in Table 2 for third graders are positive and statistically significant when all, or just the three word-level, interventions are considered, it is noteworthy that on the GRADE, which is a group-administered test for reading comprehension, the impact estimate and the estimated change in standard scores for the control group indicate that there was not a substantial improvement in reading comprehension in the intervention groups relative to the larger normative sample for the test. Instead, this evidence suggests that the interventions helped these students maintain their relative position among all students and not lose ground in reading comprehension, as measured by the GRADE test. Results from the GRADE test are particularly important, because this test, more than others in the battery, closely mimics the kinds of testing demands (group administration, responding to multiple choice comprehension questions) found in current state-administered reading accountability measures.

- ***Among key subgroups, the most notable variability in findings were observed for students who qualified for free or reduced price lunches and those who did not.*** Although the ability to compare impacts between groups is limited by the relatively small samples, we did generally find significant impacts on the reading outcomes for third graders who did not qualify and few significant impacts for those who did qualify (see Tables 3 and 4), when all four interventions are considered together and when the three word-level interventions are considered together. These findings for third graders may be driven in part by particularly large negative gains among the control group students in the schools assigned to one intervention.

- ***At the end of the first year, the reading gap for students in the intervention group was generally smaller than the gap for students in the control group when considering all four interventions together.*** The reading gap describes the extent to which the average student in one of the two evaluation groups (intervention or control) is lagging behind the average student in the population (see Figures 1-12 and Table 5). The reduction in the reading gap attributable to the interventions at the end of the school year is measured by the interventions' impact relative to the gap for the control group, the latter showing how well students would have performed if they had not been in one of the interventions. Being in one of the interventions reduced the reading gap on Word Attack skills by about two-thirds for third graders. On other word-level tests and a measure of reading comprehension, the interventions reduced the gap for third graders by about one-fifth to one-quarter. For fifth

graders, the interventions reduced the gap for Word Attack and Sight Word Efficiency by about 60 and 12 percent, respectively.[3]

Future reports will focus on the impacts of the interventions one year after they ended. At this point, it is still too early to draw definitive conclusions about the impact of the interventions assessed in this study. Based on the results from earlier research (Torgesen et al. 2001), there is a reasonable possibility that students who substantially improved their phonemic decoding skills will continue to improve in reading comprehension relative to average readers. Consistent with the overall pattern of immediate impacts, we would expect more improvement in students who were third graders when they received the intervention relative to fifth graders. We are currently processing second-year data (which includes scores on the Pennsylvania state assessments) and expect to release a report on that analysis within the next year.

[3] In future analyses, we plan to explore another approach for estimating the impact of the interventions on closing the reading gap. This approach will contrast the percentage of students in the intervention groups and the control groups who scored within the "normal range" on the standardized tests.

Table 1

Baseline Characteristics of the Analysis Sample
3rd Grade and 5th Grade

Baseline Means	Grade Level					
	Combined		3rd		5th	
Student Characteristics						
Age	9.7		8.7		10.7	
Male (%)	54		52		56	
Hispanic (%)	2		2		1	
Race--White (%)	73		71		74	
Race--African American (%)	27		29		26	
Race--Other (%)	a		a		a	
Family income less than $30,000 (%)	50		49		50	
Family income between $30,000 and $60,000 (%)	34		33		35	
Family income over $60,000 (%)	16		18		14	
Eligible for Free or Reduced Price Lunch (%)	45		46		45	
Has any learning or other disability (%)	33		34		32	
Mother has bachelor's degree or higher (%)	12		12		12	
Reading Tests	Standard Score	Percentile	Standard Score	Percentile	Standard Score	Percentile
Screening Tests						
TOWRE Sight Word Efficiency	84.3	15	84.4	15	84.2	15
TOWRE Phonemic Decoding Efficiency	82.9	13	85.6	17	80.5	10
Peabody Picture Vocabulary Test--Revised	94.8	36	94.6	36	94.9	37
Baseline Tests						
WRM Word Identification	88.7	23	88.7	23	88.7	22
TOWRE Phonemic Decoding Efficiency	83.2	13	85.6	17	81.0	10
WRM Word Attack	92.9	32	92.6	31	93.1	32
TOWRE Sight Word Efficiency	85.3	16	86.5	18	84.2	15
AIMSWeb (Raw score)	NA	NA	40.9	NA	77.4	NA
WRM Passage Comprehension	92.3	30	91.8	29	92.7	31
GRADE	89.0	23	86.3	18	91.4	28
Woodcock Johnson Spelling	89.7	25	88.6	22	90.8	27
Woodcock Johnson Calculation	94.9	37	95.4	38	94.6	36
Other Baseline Tests Administered						
RAN Colors	89.0	23	87.7	21	90.2	26
RAN Letters	89.7	25	87.0	19	92.1	30
RAN Numbers	92.0	30	89.6	24	94.3	35
RAN Objects	88.8	23	87.7	21	89.8	25
RAS Numbers and Letters	89.3	24	87.1	19	91.4	28
RAS Colors, Numbers, and Letters	88.9	23	86.6	19	91.0	27
CTOPP Blending Words	7.5	20	7.7	22	7.3	18
CTOPP Elision	7.7	22	7.9	25	7.5	20
CTOPP Rapid Digit Naming	7.9	24	7.8	24	8.0	25
CTOPP Rapid Letter Naming	8.5	30	8.5	31	8.4	30
Clinical Evaluation of Language Fundamentals-IV	7.8	23	7.6	21	8.0	25
Sample Size	742		335		407	

Note: Weights used to account for differential randomization probabilities and nonresponse.

Note: All standard scores have mean 100 and standard deviation 15, except for CTOPP and Clinical Evaluation of Language Fundamentals-IV, which have mean 10 and standard deviation 3. Standard scores unavailable for the Aimsweb test.

Note: The percentile score shown for each test is the percentile corresponding with the mean standard score.

a Values suppressed to protect student confidentialiy.

Table 2

Impacts for 3rd and 5th Graders

		All Interventions		Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading			Corrective Reading	
Grade 3	Baseline	Control Gain	ABCD Impact	Control Gain	BCD Impact	Control Gain	A Impact	Control Gain	B Impact	Control Gain	C Impact		Control Gain	D Impact
Word Attack	92.6	0.2	5.0 *	0.0	6.8 *	0.7	-0.5	2.5	6.5 *	-3.0	8.8 *		0.5	5.2 *
TOWRE PDE	85.6	3.0	3.0 *	2.6	4.4 *	4.1	-1.3	4.1	7.1 *	0.2	5.8 *		3.6	0.4
Word Identification	88.7	-0.6	2.3 *	-0.6	2.6 *	-0.5	1.3	0.4	2.0	-2.3	2.5		0.1	3.3 *
TOWRE SWE	86.5	3.4	2.7 *	3.6	2.8 *	2.9	2.6	4.9	0.7	3.5	3.1		2.4	4.6 *
Aimsweb	40.9	20.6	4.9 *	20.3	5.9 *	21.5	1.9	22.6	1.0	17.5	6.0		20.9	10.7 *
Passage Comprehension	91.8	0.9	1.2	1.5	0.7	-0.8	2.7	2.4	0.2	-0.5	1.0		2.6	0.9
GRADE	86.2	-4.0	4.6 *	-3.1	4.4	-6.5	5.3	-4.2	4.9	-4.3	4.2		-0.9	4.2
Sample Size	335		335		242		93		92		71			79

		All Interventions		Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading			Corrective Reading	
Grade 5	Baseline	Control Gain	ABCD Impact	Control Gain	BCD Impact	Control Gain	A Impact	Control Gain	B Impact	Control Gain	C Impact		Control Gain	D Impact
Word Attack	93.1	2.2	2.7 *	2.4	3.9 *	1.3	-0.9	3.2	5.3 *	2.0	4.4 *		2.1	1.9
TOWRE PDE	81.0	5.9	1.4	6.3	1.5	4.6	1.1	7.9	4.1 *	6.8	-1.4	#	4.3	1.9
Word Identification	88.7	2.9	0.5	2.8	0.9	3.1	-0.6	2.8	0.1	2.6	2.1		3.1	0.3
TOWRE SWE	84.2	4.0	1.4 *	4.5	1.3	2.4	1.7	5.6	2.1	4.6	-0.5		3.4	2.2
Aimsweb	77.4	19.1	2.0	18.7	2.8	20.5	-0.3	19.6	3.6	19.4	-0.1		17.1	4.9
Passage Comprehension	92.7	-1.7	1.3	-2.1	1.6	-0.6	0.3	-1.2	0.6	-3.7	2.5		-1.4	1.8
GRADE	91.5	1.0	-0.2	0.8	0.3	1.6	-1.6	-0.5	-0.7	-0.7	1.3		3.6	0.3
Sample Size	407		407		281		126		104		91			86

* Impact is statistically significant at the 0.05 level.

Impact is statistically different from the 3rd grade impact at the 0.05 level.

Note: Sample sizes indicate the number of students randomly assigned to the intervention or control group, excluding students with missing test scores at the beginning or end of the school year.

Table 3

Impacts for 3rd and 5th Graders Eligible for Free or Reduced Price School Lunch

		All Interventions		Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
		Control	ABCD	Control	BCD	Control	A	Control	B	Control	C	Control	D
Grade 3	Baseline	Gain	Impact	Gain	Impact	Gain	Impact	Gain	Impact	Gain	Impact	Gain	Impact
Word Attack	92.2	1.3	4.7 *	1.6	5.9 *	0.7	1.3	1.7	8.4 *	0.2	6.0 * #	2.8	3.3
TOWRE PDE	85.3	4.6	1.8	4.5	2.6 #	4.9	-0.7	5.1	6.2 *	1.9	3.6 #	6.5	-2.0
Word Identification	88.0	0.2	1.1	0.3	1.1	-0.2	1.0	2.3	-0.6	-1.4	1.2	0.0	2.8
TOWRE SWE	85.5	3.5	1.3	4.0	0.7	2.2	3.0	4.1	-0.8	3.9	2.5	3.9	0.4 #
Aimsweb	38.6	20.3	2.0	19.6	3.1	22.5	-1.1	22.0	-1.9	16.1	6.4	20.7	4.7
Passage Comprehension	90.4	3.3	-0.8 #	4.2	-1.2 #	0.7	0.4	3.5	0.5	4.5	-2.6 #	4.5	-1.5
GRADE	84.4	-2.0	0.1 #	-0.7	-0.8 #	-6.0	2.5	-2.6	1.6	-1.4	-2.1 #	1.8	-1.7
Sample Size	193												

		All Interventions		Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
		Control	ABCD	Control	BCD	Control	A	Control	B	Control	C	Control	D
Grade 5	Baseline	Gain	Impact	Gain	Impact	Gain	Impact	Gain	Impact	Gain	Impact	Gain	Impact
Word Attack	92.5	3.5	0.6	4.1	1.5	1.7	-2.3	5.7	0.8 #	3.7	3.0	2.8	0.8
TOWRE PDE	80.1	6.5	0.6	6.6	1.0	6.2	-0.5	8.9	2.9	7.2	-1.2	3.8	1.3
Word Identification	87.8	2.4	0.4	2.5	0.7	2.2	-0.4	2.5	-1.2	2.1	3.0 *	3.0	0.2
TOWRE SWE	83.2	2.6	3.7 * #	2.9	3.8 * #	1.6	3.2	4.5	3.9 *	4.1	1.0	0.3	6.5 * #
Aimsweb	73.4	14.7	3.1	14.0	4.5	16.6	-1.1	16.0	8.6 *	13.7	0.7	12.4	4.4
Passage Comprehension	90.6	-0.1	-0.3	-0.3	-0.1	0.5	-0.9	0.6	-0.8	-0.8	-0.8	-0.8	1.3
GRADE	88.6	3.2	-4.1 * #	3.1	-3.7	3.3	-5.4	4.9	-6.1 *	1.0	-4.2	3.3	-0.8
Sample Size	230												

* Impact is statistically significant at the 0.05 level.

Impact is statistically different from the impact for all students in that grade at the 0.05 level.

Table 4

Impacts for 3rd and 5th Graders Not Eligible for Free or Reduced Price School Lunch

		All Interventions		Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 3	Baseline	Control Gain	ABCD Impact	Control Gain	BCD Impact	Control Gain	A Impact	Control Gain	B Impact	Control Gain	C Impact	Control Gain	D Impact
Word Attack	93.3	-2.7	7.8 *	-3.8	10.9 *	0.7	-1.7	0.8	8.3 *	-13.2	19.5 * #	0.9	5.0
TOWRE PDE	86.1	0.1	5.3 *	-1.2	8.0 * #	4.1	-3.1	4.8	6.2 *	-12.1	17.6 * #	3.7	0.3
Word Identification	89.9	-2.4	3.6 *	-3.1	4.6 *	-0.2	0.5	-1.1	2.4	-7.8	7.8	-0.3	3.6
TOWRE SWE	87.9	3.0	3.0 *	2.6	3.9 *	4.1	0.2	6.8	-0.5	-0.1	5.2	1.1	6.9 * #
Aimsweb	44.1	19.0	7.6 *	19.0	8.4 *	19.1	5.1	23.1	1.1	13.0	9.6	20.9	14.5 *
Passage Comprehension	93.8	-5.0	6.1 * #	-5.9	6.7 * #	-2.1	4.2	2.7	-2.8	-20.9	19.5 * #	0.5	3.6
GRADE	88.9	-8.6	9.5 * #	-8.9	10.6 * #	-7.5	6.4	-5.5	6.0	-17.9	19.2 * #	-3.4	6.6
Sample Size	142												

		All Interventions		Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 5	Baseline	Control Gain	ABCD Impact	Control Gain	BCD Impact	Control Gain	A Impact	Control Gain	B Impact	Control Gain	C Impact	Control Gain	D Impact
Word Attack	94.0	1.4	3.7 *	1.5	5.1 *	0.9	-0.5	1.3	8.9 * #	1.4	4.1	1.9	2.2
TOWRE PDE	82.0	5.3	1.2	6.1	1.0	3.0	1.6	6.3	4.8 *	6.9	-2.1	5.0	0.5
Word Identification	89.7	3.6	0.0	3.1	0.5	4.8	-1.6	2.5	0.9	3.8	0.5	3.1	0.0
TOWRE SWE	85.4	4.8	0.0 #	5.7	-0.7 #	1.9	2.0	5.3	1.1	5.0	-0.4	6.8	-2.8 #
Aimsweb	82.2	22.1	0.3	21.7	0.2	23.5	0.5	21.0	-0.7	22.0	0.0	22.0	1.4
Passage Comprehension	95.1	-2.9	2.1	-3.2	2.4	-1.9	1.4	-2.4	1.3	-6.9	5.3 *	-0.3	0.5
GRADE	94.9	0.3	1.2 #	-0.2	1.9	1.9	-0.7	-4.5	1.8	0.1	2.8	3.8	1.0
Sample Size	177												

* Impact is statistically significant at the 0.05 level.

Impact is statistically different from the impact for all students in that grade at the 0.05 level.

Table 5

Relative Gap Reduction: All Interventions Combined

3rd Grade	Average at baseline	Gap at baseline (Std. Units)	Average at follow-up		Gap at follow-up (Std. Units)		Impact	RGR
			Intervention Group	Control Group	Intervention Group	Control Group		
Word Attack	92.6	0.49	97.8	92.8	0.15	0.48	5.0 *	0.69
TOWRE PDE	85.6	0.96	91.6	88.6	0.56	0.76	3.0 *	0.26
Word Identification	88.7	0.75	90.4	88.1	0.64	0.79	2.3 *	0.19
TOWRE SWE	86.5	0.90	92.6	89.9	0.49	0.67	2.7 *	0.27
Aimsweb	NA	NA	NA	NA	NA	NA	NA	NA
Passage Comprehension	91.8	0.55	93.9	92.7	0.40	0.48	1.2	0.17
GRADE	86.2	0.92	86.9	82.3	0.87	1.18	4.6 *	0.26

5th Grade	Average at baseline	Gap at baseline (Std. Units)	Average at follow-up		Gap at follow-up (Std. Units)		Impact	RGR
			Intervention Group	Control Group	Intervention Group	Control Group		
Word Attack	93.1	0.46	98.0	95.3	0.14	0.31	2.7 *	0.56
TOWRE PDE	81.0	1.27	88.3	86.9	0.78	0.87	1.4	0.11
Word Identification	88.7	0.76	92.1	91.6	0.53	0.56	0.5	0.06
TOWRE SWE	84.2	1.05	89.6	88.2	0.69	0.78	1.4 *	0.12
Aimsweb	NA	NA	NA	NA	NA	NA	NA	NA
Passage Comprehension	92.7	0.49	92.2	90.9	0.52	0.60	1.3	0.14
GRADE	91.5	0.57	92.3	92.5	0.51	0.50	-0.2	-0.02

* Impact is statistically significant at the 0.05 level.

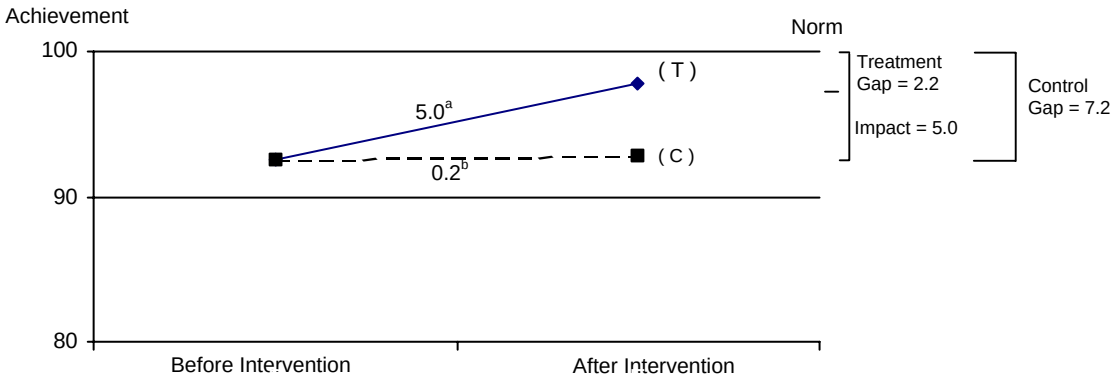
Note: RGR defined as RGR = (Impact/(100-Average for Control Group at follow-up).

Note: Gap defined as (100-Average Score)/15, where 100 is the population average and 15 is the population standard deviation.

Note: Values for Aimsweb not available because normed standard scores were unavailable.

Figure 1

Third Grade Gains in Word Attack



[a]Treatment group gain
[b]Control group gain

Figure 2

Third Grade Gains in Phonemic Decoding Efficiency

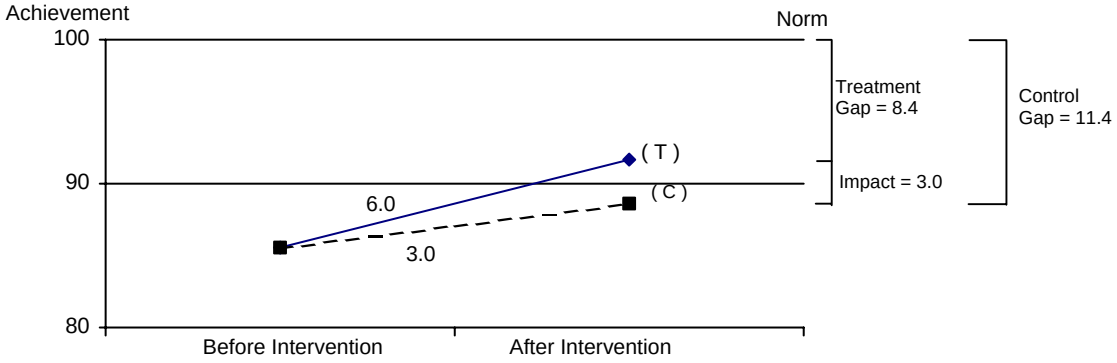


Figure 3

Third Grade Gains in Word Identification

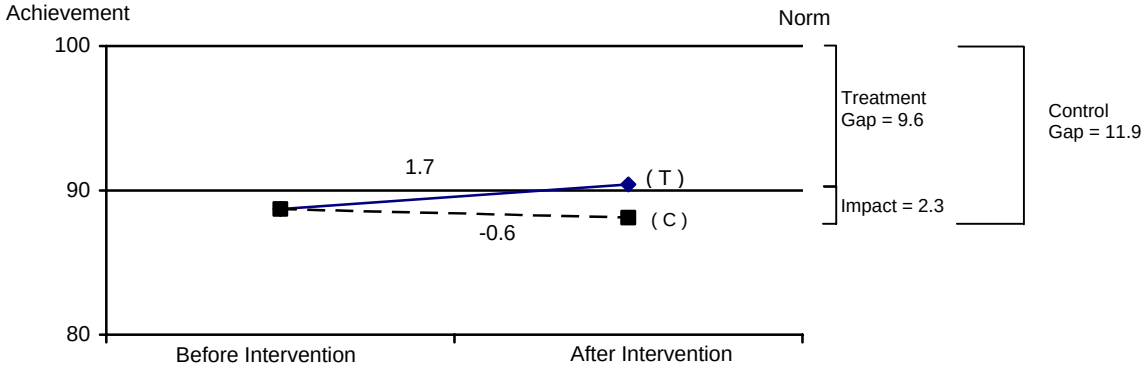


Figure 4

Third Grade Gains in Sight Word Efficiency

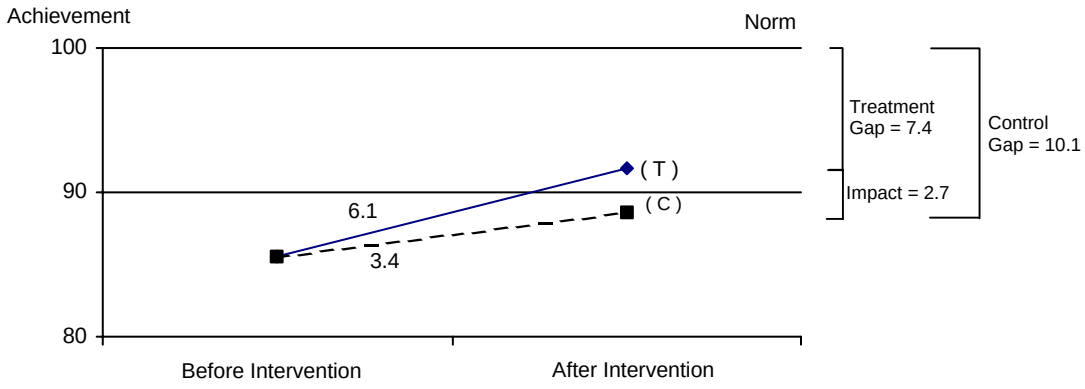


Figure 5

Third Grade Gains in Passage Comprehension

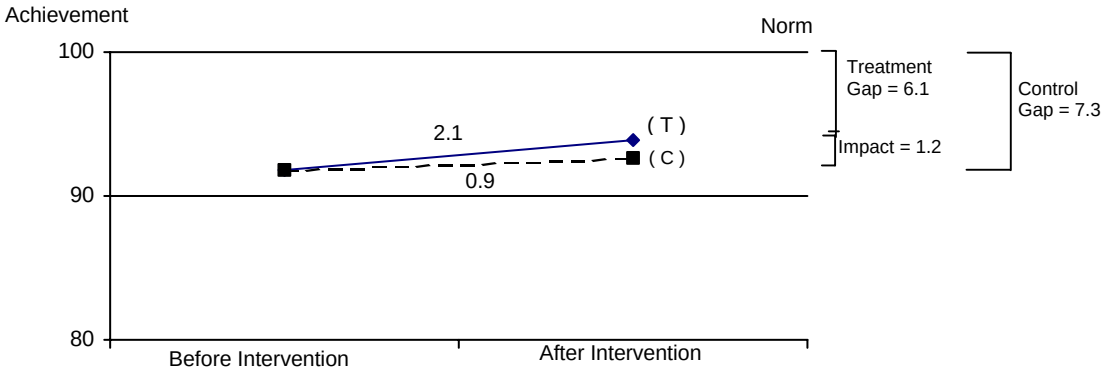


Figure 6

Third Grade Gains in GRADE Test

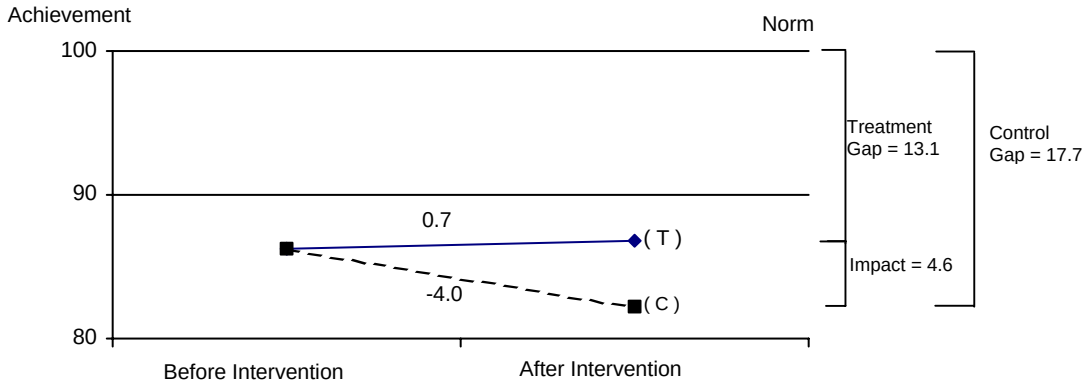


Figure 7

Fifth-Grade Gains in Word Attack

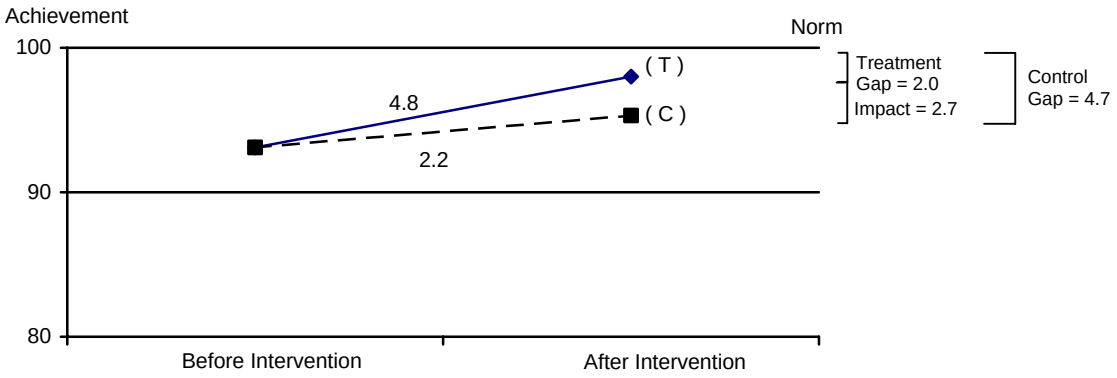


Figure 8

Fifth-Grade Gains in Phonemic Decoding Efficiency

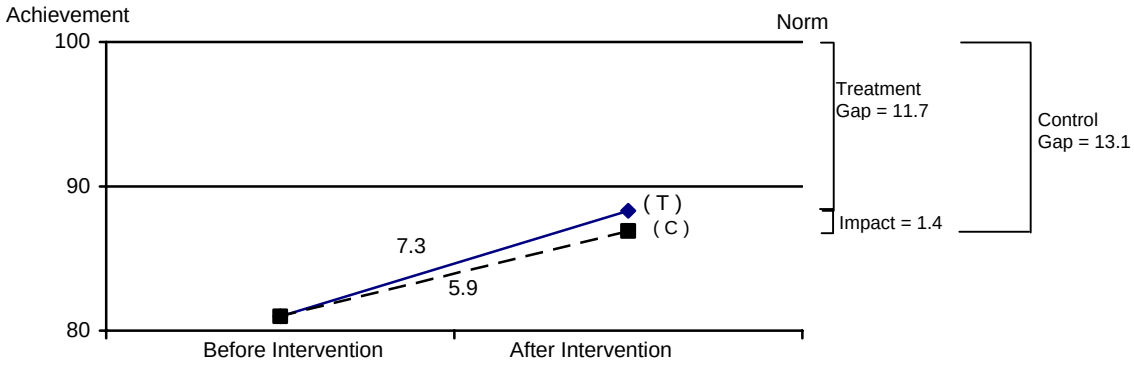


Figure 9

Fifth-Grade Gains in Word Identification

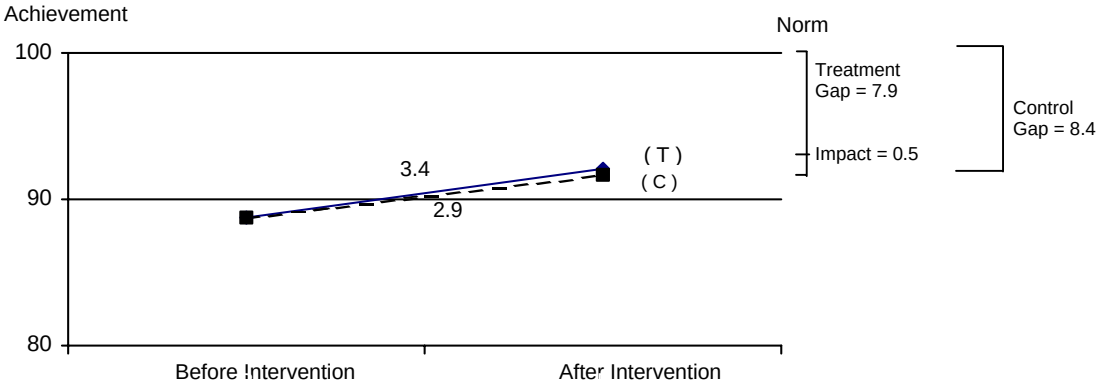


Figure 10

Fifth-Grade Gains in Sight Word Efficiency

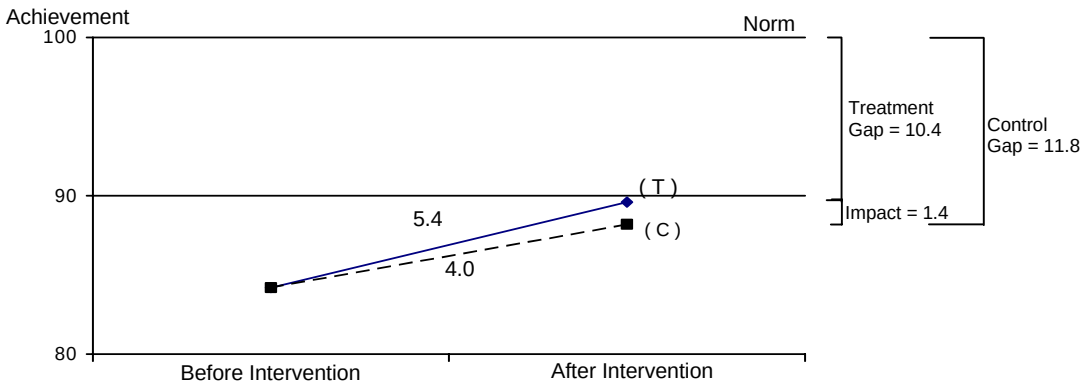


Figure 11

Fifth-Grade Gains in Passage Comprehension

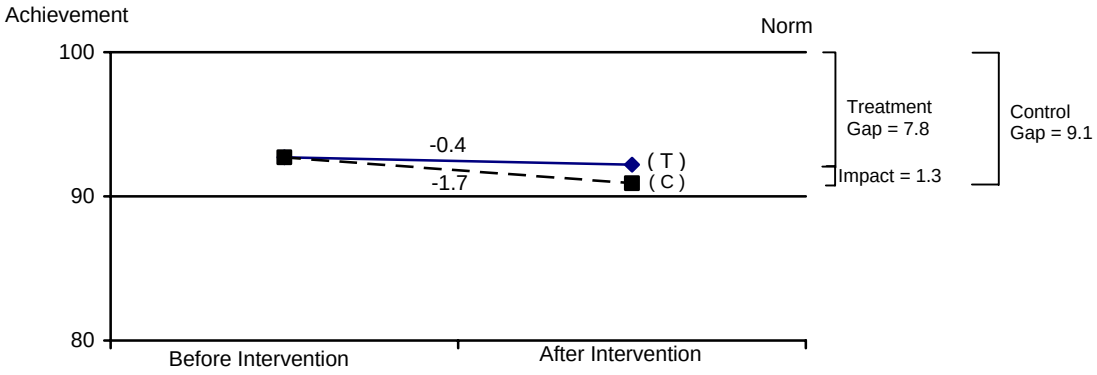
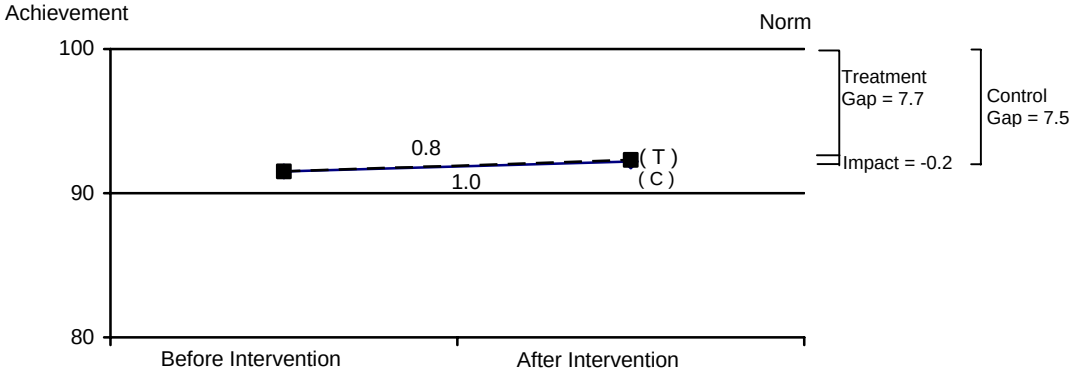


Figure 12

Fifth-Grade Gains in GRADE Test



I. INTRODUCTION

A. OVERVIEW

According to the National Assessment of Educational Progress (U.S. Department of Education 2003), nearly 4 in 10 fourth graders read below the basic level. Unfortunately, such literacy problems get worse as students advance through school and are exposed to progressively more complex concepts and courses. Historically, nearly three-quarters of these students never attain average levels of reading skill, and the consequences are life changing. Young people entering high school in the bottom quartile of achievement are substantially more likely than students in the top quartile to drop out of school, setting in motion a host of negative social and economic outcomes for students and their families.

To address this problem, many school districts have created remedial programs that aim to produce, on average, about one year's gain in reading skills for each year of instruction. However, if children begin such programs two years below grade level, they will never "close the gap" between themselves and average readers. Recent studies have found that children placed in special education after third grade typically achieve a year's gain or less in reading skill for each year in special education (McKinney 1990; Zigmond 1996). Thus, it is not surprising that most special education programs in the United States fail to close the gap in reading skills for the children they serve (Hanushek, Kain, and Rivkin 1998; Vaughn, Moody, and Schuman 1998).

As an alternative to such special education programs, many of the nation's school districts are spending substantial resources—hundreds of millions of dollars—on educational products and services developed by textbook publishers, commercial providers, and nonprofit organizations. Several studies have recently shown that intensive, skillfully-delivered instruction can accelerate the development of reading skills in children with very severe reading disabilities, and do so at a much higher pace than is typically observed in special education programs (Lovett et al. 2000; Rashotte, Torgesen, and McFee 2001; Torgesen et al. 2001; Wise, Ring, and Olson 1999). Yet, we know little about the effectiveness of these interventions for broader populations of struggling readers in regular school settings. Which interventions work best, and for whom? Under what conditions are they most effective? Do these programs have the potential to close the reading gap between struggling and average readers?

To help answer these questions, we designed an experimental evaluation of four widely used programs for elementary school students with reading problems. Before describing these programs and the evaluation in detail, we review the findings from studies that have assessed the specific reading difficulties encountered by struggling readers.

B. READING DIFFICULTIES AMONG STRUGGLING READERS

The available data demonstrate that a large fraction of students in the late elementary school grades are unable to read at a basic level. However, to design effective instructional approaches that will substantially improve these students' reading skills, we must understand the specific nature of their reading difficulties. Research on this issue has revealed that struggling readers in late elementary school typically have problems with (1) accuracy, (2) fluency, and (3) comprehension.

When asked to read passages at their grade level, struggling readers make many more errors in reading the words as compared with average readers (Manis, Custodio, and Szeszulski 1993; Stanovich and Siegel 1994). Two limitations in reading skill typically underlie these *accuracy* problems. When struggling readers

encounter an unfamiliar word, they tend to place too much reliance on guessing it based primarily on the context or meaning of the passage (Share and Stanovich 1995). They are typically forced to guess from context because their phonemic analysis skills—their ability to use "phonics" to assist in the word identification process—are significantly impaired (Bruck 1990; Siegel 1989). The other underlying limitation is that in grade-level text, children with reading difficulties encounter more words that they cannot read "by sight" than do average readers (Jenkins et al. 2003).

Lack of ability to accurately recognize many words that occur in grade-level text (limited "sight word" vocabulary) also limits these children's reading *fluency*. In fact, recent research has demonstrated that the primary factor that limits struggling readers' fluency is the high proportion of words in grade-level text that they cannot recognize at a single glance (Jenkins, Fuchs, van den Broek, Espin, and Deno 2003; Torgesen and Hudson in press; Torgesen, Rashotte, and Alexander 2001). Problems with reading fluency are emerging as one of the most common and difficult to remediate traits of older struggling readers (Torgesen and Hudson in press). For example, a recent study of the factors associated with unsatisfactory performance on one state's third-grade reading accountability measure—a measure of comprehension of complex text—found that students reading at the lowest of five levels on the test had reading fluency scores at the 6th percentile (Schatschneider et al. 2004).

The third type of reading problem experienced by almost all struggling readers in late elementary school involves difficulties *comprehending* written text. For many poor readers, comprehension difficulties are caused primarily by accuracy and fluency problems (Share and Stanovich 1995). Children in this group often have average to above-average general verbal or language comprehension skills, but their ability to comprehend text is hampered by their limited ability to read words accurately and fluently. When their word-level reading problems are remediated, their reading comprehension skills tend to improve to a level that is more consistent with their general verbal skills (Snowling 2000; Torgesen et al. 2001). The weak comprehension skills of children in another large group of poor readers are attributable to not only accuracy and fluency problems but also general verbal skills—particularly vocabulary skills—that are significantly below average (Snow, Burns, and Griffen 1998), often because their home environments have not exposed them to rich language learning opportunities (Hart and Risley 1995). Even when the word-level reading skills of these children are brought into the average range, they may continue to struggle with comprehension because they lack the vocabulary and background knowledge necessary to understand complex text at the upper elementary level. Finally, poor readers in mid- to late elementary school are also frequently deficient in the use of effective comprehension strategies because they missed opportunities to acquire them while struggling to read words accurately or were not taught them explicitly by their reading teachers (Brown, Palincsar, and Purcell l986; Mastropieri and Scruggs 1997).

C. STRATEGIES FOR HELPING STRUGGLING READERS

In light of what has been learned about the specific reading problems of poor readers, we designed this evaluation to contrast two intervention classifications. One of these intervention classifications—referred to as w*ord level*—includes methods that focus on improving word-level reading skills so that they no longer limit children's ability to comprehend text. Such methods devote the majority of their instructional time to establishing phonemic awareness, phonemic decoding skills, and word and passage reading fluency. Methods in this classification sometimes include activities to check comprehension (such as asking questions and discussing the meaning of what is read), but this instruction is incidental to the primary focus on improving word-level reading skills. The bulk of instructional and practice time in methods included within this classification is focused on building children's ability to read text accurately and fluently. The second intervention classification—referred to as *word level plus comprehension*—includes methods that more evenly balance instructional time between activities to build word-level skills and activities devoted to building vocabulary and reading comprehension strategies. These interventions include extended activities that are designed to increase comprehension and word knowledge

(vocabulary), and these activities would take roughly the same amount of instructional time as the activities designed to increase word reading accuracy and fluency.

Although we sought to contrast word level and word level plus comprehension methods, we did not design new instructional programs to fit these two classifications. Rather, we employed either parts or all of four existing and widely used remedial reading instructional programs: Corrective Reading, Failure Free Reading, Spell Read P.A.T, and Wilson Reading. These four interventions were selected from more than a dozen potential program providers. The selection was done by members of the Scientific Advisory Board of the Haan Foundation for Children. The Haan Foundation coordinated the selection process and funding for the interventions.[4] The decision to modify these intact programs was justified both because it created two treatment classes that were aligned with the different types of reading deficits observed in struggling readers (discussed above) and because it gave us sufficient statistical power to contrast the relative effectiveness of the two classes. There were not enough schools available in the sample to support direct contrasts of effectiveness between the programs considered individually. Because Corrective Reading and Wilson Reading were both modified in order to fit them within the two treatment classes, results from this study do not provide complete evaluations of these interventions; instead, the results suggest how interventions using primarily the word level components of these programs will affect reading achievement.

Another potentially important difference between the instructional emphases of the interventions in this evaluation and how such programs might be implemented in a nonresearch school setting or a clinical setting is that in these other settings, the balance of activities within a program can be varied to suit the needs of individual students. Within the context of this study, however, the relative balance of instructional activities between word-level skills and vocabulary/comprehension skills was to be held constant across students within each program. Despite this restriction, it was still possible for instructors to vary, for example, the rate of movement through the instructional content or the specific vocabulary taught according to children's needs.

Finally, all four interventions delivered instruction to groups of three students "pulled out" of their regular classroom activities. Although "pull out" methods for remedial instruction have received some criticism over the last 20 years (Speece and Keogh 1996), we specified this approach for several reasons. First, all of the smaller-scale research that has produced significant acceleration of reading growth in older students used some form of a "pull out" method, with instruction delivered either in small groups or individually. Second, we are aware of no evidence that the level of intensity of instruction required to significantly accelerate reading growth in older students can be achieved by inclusion methods or other techniques that do not teach students in relatively small, homogeneous groups for regular periods of time every day (Zigmond 1996). Although the type of instruction offered in this study might be achieved by "push in" programs in which small groups are taught within their regular classroom, this was not a practical solution for this study because our instructional groups of struggling readers were comprised of children assigned to several different regular classrooms within each school.[5]

From this discussion, it is evident that this study is an evaluation of interventions that both focus on particular content and are delivered in a particular manner. Our decision to manipulate both of these dimensions simultaneously is consistent with one of the most important goals of the study: to examine

[4] A complete list of members of the advisory board is provided in Appendix Q.

[5] One implication of providing pull out instruction is that the intervention students might receive less reading instruction in their regular classrooms or through other instruction provided by their schools. The implementation study revealed that this occurred.

the extent to which the reading skills of struggling readers in grades three and five could be significantly accelerated if high quality instruction was delivered with sufficient intensity and skill. It also means, of course, that if there is a significant impact of an intervention compared to the control group, the impact could be related to either the increased intensity of instruction or to the particular focus of the intervention.

D. EVALUATION DESIGN AND IMPLEMENTATION

We designed the evaluation to address a number of different questions, only some of which are addressed in this initial report. In this report, we provide preliminary answers to the following questions:

1. What is the impact of being in any of the four remedial reading interventions, considered as a group, relative to the instruction provided by the schools? What is the impact of being in one of the remedial reading programs that focuses primarily on developing word-level skills, considered as a group, relative to the instruction provided by the schools? What is the impact of being in each of the four particular remedial reading interventions, considered individually, relative to the instruction provided by the schools?
2. Do the impacts of programs vary across students with different baseline characteristics?
3. To what extent can the instruction provided in this study close the reading gap and bring struggling readers within the normal range, relative to the instruction provided by their schools?

We implemented the evaluation in the Allegheny Intermediate Unit (AIU), which is located just outside Pittsburgh, Pennsylvania. The evaluation is a large-scale, longitudinal evaluation comprising two main elements. The first element of the evaluation is an impact study of the four interventions based on a scientifically rigorous design—an experimental design that uses random assignment at two levels: (1) 50 schools from 27 school districts in the AIU were randomly assigned to one of the four interventions and (2) within each school, eligible children in grades 3 and 5 were randomly assigned to a treatment group or to a control group. Students assigned to the intervention group (treatment group) were placed by the program providers and local coordinators into instructional groups of three students. Students in the control groups received the same instruction in reading that they would have ordinarily received.

Children were defined as eligible if they were identified by their teachers as struggling readers and if they scored at or below the 30th percentile on a word-level reading test and at or above the 5th percentile on a vocabulary test. From an original pool of 1,576 3rd and 5th grade students identified as struggling readers, 1,042 also met the test-score criteria. Of these eligible students, 772 were given permission by their parents to participate in the evaluation.

The second element of the evaluation is an implementation study that has two components: (1) an exploration of the similarities and differences in reading instruction offered in the four interventions and (2) a description of the regular instruction that students in the control group received in the absence of the interventions and the regular instruction received by the treatment group beyond the interventions.

The interventions provided instruction to students in the treatment group from the first week of November 2003 through the first weeks in May 2004. During this time, the students received, on average, about 90 hours of instruction, which was delivered five days a week to groups of three students in sessions that were approximately 50 minutes long. A small amount of the instruction was delivered in groups of two, or one on one, because of absences and make-up sessions.

The teachers who provided intervention instruction were recruited from participating schools on the basis of experience and the personal characteristics relevant to teaching struggling readers. They received, on average, nearly 70 hours of professional development and support during the implementation year.

To address the research questions presented above, we are collecting test data and other information on students, parents, teachers, classrooms, and schools several times over a three-year period. Key data collection points pertinent to this initial report include the period just before the interventions began, when baseline information was collected, and the period immediately after the interventions ended, when follow-up data were collected. Additional follow-up data for students and teachers are being collected in 2005 and again in 2006. In this report, we present findings from the implementation study and estimates of the impacts of the interventions just after the interventions ended.

II. DESIGN AND IMPLEMENTATION OF STUDY

This evaluation has two main elements: (1) an impact study and (2) an implementation study. The implementation study examines the instruction provided by the four interventions and the instruction provided outside of the interventions to both the students who participated in the interventions and those who did not. Although this chapter describes some of the data that we have collected for the implementation study, we describe the design and findings of that study in detail in the next chapter.

This chapter focuses mainly on the impact study. The impact study is based on a scientifically rigorous design—an experimental design that uses random assignment at two levels: (1) schools were randomly assigned to one of the four interventions, and (2) within each school, eligible children in grades three and five were randomly assigned to a treatment group or to a control group. Randomization at the school-level was done so that the interventions would be implemented within similar schools. Randomization at the student-level ensures that the students in the treatment and control groups are only randomly different from one another on all background covariates, including reading ability at the beginning of the school year. Thus, differences in outcomes at the end of the school year can be attributed to the interventions and not to pre-existing differences between the groups.[6] All student-level analyses account for the clustering of students within schools, as detailed in Chapter IV.

In the remainder of this chapter, we describe how schools and students were randomized. Then, we describe the data that we have collected for the evaluation.

A. THE RANDOM ASSIGNMENT OF SCHOOLS AND STUDENTS

1. Randomization of Schools

We implemented the intervention in the Allegheny Intermediate Unit (AIU), located just outside Pittsburgh, Pennsylvania. The AIU consists of 42 school districts and about 125 elementary schools. Not all schools that agreed to participate in the study had sufficient numbers of eligible third- and fifth-grade students, and some schools had only third or fifth grade, not both. Thus, we partnered some schools to form "school units" such that each school unit would have two third-grade and two fifth-grade instructional groups consisting of three students per instructional group. From a pool of 52 schools, we formed 32 school units, and randomly assigned the 32 school units to the four interventions, within four strata defined by the percentage of students eligible for free or reduced- price school lunch.

[6] A power analysis was done to estimate the minimum detectable impacts (MDI) given the study design, the actual number of schools and students enrolled, the variability in the follow-up test scores explained by the variability in baseline test scores, and the estimated intraclass correlation. For the power calculations, the two-tailed significance level is 0.05 with a power of 0.80. Other parameters are based on the observed data for two of the main reading measures: Word Attack and GRADE. The observed R-squared values between the baseline and follow-up tests are 0.48 and 0.35 for Word Attack and GRADE, respectively. The observed intraclass correlations for Word Attack and GRADE are 0.11 and 0.15, respectively. This analysis indicated that, when estimating separate impacts for third and fifth graders, the MDI's for testing whether the four interventions combined or the three word-level interventions combined had an impact are approximately 0.3 (in standard deviation units); the MDI for testing whether an individual intervention had an impact is approximately 0.55. When testing subgroup impacts, the MDI's for all interventions combined and for each intervention individually are approximately 0.35 and 0.7, respectively. A power analysis based on assumed values for relevant parameters and a desire to detect impacts of 0.5 standard deviations guided the design of the study.

One school unit (consisting of two schools) dropped out of the study after randomization but before it learned of its random assignment, leaving 31 school units and 50 schools in the study.[7,8]

To assess the similarity of the intervention groups after randomly assigning schools, Table II.1 shows the distribution of school unit–level covariates across the four groups of school units assigned to each intervention. Appendix A also compares the schools in the study with other schools in the AIU and with schools nationwide. Tables II.2 and II.3 present comparisons based on student-level covariates, and the final columns of each of those tables also show tests of significance for differences in student-level covariates across the four interventions (for grades three and five, respectively). The only two significant differences in the school unit–level covariates across the four interventions are both attributable to differences in school size. By chance, five of the six smallest schools were assigned to Wilson Reading and so some of the variables directly related to enrollment (total enrollment and average class size) differ across the four interventions. On student-level covariates, we observe only a difference on the racial distribution in the schools. With just 32 school units randomized, it is not surprising to observe some differences among the four groups. While small differences may affect the inferences we draw from the impact analysis when comparing interventions, our impact analyses are based on the differences in reading achievement for students in treatment and control groups within school units rather than between school units. Thus, small differences among interventions are not critical and should not bias our impact estimates for individual interventions. In addition, when the student-level randomization is assessed, the students in the treatment and control groups are very similar to each other (see Tables II.2 through II.5).

2. Randomization of Students

After we randomized school units to one of the four interventions, we randomized the eligible students within each school and grade either to receive the intervention (the treatment group) or not to receive the intervention (the control group). The student-level randomization process was as follows:[9]

- ***Identify Potentially Eligible Students***. Teachers in the 50 schools identified 1,576 struggling readers in third or fifth grade for screening. Nearly all (1,502) of these students were screened.[10]

[7] Because we did not collect data from the two schools that dropped out, we cannot include those schools in the analyses. Exclusion of those schools could have affected the comparisons across the four interventions by making the distributions of students across the interventions slightly different. However, an analysis of the distributions of student-level covariates across the four interventions shows that the effects of the school exclusions were minimal (see Tables II.2 and II.3).

[8] Figure A.1 of Appendix A illustrates the selection of schools and the process of randomizing school units to the four interventions.

[9] Separately for each intervention, Figures A.2 through A.5 of Appendix A show the details of students' progression through the study. Appendix A also details the data collection process.

[10] For the following reasons, 74 students were not screened: the parents returned passive consent forms that declined screening (37), students transferred to other schools before the screening (25), or other reasons (12), such as expulsion, retention in the previous grade, home schooling, or severe disability.

Table II.1

Characteristics of School-Units Assigned to the Four Intervention Groups

School Characteristics	Failure Free Reading	Spell Read	Wilson Reading	Corrective Reading	
Measurements of School Size					
Total enrollment	506	563	389	508	*
Average enrollment per grade	118	113	68	118	
Number of grades in school	5	5	6	5	
Both 3rd and 5th grades in school	0.88	0.63	1.00	0.63	
Number of 3rd grade classes	4.4	5.0	3.4	4.4	
Number enrolled in 3rd grade	110	118	69	95	
Number of 5th grade classes	5.9	4.6	3.2	5.7	
Number enrolled in 5th grade	153	116	69	144	
Average class size	25	24	21	23	*
Characteristics of Students in the School					
Percent eligible for free or reduced price lunch	0.35	0.36	0.40	0.34	
Fraction of students who leave during the year	0.04	0.03	0.09	0.09	
Percent white	0.85	0.70	0.76	0.82	
Percent African American	0.14	0.29	0.23	0.16	
School-wide Title 1	0.88	0.71	0.71	0.88	
Sample Size	8	8	8	8	

Note: Includes all school-units randomly assigned. Within a school-unit, each school given equal weight.

* Difference across interventions is statistically significant at the 0.05 level.

- ***Determine Eligibility***. Of those 1,502 students screened, 1,042 were eligible for the study based on the following eligibility criteria:
  - Scoring at or above the fifth percentile on a test of verbal ability (Peabody Picture Vocabulary Test—Revised)
  - Scoring at or below the 30th percentile on a word-level reading ability test (Test of Word Reading Efficiency (TOWRE), Phonemic Decoding Efficiency and Sight Word Efficiency subtests combined)
  - Students were also required to have written parental consent to participate in the study; 779 of the test-score eligible students received this consent.

- ***Randomly Assign Eligible Students to the Treatment and Control Groups***. 772 of the eligible students who had parental consent were randomized to the treatment group or the control group.[11] Within each school unit and grade, 3, 6, or 12 eligible students were randomly chosen to receive the intervention.[12] A total of 458 students were assigned to the treatment group. The remaining 314 students were assigned to the control group. Once students were assigned to the treatment group within a school, program operators assigned the treatment students to instructional groups composed of three students each, based on each program's own test results and constraints regarding students' schedules.

[11] Seven of the 779 students were not randomized because they came from grades in schools from which we obtained an insufficient number of eligible students or from schools in which we did not use students from that grade (because students from another school in the same school unit were included in the study instead).

[12] The number of students in each school and grade chosen to receive the treatment depended on the number of intervention slots available (based on expectations of the number of eligible students per school).

Table II.2

Baseline Characteristics of the Four Intervention Groups and the Control Group,
Analysis Sample: 3rd Grade

Baseline Means	Failure Free Reading		Spell Read		Wilson Reading			Corrective Reading		
	Treat.	Cont.	Treat.	Cont.	Treat.	Cont.		Treat.	Cont.	
Student Characteristics										
Age	8.6	8.7	8.8	8.7	8.7	8.6		8.7	8.7	
Male (%)	53	58	73	59	39	21		56	48	#
Hispanic (%)	a	a	a	a	a	a		a	a	
Race--White (%)	76	81	65	68	55	68		74	82	
Race--African American (%)	24	19	35	32	45	32		26	18	
Race--Other (%)	a	a	a	a	a	a		a	a	
Family income less than $30,000 (%)	42	41	57	49	48	56		41	56	
Family income between $30,000 and $60,000 (%)	47	42	20	24	32	44		41	14	*
Family income over $60,000 (%)	11	17	23	27	a	a		18	30	
Eligible for free or reduced price lunch (%)	45	49	46	36	36	64	*	42	48	
Has any learning or other disability (%)	40	46	35	25	34	19		30	43	
Mother has bachelor's degree or higher (%)	14	9	13	15	a	a		19	11	
Screening Tests										
TOWRE Sight Word Efficiency	84.0	82.0	85.4	84.7	86.8	84.6		85.3	82.2	
TOWRE Phonemic Decoding Efficiency	84.1	85.1	85.7	85.0	86.1	86.0		85.7	87.1	
Peabody Picture Vocabulary Test--Revised	93.5	94.6	95.5	97.8	90.4	90.5		97.6	96.5	
Baseline Tests										
WRM Word Identification	88.6	87.2	89.5	87.2	90.6	89.8		89.7	87.7	
TOWRE Phonemic Decoding Efficiency	84.2	84.3	86.2	84.6	87.0	86.2		87.1	85.9	
WRM Word Attack	90.0	89.2	93.8	91.4	94.7	94.3		93.8	94.7	
TOWRE Sight Word Efficiency	86.9	84.5	89.3	86.6	89.0	84.1	*	86.9	84.0	
AIMSWeb (Raw score)	37.7	33.6	46.8	41.4	49.3	41.0		43.4	34.4	
WRM Passage Comprehension	90.7	88.5	95.2	89.9	93.8	92.9		94.2	89.7	
GRADE	86.1	84.9	87.8	83.9	88.6	85.8		89.8	84.1	*
Woodcock Johnson Spelling	90.0	86.5	89.4	89.0	89.3	87.8		90.5	85.9	
Woodcock Johnson Calculation	92.4	96.8	99.3	95.1	96.9	92.4		96.9	92.8	
Other Baseline Tests Administered										
RAN Colors	88.0	85.9	88.8	85.4	88.7	86.7		89.8	88.3	
RAN Letters	87.3	84.7	91.5	87.2	90.3	85.6	*	85.3	83.4	
RAN Numbers	88.7	86.6	94.5	88.0	94.7	88.9		90.2	84.8	
RAN Objects	87.3	85.0	89.8	83.8	90.6	87.9		91.3	86.7	
RAS Numbers and Letters	87.1	85.3	92.1	86.3	* 90.8	84.0	*	86.2	84.7	
RAS Colors, Numbers, and Letters	85.9	86.5	90.4	83.9	89.7	86.9		85.8	84.2	
CTOPP Blending Words	7.3	6.9	8.4	8.3	8.0	7.1	*	7.9	7.5	
CTOPP Elision	7.7	7.4	8.6	8.6	7.9	8.1		7.8	7.3	
CTOPP Rapid Digit Naming	7.7	7.4	8.3	7.5	8.2	7.2	*	8.6	7.7	
CTOPP Rapid Letter Naming	8.5	8.1	8.9	8.6	8.7	7.8		9.0	8.5	
Clinical Evaluation of Language Fundamentals-IV	7.1	7.1	6.9	7.3	8.3	9.3		8.2	6.5	*
Sample Size	55	38	56	36	53	18		44	35	

Note: Weights used to account for differential randomization probabilities and nonresponse.

Note: All test scores are shown as standard scores, unless otherwise indicated. All standard scores have mean 100 and standard deviation 15, except for CTOPP and Clinical Evaluation of Language Fundamentals-IV, which have mean 10 and standard deviation 3.

* Difference between treatment and control groups is statistically significant at the 0.05 level.

Difference across the four interventions (with treatment and control groups pooled within each intervention) is statistically significant at the 0.05 level.

a Values suppressed to protect student confidentiality.

Table II.3

Baseline Characteristics of the Four Intervention Groups and the Control Group,
Analysis Sample: 5th Grade

Baseline Means	Failure Free Reading		Spell Read		Wilson Reading			Corrective Reading		
	Treat.	Cont.	Treat.	Cont.	Treat.	Cont.		Treat.	Cont.	
Student Characteristics										
Age	10.6	10.7	10.8	10.7	10.8	10.5	*	10.7	10.7	
Male (%)	53	51	54	58	54	66		49	64	
Hispanic (%)	a	a	a	a	a	a		a	a	
Race--White (%)	78	83	75	67	55	59		83	88	
Race--African American (%)	22	17	25	33	45	41		18	12	
Race--Other (%)	a	a	a	a	a	a		a	a	
Family income less than $30,000 (%)	41	50	51	59	73	47	*	32	52	
Family income between $30,000 and $60,000 (%)	43	33	39	34	23	36		43	30	
Family income over $60,000 (%)	16	17	11	7	a	a	*	25	18	
Eligible for free or reduced price lunch (%)	42	45	52	43	55	42		41	41	
Has any learning or other disability (%)	27	38	26	35	31	30		30	37	
Mother has bachelor's degree or higher (%)	12	17	5	9	a	a		15	23	
Screening Tests										
TOWRE Sight Word Efficiency	84.0	85.1	83.8	85.3	83.9	84.8		82.7	83.8	
TOWRE Phonemic Decoding Efficiency	81.5	79.7	78.4	80.4	81.9	82.2		79.9	79.9	
Peabody Picture Vocabulary Test--Revised	94.6	95.2	92.0	92.1	91.6	100.0	*	95.1	98.9	
Baseline Tests										
WRM Word Identification	90.3	89.0	87.1	88.0	87.9	90.0		87.5	89.5	
TOWRE Phonemic Decoding Efficiency	82.0	81.8	77.9	80.5	82.8	81.2		80.7	81.1	
WRM Word Attack	93.4	92.9	90.7	93.5	93.4	94.4		93.6	93.4	
TOWRE Sight Word Efficiency	84.1	85.5	82.9	85.8	84.2	84.6		83.6	83.1	
AIMSWeb (Raw score)	78.7	75.0	79.5	80.3	75.0	80.2		75.8	74.8	
WRM Passage Comprehension	92.4	92.2	91.4	93.2	90.6	96.4	*	91.8	93.4	
GRADE	91.4	92.1	89.9	90.4	92.1	95.2		88.2	92.4	
Woodcock Johnson Spelling	93.9	92.1	89.7	91.9	91.2	92.3		88.4	86.8	
Woodcock Johnson Calculation	94.3	93.3	94.9	95.5	94.0	95.2		94.0	95.2	
Other Baseline Tests Administered										
RAN Colors	89.7	91.8	90.6	90.8	90.8	86.9		93.4	87.0	*
RAN Letters	92.9	93.9	93.0	92.3	91.2	90.0		92.4	90.7	
RAN Numbers	95.2	94.9	95.7	94.5	93.5	92.9		94.1	93.2	
RAN Objects	90.0	94.4	89.9	88.2	90.4	85.7		91.6	88.1	
RAS Numbers and Letters	91.2	92.0	92.3	93.5	90.2	90.2		91.9	89.4	
RAS Colors, Numbers, and Letters	92.0	91.5	93.1	92.9	89.9	91.9		89.3	87.5	
CTOPP Blending Words	7.3	7.2	7.4	7.9	6.8	6.9		7.3	7.1	
CTOPP Elision	8.0	8.0	6.7	7.9	* 6.8	7.4		7.2	7.8	
CTOPP Rapid Digit Naming	8.0	8.1	7.8	7.9	7.8	7.7		8.4	8.0	
CTOPP Rapid Letter Naming	8.2	8.8	8.5	8.2	7.7	8.4		8.8	8.6	
Clinical Evaluation of Language Fundamentals-IV	9.0	8.4	9.0	8.9	9.1	8.1		6.0	5.8	
Sample Size	61	65	59	45	53	38		55	31	

Note: Weights used to account for differential randomization probabilities and nonresponse.

Note: All test scores are shown as standard scores, unless otherwise indicated. All standard scores have mean 100 and standard deviation 15, except for CTOPP and Clinical Evaluation of Language Fundamentals-IV, which have mean 10 and standard deviation 3.

* Difference between treatment and control groups is statistically significant at the 0.05 level.

Difference across the four interventions (with treatment and control groups pooled within each intervention) is statistically significant at the 0.05 level.

a Values suppressed to protect student confidentiality.

Table II.4

Baseline Characteristics of Full Sample and Three Word-level Interventions, by Treatment Status, Analysis Sample: 3rd Grade

Baseline Means	All Interventions			Word-level Interventions		
	Treatment	Control		Treatment	Control	
Student Characteristics						
Age	8.7	8.6		8.7	8.6	
Male (%)	56	47		57	43	
Hispanic (%)	2	2		a	a	
Race--White (%)	68	75		65	72	
Race--African American (%)	32	25		35	28	
Race--Other (%)	a	a		a	a	
Family income less than $30,000 (%)	48	50		50	54	
Family income between $30,000 and $60,000 (%)	34	32		30	28	
Family income over $60,000 (%)	18	18		20	18	
Eligible for free or reduced price lunch (%)	42	49		42	49	
Has any learning or other disability (%)	35	33		33	29	
Mother has bachelor's degree or higher (%)	13	10		12	11	
Screening Tests						
TOWRE Sight Word Efficiency	85.3	83.4		85.8	83.9	
TOWRE Phonemic Decoding Efficiency	85.4	85.8		85.9	86.0	
Peabody Picture Vocabulary Test--Revised	94.2	94.9		94.5	95.0	
Baseline Tests						
WRM Word Identification	89.6	87.9		89.9	88.2	
TOWRE Phonemic Decoding Efficiency	86.0	85.2		86.7	85.5	
WRM Word Attack	93.0	92.2		94.1	93.4	
TOWRE Sight Word Efficiency	88.1	84.9	*	88.5	85.0	*
AIMSWeb (Raw score)	44.2	37.6	*	46.5	39.1	*
WRM Passage Comprehension	93.4	90.2		94.4	90.8	
GRADE	88.0	84.7	*	88.7	84.6	*
Woodcock Johnson Spelling	89.8	87.4		89.7	87.7	
Woodcock Johnson Calculation	96.3	94.4	*	97.8	93.5	*
Other Baseline Tests Administered						
RAN Colors	88.8	86.5		89.1	86.7	
RAN Letters	88.7	85.3	*	89.2	85.5	*
RAN Numbers	92.0	87.1	*	93.2	87.3	*
RAN Objects	89.6	85.8		90.5	86.0	
RAS Numbers and Letters	89.1	85.1	*	89.8	85.0	*
RAS Colors, Numbers, and Letters	87.9	85.4		88.7	84.9	
CTOPP Blending Words	7.9	7.5		8.1	7.7	
CTOPP Elision	8.0	7.8		8.1	8.0	
CTOPP Rapid Digit Naming	8.2	7.5	*	8.4	7.5	*
CTOPP Rapid Letter Naming	8.8	8.3	*	8.9	8.3	
Clinical Evaluation of Language Fundamentals-IV	7.6	7.5		7.8	7.7	
Sample Size	208	127		153	89	

Note: Weights used to account for differential randomization probabilities and nonresponse.

Note: All test scores are shown as standard scores, unless otherwise indicated. All standard scores have mean 100 and standard deviation 15, except for CTOPP and Clinical Evaluation of Language Fundamentals-IV, which have mean 10 and standard deviation 3.

* Difference between treatment and control groups is statistically significant at the 0.05 level.

a Values suppressed to protect student confidentiality.

Table II.5

Baseline Characteristics of Full Sample and Three Word-level Interventions, by Treatment Status, Analysis Sample: 5th Grade

Baseline Means	All Interventions Treatment	All Interventions Control		Word-level Interventions Treatment	Word-level Interventions Control	
Student Characteristics						
Age	10.7	10.6		10.8	10.6	
Male (%)	53	60		53	62	
Hispanic (%)	a	a		a	a	
Race--White (%)	73	75		72	72	
Race--African American (%)	27	25		28	28	
Race--Other (%)	a	a		a	a	
Family income less than $30,000 (%)	48	52		51	53	
Family income between $30,000 and $60,000 (%)	38	33		36	34	
Family income over $60,000 (%)	14	15		13	14	
Eligible for free or reduced price lunch (%)	47	43		49	42	
Has any learning or other disability (%)	28	35		29	34	
Mother has bachelor's degree or higher (%)	8	16	*	7	16	*
Screening Tests						
TOWRE Sight Word Efficiency	83.6	84.8		83.4	84.6	
TOWRE Phonemic Decoding Efficiency	80.4	80.5		80.0	80.8	
Peabody Picture Vocabulary Test--Revised	93.4	96.5	*	92.9	96.9	*
Baseline Tests						
WRM Word Identification	88.2	89.1		87.5	89.2	
TOWRE Phonemic Decoding Efficiency	80.8	81.1		80.4	80.9	
WRM Word Attack	92.8	93.5		92.5	93.8	
TOWRE Sight Word Efficiency	83.7	84.8		83.5	84.5	
AIMSWeb (Raw score)	77.3	77.5		76.8	78.4	
WRM Passage Comprehension	91.6	93.7		91.3	94.3	*
GRADE	90.4	92.5		90.0	92.6	
Woodcock Johnson Spelling	90.8	90.7		89.7	90.3	
Woodcock Johnson Calculation	94.3	94.8		94.4	95.3	
Other Baseline Tests Administered						
RAN Colors	91.2	89.2	*	91.7	88.3	*
RAN Letters	92.4	91.8		92.2	91.0	
RAN Numbers	94.7	93.9		94.5	93.5	
RAN Objects	90.5	89.2		90.6	87.4	*
RAS Numbers and Letters	91.4	91.3		91.5	91.0	
RAS Colors, Numbers, and Letters	91.1	90.9		90.8	90.7	
CTOPP Blending Words	7.2	7.3		7.2	7.3	
CTOPP Elision	7.2	7.8		6.9	7.7	*
CTOPP Rapid Digit Naming	8.0	7.9		8.0	7.9	
CTOPP Rapid Letter Naming	8.3	8.5		8.4	8.4	
Clinical Evaluation of Language Fundamentals-IV	8.3	7.8		8.0	7.6	
Sample Size	228	179		167	114	

Note: Weights used to account for differential randomization probabilities and nonresponse.

Note: All test scores are shown as standard scores, unless otherwise indicated. All standard scores have mean 100 and standard deviation 15, except for CTOPP and Clinical Evaluation of Language Fundamentals-IV, which have mean 10 and standard deviation 3.

* Difference between treatment and control groups is statistically significant at the 0.05 level.

a Values suppressed to protect student confidentiality.

Using all 1,502 students screened, Table II.6 compares the test scores of the 1,042 students eligible based on test scores with the 460 students ineligible based on test scores. As the eligibility criteria would suggest, the eligible students demonstrated lower word-level reading ability (as measured by the TOWRE test) than the ineligible students but higher verbal ability (as measured by the Peabody Picture Vocabulary test).[13] Table II.7 compares the test scores of the 263 students eligible based on test scores but whose parents did not give consent with the 779 students fully eligible based on test scores and consent; 772 of the eligible students were randomly assigned to the treatment or control group. There is only one statistically significant difference in the average screening test scores of the two groups, indicating that the students who received consent are similar to the students who did not receive consent, at least on these measures of word-level reading and verbal ability.

The study had almost no nonresponse at baseline or follow-up data collection, and most students received the instruction for the group to which they were assigned. That is, no control students received the intervention, and few treatment students did not receive any intervention. In particular, 13 students assigned to the treatment group did not receive any intervention; of the 13, 9 did not receive the intervention but remained in the study while 4 withdrew from the study. An additional 3 treatment students and 2 control students withdrew from the study after the first week.[14]

The final analysis sample contains fewer students (742) than the 772 students randomized to one of the interventions. The study dropped 30 students for one of two reasons: either they were in one school unit that did not have any control students, or they did not take the follow-up tests at the end of the school year. Specifically, in the Corrective Reading group, one school unit did not have enough eligible students to allow for any control students. Given that the absence of controls prevents a comparison of treatment and control outcomes in that school unit, we dropped the 9 treatment students in the school unit from the analysis.[15] In addition, 21 students (13 treatments and 8 controls) did not take any of the reading tests at the end of the school year.[16] For each intervention and grade, Tables II.2 and II.3 separately compare the covariates of students in the treatment and control groups in the final analysis sample; Tables II.4 and II.5 do the same for all interventions combined and the three word-level interventions combined.

Even though all the mean scores for intervention and control group students are below average for the students' grade level, Tables II.4 and II.5 demonstrate that these students are, on average, only moderately impaired in word-level reading skills. For example, on the widely used measures from the Woodcock Reading Mastery Test-Revised (WRMT-R, Woodcock 1998), the third-grade students in the treatment groups achieved average standard scores of 90, 93, and 93 on the Word Identification, Word Attack, and Passage Comprehension tests, respectively. These scores fall between the 25th and 32nd

[13] Among third graders, the difference in Peabody Picture Vocabulary test scores between eligible and ineligible students was not statistically significant at the .05 level. The scores were significantly different between eligible and ineligible fifth-grade students.

[14] The 9 withdrawals resulted from students' moves to a new school, parents not wanting their child in the control group, emotional issues, a student scoring well on the intervention's test, the student missing out on something in the regular classroom, and other unspecified reasons. The 13 treatment group drop-outs were the result of severe behavioral issues, parents not consenting to separating siblings, students' requests to leave the intervention, student stress/medication issues, students' moving, and other unspecified reasons.

[15] To permit estimation of school unit–level parameters, the hierarchical model used to estimate impacts requires treatment and control students within each school.

[16] Nearly half of these 21 students (9) had withdrawn from the study. Other nonrespondents at the end of the school year were not tested because of illness, difficulties in contacting the students, or because the student had moved.

Table II.6

Comparison of Eligible and Ineligible Students

Screening test scores	Eligible based on test scores Mean	Ineligible based on test scores Mean	Difference
Full Sample			
TOWRE Sight Word Efficiency	85	97	-12 *
TOWRE Phonemic Decoding Efficiency	83	96	-13 *
Peabody Picture Vocabulary Test--Revised (PPVT)	94	91	4
In Grade 3 (%)	44	59	-15 *
3rd Graders			
TOWRE Sight Word Efficiency	85	99	-14 *
TOWRE Phonemic Decoding Efficiency	85	97	-12 *
Peabody Picture Vocabulary Test--Revised (PPVT)	95	93	2
5th Graders			
TOWRE Sight Word Efficiency	84	93	-9 *
TOWRE Phonemic Decoding Efficiency	81	95	-14 *
Peabody Picture Vocabulary Test--Revised (PPVT)	94	87	7 *
Sample Size	1,042	460	

Note: The numbers in the "Difference" column may not exactly equal the difference between the numbers in the "Eligible" and "Ineligible" columns because of rounding. Estimates are unweighted.

Note: All test scores are shown as standard scores, unless otherwise indicated.

* Difference across groups is statistically significant at the 0.05 level.

percentiles, meaning that approximately half the students in the third-grade sample began the study with phonemic decoding scores above the 30th percentile and that many had scores solidly within the average range (between the 40th and 60th percentiles). The scores for fifth grade were similar: 88 for Word Identification, 93 for Word Attack, and 92 for Passage Comprehension. These baseline scores for word-level skills are much higher than corresponding scores from a set of 13 intervention samples recently reviewed by Torgesen (2005). The students in those studies were of approximately the same ages as those in the present study, and their average baseline standard score for Word Attack was 75 and their average baseline score for Word Identification was 73. These scores, which are below the fifth percentile, indicate that the average students in these other studies had reading skills that were substantially more impaired than the reading skills of the students in our sample and the population of struggling readers in the United States.

Within each intervention and grade, we observed a few significant differences in student characteristics at baseline between students assigned to the treatment group and students assigned to the control group (see Tables II.2 and II.3). Most of the differences are scattered across tests and interventions and are not surprising; a few differences would be expected even with random assignment. There are more significant differences when we compare the treatment and control groups in the

Table II.7

Comparison of Consenting and Nonconsenting Students, Among All Eligible

Screening test scores	Consenting Mean	Not consenting Mean	Difference
Full Sample			
TOWRE Sight Word Efficiency	84	85	-1
TOWRE Phonemic Decoding Efficiency	83	83	0
Peabody Picture Vocabulary Test--Revised (PPVT)	94	95	-1
In Grade 3 (%)	45	38	7 *
3rd Graders			
TOWRE Sight Word Efficiency	85	86	-1 *
TOWRE Phonemic Decoding Efficiency	85	85	1
Peabody Picture Vocabulary Test--Revised (PPVT)	95	97	-2
5th Graders			
TOWRE Sight Word Efficiency	94	95	-1
TOWRE Phonemic Decoding Efficiency	84	85	-1
Peabody Picture Vocabulary Test--Revised (PPVT)	81	82	-2
Sample Size	779	263	

Note: The numbers in the "Difference" column may not exactly equal the difference between the numbers in the "Eligible" and "Ineligible" columns because of rounding. Estimates are unweighted.

Note: All test scores are shown as standard scores, unless otherwise indicated.

* Difference across groups is statistically significant at the 0.05 level.

combined group of all interventions and the combined group of the three word-level interventions, particularly among third graders (see Tables II.4 and II.5).[17]

We also compared the distributions of covariates between the treatment and control groups within key subgroups defined by students' scores on the Word Attack test and by free or reduced-price school lunch eligibility. The results are broadly similar to those shown in Tables II.2 through II.5, with scattered differences across interventions but no apparent systematic differences between the treatment

[17] In fact, even if the covariate distributions were exactly the same in the treatment and control groups, we would expect 5 percent of the differences (1 of 20 characteristics) to be significantly different at the 0.05 level given the design of the statistical tests used here. When adjustments for multiple comparisons are made, many of the significant differences that are scattered across characteristics and interventions are no longer significant, although many of the differences seen among third graders in the four interventions combined remain. See Chapter IV and Appendix D for more discussion of the techniques used to adjust for multiple comparisons. We focus here on the results derived without any adjustment for multiple comparisons because not doing such an adjustment is in fact conservative when assessing balance in baseline covariates, unlike the situation when estimating impacts, where it is more conservative to do an adjustment.

and control groups. For third-grade students with low Word Attack scores, there are statistically significant differences in some test scores when comparing students in the Corrective Reading schools, and when comparing treatment and control students across the interventions combined. Almost no significant differences are seen for fifth-grade students with low Word Attack scores. For students with high Word Attack scores, almost no significant differences are seen for third-grade students, however there are some differences in the test scores of fifth-grade treatment and control group students in the Wilson Reading and Spell Read schools and when examining the interventions combined. Within the subgroup of students eligible for free or reduced-price school lunch, there are almost no differences between third-grade students in the treatment and control groups within each of the four interventions, but a few differences for fifth-grade students in the Spell Read and Corrective Reading schools. The results for students not eligible for free or reduced-price school lunch are very similar to those shown in Tables II.2-II.5 for the full sample, with some differences among third-grade students in Wilson Reading and when considering the interventions combined, and a few differences for fifth-grade students in Wilson Reading schools.

It is important to note that many of these reading tests are highly correlated with one another and thus the significance tests performed are not independent. For example, the Rapid Automatized Naming tests are all done at the same point in time and are testing similar skills (see Section B). Also, because students were randomly assigned to treatment or control status, the differences between the treatment and control groups are due entirely to chance. To adjust for these chance differences, we include the baseline value of each test as a predictor variable in the outcome models used to estimate impacts, a specification that was chosen before these differences were seen.

Depending on the number of eligible students in their school and grade, students had varying probabilities of assignment to the treatment group. Thus, all student-level analyses are conducted using weights that account for the unequal treatment probabilities and ensure that the treatment and control students weight up to represent the same population: that of all students in the study, where the students from each school are weighted proportional to the number of treatment slots given to that school. The weights also adjust for student dropout and nonresponse, and account for the randomization strata without any control students. Full details of the weighting procedure are given in Appendix C.

B. DATA

Test data and other information on students, parents, teachers, classrooms, and schools is being collected several times over a three-year period. Key data collection points pertinent to this report include the period just before the interventions began, when baseline information was collected, and the period immediately after the interventions ended, when follow-up data were collected. Additional follow-up data for students and teachers are being collected in 2005 and again in 2006. There are three major types of information used in this report: measures of student performance, measures of student characteristics and the instruction they received, and measures of study implementation and fidelity.

1. Measures of Student Performance

The tests used to assess student performance fall into three categories. First, seven measures of reading skill were administered at baseline and follow-up to assess student progress in learning to read. Second, measures of language skills were administered only at baseline in order to assess the relationship between individual differences in performance on these measures and individual differences in response to the interventions. Third, two other academic measures were administered at baseline and follow-up. A measure of spelling skill assessed the impact of remedial reading instruction on spelling ability, and a

measure of mathematical calculation skill assessed the impact of receiving the interventions in reading on an academic skill that is theoretically unrelated to improvements in reading. In a sense, the last measure is a "control" measure for effects of participation in the interventions on a skill that was not directly taught. The following describes each measurement category. Descriptions of each of these tests can be found in Exhibit 1 at the end of this chapter, and examples of items from the seven measures of reading skill can be found in Appendix L.

a. Measures of Reading

The measures of reading skills assessed phonemic decoding, word reading accuracy, text reading fluency, and reading comprehension. A sample test item from each of these tests is given in Appendix L. The seven tests, classified into three categories of reading skills, are:

Phonemic Decoding

- Word Attack (WA) subtest from the Woodcock Reading Mastery Test-Revised (WRMT-R; Woodcock 1998)
- Phonemic Decoding Efficiency (PDE) subtest from the Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, and Rashotte 1999)

Word Reading Accuracy and Fluency

- Word Identification (WI) subtest from the WRMT-R
- Sight Word Efficiency (SWE) subtest from the TOWRE
- Oral Reading Fluency subtest from Edformation, Inc., (Howe and Shinn, 2002). The text of this report refers to these passages as Aimsweb passages, which is the term used broadly in the reading practice community.

Reading Comprehension

- Passage Comprehension (PC) subtest from the WRMT-R
- Passage Comprehension from the Group Reading Assessment and Diagnostic Evaluation (GRADE; Williams 2001)

b. Measures of Language

These measures assessed phonemic awareness, rapid automatic naming ability, syntactic skill, and vocabulary. The tests included (1) the Peabody Picture Vocabulary test (PPVT-III; Dunn and Dunn 1997), (2) subtests from the Comprehensive Test of Phonological Processes (CTOPP; Wagner, Torgesen, and Rashotte 1999), (3) subtests from the Rapid Automatized Naming and Rapid Alternating Stimulus Tests (RAN/RAS; Wolf and Denkla 2005), and (4) a subtest from the Clinical Evaluation of Language Fundamentals-Fourth Edition (CELF; Semel, Wiig, and Secord 2003).

c. Measures of Spelling and Mathematics Calculation Ability

The spelling and calculation subtests from the Woodcock-Johnson III Tests of Achievement (WJ-III; Woodcock, McGrew, and Mather 2001) assessed spelling and mathematics calculation abilities.

2. Timing of Student-Level Data Collection and Correlations Among Measures

Table II.8 shows the time points during the study at which the above tests were administered, as well as estimates of the test reliability. Even though the above tests are grouped by the skills they measure, the correlations of the tests—even among tests measuring similar constructs—were not always large. For example, the correlation between the Word Attack and Phonemic Decoding Efficiency tests was .64, the average correlation among the three tests measuring word reading accuracy and fluency was .55, and the correlation between the Passage Comprehension and GRADE tests was .44. These correlations are somewhat lower in the present sample than those reported elsewhere for the same tests. For example, the manual for the TOWRE test (Torgesen, Wagner, and Rashotte 1999) reports a correlation of .91 between the Word Attack and Phonemic Decoding Efficiency tests for a sample of at-risk third-grade students. A correlation of .87 between the two tests was reported in the same manual for a large random sample of fifth-grade students. Similarly, the test manual also reported correlations between the Word Identification and Sight Word Efficiency tests for the same samples of third- and fifth-grade students at .92 and .86, respectively. The manual for the Woodcock Reading Mastery Test-Revised (Woodcock 1998) reports a correlation between the Word Identification measure and Passage Comprehension measure of .67 for third graders and .59 for fifth graders. The lack of a strong correlation between the two measures of reading comprehension may reflect several differences in the way the tests are administered and the types of required responses. Table II.9 presents the full set of correlations among the seven measures of reading. The shaded boxes indicate tests that measure similar constructs: baseline tests measuring phonemic decoding skills, baseline tests measuring reading fluency and accuracy, and baseline tests measuring reading comprehension.

For all tests except the Aimsweb passages, the analysis used grade-normalized standard scores, which indicate where a student falls within the overall distribution of reading ability among students in the same grade.[18,19] Scores above 100 indicate above-average performance; scores below 100 indicate below-average performance. In the population of students across the country at all levels of reading ability, standard scores are constructed to have a mean of 100 and a standard deviation of 15, implying that approximately 70 percent of all students' scores will fall between 85 and 115 and that approximately 95 percent of all students' scores will fall between 70 and 130.[20] For the Aimsweb passages, the score used in this analysis is the median correct words per minute from three grade-level passages.

[18] When possible, we standardized scores to the grade and month (e.g., we used different standardizations for fall and spring test administrations, when possible).

[19] We could not calculate standard scores for the Aimsweb test because the timing of the test administrations made it difficult to standardize the tests appropriately. Instead, the present report presents raw scores. As contrasted with the other tests, the raw score for the Aimsweb has a simple substantive meaning in that it corresponds to the number of words read correctly.

[20] The test standardizations use a "norming" population for each test, with data collected and analyzed by each test's publisher. The norming populations are selected to be representative of the national population of students at a given age or grade level.

Table II.8

Tests Administered at Beginning and End of the School Year

Test Administered at Screening, Baseline, and/or Follow-up	Screening (September-October)	Baseline (October-November)	Follow-up (May-June)	Reliability
Measures of Reading				
Phonemic Decoding				
Woodcock Test-R (WRMT-R) Word Attack (WA)		✓	✓	0.90[a]
Test of Word Reading Efficiency (TOWRE) Phonemic Decoding Efficiency (PDE)	✓	✓	✓	0.93[b]
Word Reading Accuracy and Fluency				
WRMT-R Word Identification (WI)		✓	✓	0.94[a]
TOWRE Sight Word Efficiency (SWE)	✓	✓	✓	0.95[b]
Aimsweb Oral Reading Passages (AIMS)		✓	✓	0.92[b]
Reading Comprehension				
WRMT-R Passage Comprehension (PCG)		✓	✓	0.82[a]
Group Reading Assessment and Diagnostic Evaluation Passage Comprehension (GRADE)		✓	✓	Grade 3: 0.88[c] Grade 5: 0.90[c]
Measures of Language				
Comprehensive Test of Phonological Processes (CTOPP)				
Phoneme Blending		✓		0.84[c]
Phoneme Elision		✓		0.89[c]
Rapid Automatic Naming of Letters		✓		0.92[c]
Rapid Automatic Naming of Numbers		✓		0.87[c]

TABLE II.8 (*continued*)

Test Administered at Screening, Baseline, and/or Follow-up	Screening (September-October)	Baseline (October-November)	Follow-up (May-June)	Reliability
Rapid Automated Naming (RAN)				
Colors		✓		0.90[d]
Objects		✓		0.84[d]
Numbers		✓		0.92[d]
Letters		✓		0.90[d]
Rapid Alternating Stimulus (RAS)				
2-set		✓		0.90[d]
3-set		✓		0.91[d]
Peabody Picture Vocabulary Test—Revised (PPVT-III)	✓			0.95[c]
Clinical Evaluation of Language Fundamentals–IV (CELF-IV) Formulated Sentences		✓		0.87[c]
Other Tests				
Woodcock Johnson Tests of Achievement-III (WJ-III)				
Spelling		✓	✓	0.89[c]
Calculation		✓	✓	0.85[c]

a: Split-half reliability
b: Alternate-form reliability
c: Internal consistency reliability
d: Test-retest reliability

Table II.9

Correlations among Reading Tests at Baseline (All Students)

	Word Attack	TOWRE PDE	Word Identification	TOWRE SWE	Aimsweb	Passage Comprehension	Grade
Word Attack	1.00	0.64	0.64	0.46	0.36	0.53	0.34
TOWRE PDE		1.00	0.59	0.62	0.28	0.43	0.26
Word Identification			1.00	0.66	0.48	0.58	0.40
TOWRE SWE Baseline				1.00	0.50	0.58	0.36
Aimsweb					1.00	0.44	0.45
Passage Comprehension						1.00	0.44
GRADE							1.00

3. Measures of Student Characteristics and Instruction Received

a. Parent Survey

A parent survey was administered at the time the letters of permission were sent to students' homes. The survey asked a range of questions concerning student background and demographic characteristics such as socioeconomic status (parental education and employment), school history (mobility), medical history, and primary language spoken in the home. In addition, the survey asked parents about their child's history of special tutoring in reading that occurred outside school.

b. Classroom Teacher Survey

Each child's regular classroom teacher completed a survey twice during the intervention year. The first survey, administered in the fall, asked the teacher to characterize the reading instruction each child received in the regular classroom as well as any special reading instruction or reading programs the child attended outside the regular classroom. If the student had an individual education plan (IEP) for special education, the teacher detailed the type of instruction specified. In addition to describing the instruction received by each child, the teacher reported on the instruction that each child in the intervention group typically missed when attending intervention sessions. As for the second survey administered in the spring, the teacher not only answered the same questions about instruction asked by the first survey but also filled out a classroom behavior rating form for each child. The behavior rating scales were adapted from the Multigrade Behavior Inventory (Agronin, Holahan, Shaywitz, and Shaywitz 1992) and Iowa-Connors Teacher Rating Scale (Loney and Milich l982).

c. Intervention Attendance Logs

To detail the amount of intervention instruction received by each student in the intervention group, each intervention teacher maintained an attendance log indicating the number of minutes of instruction received by each student each day.

4. Measures of Study Implementation and Fidelity

A variety of data sources were utilized in the implementation and fidelity analyses, including videotapes of instructional group sessions and ratings of teacher quality and program fidelity. To assess the intervention teachers, trainers from the individual reading programs and staff from the AIU rated each intervention teacher on multiple occasions during the year. The AIU staff ratings were based on observations of specific class sessions, while the trainers' ratings were based on impressions formed over the course of extended interactions with the intervention teachers. In addition, each intervention teacher was videotaped twice, with the videotapes used to assess teacher quality as well as to detail the amount of time, on average, that each of the four interventions spent on various reading activities. Finally, intervention teachers kept a log of the training they received throughout the school year. These data sources are described further in Chapter III.

EXHIBIT 1. STUDENT PERFORMANCE MEASURES

READING MEASURES

Phonemic Decoding

- **Word Attack** subtest from the Woodcock Reading Mastery Test-Revised (WRMT-R; Woodcock 1998) requires students to pronounce printed nonwords that are spelled according to conventional English spelling patterns.

- **Phonemic Decoding Efficiency** subtest from the Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, and Rashotte 1999) requires students to pronounce nonwords from a list of increasing difficulty as fast as they can. The score is the number of words correctly pronounced within 45 seconds.

Word Reading Accuracy and Fluency

- **Word Identification** subtest from the WRMT-R requires students to pronounce real words from a list of increasing difficulty. The child's score is the total number of words read correctly before reaching a ceiling, which is determined when the child makes a specific number of errors in a row.

- **Sight Word Efficiency** subtest from the TOWRE requires students to pronounce real words from a list of increasing difficulty as fast as they can. The score is the number of words correctly pronounced within 45 seconds.

- **Oral Reading Fluency** subtest from Edformation, Inc., (Howe and Shinn, 2002) requires students to read three passages at their grade level (third or fifth); their score is the median number of correct words per minute for the three passages. The text of this report refers to these passages as Aimsweb passages, which is the term used broadly in the reading practice community.

Reading Comprehension

- **Passage Comprehension** subtest from the WRMT-R requires students to read short passages that contain a blank substituted for one of the words. The task is to use the context of the passage to determine what word should fill the blank. The subtest uses the cloze procedure for estimating reading comprehension ability. This measure of reading comprehension has been widely used in other intervention research with older students, so it provides one basis for comparing results from this study with those from earlier research.

- **Passage Comprehension** subtest from the Group Reading Assessment and Diagnostic Evaluation (GRADE; Williams 2001) requires students to read short passages and answer multiple-choice questions. The present study used this test because it relies on a method for assessing reading comprehension that is similar to methods widely used in the United States for state level accountability testing. It is administered in a group setting and requires students to read passages and answer questions independently. Despite a time limit, most students are able to complete all of the items.

SPELLING AND MATHEMATICS CALCULATION ABILITY MEASURES

- **Spelling** subtest from the Woodcock-Johnson III Tests of Achievement (WJIII; Woodcock, McGrew, and Mather 2001) requires students to spell words that are dictated to them

- **Calculation** subtest from the WJIII requires students to perform mathematical calculations of increasing difficulty until they miss a certain number of problems in a row.

LANGUAGE MEASURES

- **Peabody Picture Vocabulary Test, Third Edition** (PPVT-III; Dunn and Dunn 1997) is a measure of receptive vocabulary in which the subject is required to select a picture that best depicts the verbal stimulus given by the examiner.

- Subtests from the **Comprehensive Test of Phonological Processes** (CTOPP; Wagner, Torgesen, and Rashotte 1999)

  - **Blending subtest**. Measures a student's ability to blend together separate phonemes to form words.

  - **Elision subtest**. Measures a student's ability to manipulate the sounds in orally presented words. For example, the student might be asked to indicate the word that is made when the word *split* is pronounced without saying the phoneme /l/.

  - **Rapid naming for letters/numbers**. Each subtest requires the student to name a matrix of six letters/numbers each randomly repeated six times, for a total of 36 items. The child's score is the time required to name all the items. The test is administered twice, and the student's score is the average of the two administrations.

- Subtests from the **Rapid Automatized Naming and Rapid Alternating Stimulus Tests** (RAN/RAS; Wolf and Denkla 2005.)

- **Rapid Automatized Naming.** Each subtest requires the student to name 5 high-frequency items randomly repeated 10 times in an array of 5 rows for a total of 50 stimulus items. Each row of 10 items contains two examples of each of the 5 items. The student's score is the time required to name all the items.

  - Colors—each item is a color
  - Objects—each item is an object
  - Numbers—each item is a number
  - Letters—each item is a letter

- **Rapid Alternating Stimulus**—each subtest requires the student to name items from the previous subtests that are randomly repeated 10 times in an array of 5 rows for a total of 50 stimulus items. The student's score is the time required to name all of the items.

  - 2-set numbers and letters—each row of 10 items contains one example of each of the 5 numbers and letters used in the subtests above.

  - 3-set colors, numbers, and letters—each row of 10 items contains colors, numbers, and letters used in the subtests above. Each item occurs 3 or 4 times in the array.

- **Sentence Assembly Test from the Clinical Evaluation of Language Fundamentals, Fourth Edition** (Semel, Wiig, and Secord 2003) requires the student to arrange words in a grammatically correct form to make a statement or ask a question.

III. IMPLEMENTATION ANALYSIS

The purpose of this evaluation is to estimate the impact of four reading interventions on student reading achievement, given that each of the interventions was delivered with as much fidelity and skill as could be attained in a standard school setting. Procedures to ensure high quality implementation included careful selection of teachers to deliver the interventions, initial training and on-going supervision of the instructors by the program developers, and the use of a full-time study coordinator whose duties included working with school personnel to facilitate the scheduling of intervention sessions and to minimize student absences. Although these preconditions for successful implementation were established, we also evaluated the quality and fidelity of the instructional implementation. In this way, we could be assured that observed impacts could be attributed to an intervention that was implemented as planned. Overall, the training and supervision produced instruction that was judged to be highly faithful to each intervention model.

This chapter documents in detail the procedures that were undertaken to ensure such high quality implementation, describes the instruction provided to students in the treatment and control groups, and presents the analyses supporting the conclusion that the interventions were implemented with high fidelity. This implementation and fidelity analysis utilized teacher surveys and ratings of intervention group teachers (by both AIU and reading program staff), as well as videotapes of instructional group sessions. The videotapes provide information on the quality of instruction as well as on the amount of time spent by each program on particular reading activities, thus allowing an exploration of the similarities and differences in reading instruction offered in the four interventions.

A. INSTRUCTION PROVIDED TO STUDENTS IN THE TREATMENT GROUP

The following three criteria informed the selection of interventions evaluated in this study: (1) the extent to which program providers had the capability to provide the teacher training and supervision required by the study design; (2) the extent of existing evidence of the method's effectiveness in remediating reading difficulties in older children; and (3) the "fit" of the instructional methods within the two instructional contrasts.

We circulated a request for applications to all known program providers with the capacity to participate in the study and, in return, received 12 applications. Nine applications characterized themselves as word level plus comprehension interventions (WL+C) and 3 as word level (WL) interventions. Two members of the study's scientific advisory board rated the quality of the research evidence available establishing the efficacy of each of the instructional programs, and the methods were then ranked by their scores on this dimension. With too few qualified applicants in the WL instructional category, the advisory board invited one of the highly qualified applications in the WL+C category to submit the word-level component of its program under the WL category. One of the applicants in the WL+C category who was initially invited to participate had to decline because of other commitments during the study's time frame. One initial difficulty that became apparent early in the selection process was that the remaining two highest-rated WL+C interventions used substantially different methods to teach word-level reading skills. However, given that this initial difference did not violate the basic premise of the instructional category, we included both methods in the WL+C category. The interventions within each intervention category were as follows:

Word Level Plus Comprehension Interventions. The two interventions in the WL+C category were Spell Read Phonological Auditory Training (Spell Read P.A.T.; MacPhee, 1990) and Failure Free Reading (Lockavitch 1996).

Word-Level Interventions. The two interventions in the word-level category were Corrective Reading (Engelmann, Carnine, & Johnson, 1999; Engelmann, Meyer, Carnine, Becker, Eisele, & Johson, 1999; Engelmann, Meyer, Johnson, & Carnine, 1999) and the Wilson Reading System, Third Edition (Wilson 2002). It is important to note that complete versions of both of these interventions contain instructional routines and materials that focus directly on comprehension and vocabulary, but, for purposes of this study, the program providers agreed to focus exclusively on word-level skills.

Below, we briefly describe the four interventions.

Spell Read Phonological Auditory Training (P.A.T.) provides systematic and explicit fluency-oriented instruction in phonemic awareness and phonics along with everyday experiences in reading and writing for meaning. The phonemic activities involve a wide variety of specific tasks based on specific skill mastery, including, for example, building syllables from single sounds, blending consonant sounds with vowel sounds, and analyzing or breaking syllables into their individual sounds. Each lesson also includes language-rich reading and writing activities intended to ensure that students use their language in combination with phonologically based reading skills when reading and writing.

The program consists of 140 sequential lessons divided into three phases. The lesson sequence begins by teaching the sounds that are easiest to hear and manipulate and then progresses to the more difficult sounds and combinations. More specifically, Phase A introduces the primary spelling of 18 vowels and 26 consonants and the consonant-vowel, vowel-consonant, and consonant-vowel-consonant patterns. The goals of Phase B are to teach the secondary spellings of sounds and consonant blends and to bring students to fluency at the two-syllable level. In Phase C, students learn beginning and ending clusters and work toward mastery of multisyllabic words. A part of every lesson involves "shared reading" of leveled trade books and discussion of content. Students also spend time at the end of every lesson writing in response to what they read that day. All groups began with the first lesson but then progressed at a pace commensurate with their ability to master the material. By the end of the intervention period, the students receiving Spell Read instruction had reached points ranging from the end of phase A to the initial lessons of level C.

Failure Free Reading uses a combination of computer-based lessons, workbook exercises, and teacher-led instruction to teach sight vocabulary, fluency, and comprehension. Students spend approximately a third of each instructional session working within each of these formats, so that they spend most of their time working independently rather than in a small group. Unlike the other three interventions, Failure Free Reading does not emphasize phonemic decoding strategies. Rather, it builds the student's vocabulary of "sight words" through a program involving several exposures and text that is engineered to support learning of new words. Students read material that is designed to be of interest to their age level while challenging their current independent and instructional reading level. Lessons are based on story text controlled for syntax and semantic content. Each lesson progresses through a cycle of previewing text content and individual word meanings, listening to text read aloud, discussing text context, reading the text content with support, and reviewing the key ideas in the text in worksheet and computer formats. Teachers monitor student success and provide as much repetition and support as students need to read the day's selection.

Although the students are grouped for instruction as in the other three interventions, the lessons in Failure Free Reading are highly individualized, with each student progressing at his or her own pace based on initial placement testing and frequent criterion testing. Two levels of story books are available.

Students who show mastery at the second level progress to a related program called Verbal Master, which uses the same instructional principles but emphasizes vocabulary building and writing activities rather than passage reading. Verbal Master activities include listening to definitions and applications of target vocabulary words and interpreting and constructing sentences containing the target words. The curriculum also provides reinforcement exercises such as sentence completion and fill-in-the-blank activities as well as basic instruction in composition. Most of the third grade students assigned to the Failure Free condition spent all of their instructional time working within the first and second level of story sequences. On the other hand, 65 percent of the fifth grade students spent half or more of their instructional time in Verbal Master.

Corrective Reading uses scripted lessons that are designed to improve the efficiency of "teacher talk" and to maximize opportunities for students to respond to and receive feedback. The lessons involve explicit and systematic instructional sequences that include a series of quick tasks intended to focus students' attention on critical elements for successful word identification. The tasks also include exercises that build rate and fluency through oral reading of stories that have been carefully constructed to counter word-guessing habits. The decoding strand, which was the component of Corrective Reading used in the present study, includes four levels—A, B1, B2, and C. Placement testing is used to start each group at the appropriate level, although, as we will see, the instructional groups in the study were relatively heterogeneous in terms of their beginning skills; therefore, the study did not always permit an optimal match with every child's initial instructional level. The lessons provided during the study clustered in Levels B1 and B2, with some groups progressing to Level C. By the end of B1, the curriculum covers all of the vowels and basic sound combinations in written English, the "silent-e rule," and some double consonant-ending words. Students also learn to separate word endings from many words with a root-plus-suffix structure, to build and decompose compound words, and to identify underlying sounds within written words. Level B2 addresses more irregularly spelled words, sound combinations, difficult consonant blends, and compound words while Level C focuses on strengthening students' ability to read grade-level academic material and naturally occurring text such as that in magazines. Explicit vocabulary instruction is also introduced in Level C, but this component was not provided for those groups that, in fact, reached level C in this program.

The **Wilson Reading System** uses direct, multisensory structured teaching based on the Orton-Gillingham methodology. Based on 10 principles of instruction, the program teaches sounds to automaticity; presents the structure of language in a systematic, cumulative manner; presents concepts within the context of controlled and noncontrolled written text; and teaches and reinforces concepts with visual-auditory-kinesthetic-tactile methods. Each Wilson Reading lesson includes separate sections that emphasize word study, spelling, fluency, and comprehension. Given that Wilson Reading was assigned to the word-level condition in this study, teachers were not trained in the comprehension and vocabulary components of the method, nor were they included in the instructional sessions.

The program includes 12 steps. Steps 1 through 6 establish foundational skills in word reading while Steps 7 through 12 present more complex rules of language, including sound options, spelling rules, and morphological principles. In keeping with the systematic approach to teaching language structure, all students begin with Step 1, but groups are then free to move at a pace commensurate with their skill level. By the end of the intervention period, all students receiving the Wilson Reading intervention had progressed to somewhere between Steps 4 and 6.

B. INSTRUCTION PROVIDED TO STUDENTS IN THE CONTROL GROUP

Students assigned to the control group were to receive the type and amount of intervention instruction they would have received from their schools in the absence of the study. As seen when we report on the

total amount of instruction provided to all groups, the amount of small-group and individualized instruction received by students in the control group was considerable; in fact, it approached the amount provided to the students in each intervention condition. With students in the study spread across 27 school districts, with potentially different reading curricula, the nature of the instruction received by the students in the control group was probably variable. Although we have data on the amount of reading instruction received by each student in the control group, we did not collect data like we did for students in the interventions indicating how that time was distributed across different types of reading activities, such as time building word-level skills versus time developing comprehension skills or vocabulary. This limits our ability to describe the reading instruction received by students in the control group and compare that instruction to the instruction provided to students in the interventions.

C. DELIVERY OF INTERVENTION INSTRUCTION

The study plan called for delivering as close to 100 hours of instruction as possible in 60-minute sessions, five days a week, to groups of three students. After random assignment to the intervention or control group within each school unit, the intervention students were placed in instructional groups according to their classroom schedules. An attempt was also made to match students in the instructional groups as closely as possible on their initial levels of word reading skill so that instruction could be targeted on student needs more effectively, but this was not always possible given the small numbers of students assigned to the interventions at each grade. Each teacher was to teach four groups a day. The actual implementation of instruction differed in several ways from the study's plan. The major deviations pertained to amount of instruction provided, size of instructional groups, and group homogeneity in terms of beginning word-level reading skills. Each of these issues is addressed below.

1. Intensity of Interventions

In planning the study, we recognized that groups occasionally would not be able to meet or would have to cut short their instruction. In fact, occurrences such as school assemblies, snow days, and school closings for other reasons sometimes prevented groups from receiving instruction. In addition, individual students were absent on some days. To offset these unavoidable irregularities, we put into place several strategies as follows:

- First, the intervention groups were scheduled to run for more than 100 days so that, on average, students would accumulate 100 hours of intervention.
- Second, substitute teachers were hired and trained so that groups could meet when the regular teacher was absent.
- Third, the local coordinator worked with classroom teachers and administrators at the participating schools to try to minimize disruptions to the intervention groups.
- Fourth, intervention teachers were asked to conduct make-up sessions for students who missed significant amounts of group time.

A central question of implementation fidelity is whether participants received the intended dose of the intervention. To answer this question, the study asked intervention teachers to maintain attendance logs on which they recorded, for each school day during the implementation period, (1) whether the group

met, (2) which students were present or absent, (3) the number of minutes of instruction for each student, and (4) the number of minutes of make-up instruction for each student, if any.

Using the sample of videotapes collected for the instructional fidelity analysis (18 to 20 videotapes per reading program), we compared total session time recorded on the tape with the minutes of instruction recorded by the intervention teacher on the attendance log. The modal entry for the attendance log was 60 minutes, although some sessions were recorded as shorter or, occasionally, longer. On average, the time recorded on the videotape, from the moment the students entered the room to the moment they were dismissed, was 5.9 minutes shorter than the time recorded on the attendance log. No pattern in the discrepancy was associated with whether the attendance log showed a straight 60 minutes or some other number. Based on the available information, we determined that 5.9 minutes should be subtracted from each log entry in calculating the total hours of intervention for each student.

Table III.1 displays the percentage of students who reached certain benchmarks in total hours of intervention, including students who received at least 80 hours of intervention; students who received at least 40 hours of intervention but fewer than 80; and students who received at fewer than 40 hours of intervention. As can be seen, over 90 percent of students in the treatment group received at least 80 hours of instruction.

Table III.1

Percentage of Students Attaining Different Levels of Intervention Hours

More than 80	92.3
More than 40 but fewer than 80	4.5
Fewer than 40	3.2

When we considered group size, we found that, across the four reading interventions, more than three-quarters of intervention hours were delivered to groups of three students, as intended. Very few hours, on average, were delivered to only one student. We observed no significant differences between interventions with regard to average total hours or average hours by group size (see Appendix K for details).[21] However, we did note one significant difference by grade level, with fifth-grade students receiving fewer (88) total hours of intervention, on average, than third-grade students (93 hours) [$t(399) = 2.88$, $p < .01$].

Finally, we investigated the average hours of instruction delivered by substitute rather than regular teachers for each intervention: Failure Free Reading = 4, Spell Read = 3, Wilson Reading = 6, and Corrective Reading = 6. The hours did not differ significantly between interventions (see Appendix K for details). However, three of the teachers in the Wilson Reading program were permanently replaced by a teacher from the substitute teacher pool for the last two to four weeks of instruction because the regular teachers left on maternity leave. If these "permanent substitute" hours were added to the total hours delivered by substitute teachers, then Wilson Reading would clearly differ from the other interventions in terms of total number of hours delivered by substitutes.

[21] Because this evaluation was not designed to compare the individual interventions with each other, there is relatively low power for conducting a series of tests for pairwise differences between the interventions. We conduct, instead, one test for differences across all four interventions on each variable of interest.

2. Instructional Group Heterogeneity

In providing remedial instruction to older students in word-level reading skills, it is common practice to form instructional groups that are as homogeneous as possible with regard to the basic skills being taught. Clearly, appropriate grouping of students for instruction was of concern to three of the study's four program providers. Corrective Reading, for example, administers a placement test that allows students to be placed in the program at the appropriate point depending on initial skill level. Although both Spell Read and Wilson Reading start at the same point for all students, students progress through the program in accordance with their mastery of skills. If students work at different levels of knowledge and skill, teachers find it difficult to target instruction at the appropriate level for every student.

The study design called for the random selection of six students in grade three and six students in grade five, within each school unit, to participate in the intervention. The remaining students were placed in the control group and received the services they would normally receive in the absence of the intervention. In addition to the approach that we implemented, two other approaches were considered when designing the experiment: (1) do random assignment within strata defined by test scores or (2) use the approach that we implemented, but after selecting six students for the treatment group, sort them into two groups of three based on test scores. We used our approach so that program developers could form groups the way that they normally would given the mix of students who were eligible for an intervention according to the study criteria and selected at random to receive the intervention.

One approach for reducing within-group heterogeneity would have been to impose more stringent eligibility criteria, by, for example, lowering the upper threshold on the word-level screening test from the 30th percentile to the 20th percentile. That, however, would have substantially reduced the size of the evaluation sample and the power to detect impacts. Another approach to reducing heterogeneity would have been to implement the evaluation in schools with many more eligible students and create at least several instructional groups in each school—an approach that was largely infeasible in the AIU. Given the relatively small number of students selected for the intervention and the range of students identified through the eligibility screening process, program developers may have had to create groups with more heterogeneity than they would have if they were working with larger numbers of students. However, in follow-up conversations, the program developers indicated that the extent of within-group heterogeneity that existed within this study was not unusual in comparison with what they normally confront when delivering their interventions in other settings.

Table III.2 shows the average range between the highest and lowest scores on the baseline Word Attack measure for the instructional groups in each condition. There were no significant differences in the heterogeneity of the groups across methods or grades. On average, the range of scores within the instructional groups on the beginning measure of phonemic decoding skill was almost a full standard deviation.[22]

[22] Appendix N provides information on an analysis done to assess the effects of instructional group heterogeneity on students' reading outcomes. No consistent pattern in the relationship between instructional group heterogeneity and reading outcomes was found.

Table III.2

Mean Range of Baseline Word Attack Scores within Instructional Groups

	Failure Free Reading	Spell Read	Wilson Reading	Corrective Reading
Third Grade				
Mean	14.3	13.5	13.1	13.2
Standard deviation	9.3	7.0	6.6	7.3
N	56	57	51	48
Fifth Grade				
Mean	15.8	12.4	17.4	14.2
Standard deviation	8.5	9.6	9.8	6.9
N	60	60	54	59

D. SELECTION, TRAINING, AND SUPPORT OF TEACHERS

1. Teacher Selection

We selected the intervention teachers from the schools that agreed to participate in the study. The principal of each school sought volunteers and then nominated two or three teachers to be interviewed by the research coordinator. We then used four criteria to select intervention teachers from among potential participants: (1) experience and interest in providing the type of intensive instruction examined in the study; (2) willingness to be randomly assigned to one of four intervention methods, one of which would be highly scripted; (3) personality and capability as assessed informally by the interviewer; and (4) scores on tests of phonemic awareness and phonemic decoding fluency. The second criterion required careful explanation as some teachers object strongly to working within a scripted curricula. The fourth criterion was essential because three of the four interventions involved explicit instruction in phonics; moreover, two of the program providers (Spell Read and Wilson Reading) indicated that teachers who struggle with "phonics" have a more difficult time gaining proficiency in the delivery of instruction within their programs. As part of their interview, the teachers agreed to take the Elision subtest from the CTOPP and the Phonemic Decoding Efficiency subtest from the TOWRE.

Our goal was to hire 44 teachers (10 for each intervention plus 4 substitutes). Because of difficulties at two of the schools originally recruited into the study, Wilson Reading and Corrective Reading ended up with 9 rather than 10 teachers regularly leading instructional groups.[23] For the 38 teachers eventually recruited into the study (excluding substitutes), Table III.3 shows the average years of teacher experience, by intervention.

The teachers in the Failure Free program had significantly more years of teaching experience than those delivering the Wilson Reading program [Tukey's HSD (Alpha: .05, Error: 34) = 75.58].[24] Another way to

[23] A tenth Corrective Reading teacher was trained and delivered instruction, but, with no control students at the school to which she was assigned, her students were not included in the analyses for the study.

[24] Although we can not provide the details for each program for confidentiality reasons, there were no significant differences across programs in terms of the highest degree obtained by teachers [$X^2(6, N=38) = 10.09$, $p=.12$].

Table III.3

Average Years of Teaching Experience, by Intervention

	Failure Free Reading	Spell Read	Wilson Reading	Corrective Reading
Average Years of Teaching Experience	20	11.1	8.9	15.3

look at teacher training is to consider the area of certification. The most common certifications were no systematic associations between type of certification and instructional program [$X^2(12, N=38) = 10.05$, $p=.61$].

Table III.4 reports the raw scores for teachers in each condition on the measures of phonemic awareness and phonemic decoding efficiency. The groups were not significantly different with regards to either measure (phonemic awareness [$F(3,34 = 0.72$, $p = 0.5447$]; phonemic decoding efficiency [$F(3,34) = 2.80$, $p = .0549$]).

Although the age of the teachers in this study fell outside the range of the standardization sample for both of these tests, it is possible to provide some perspective on the above scores by comparing them to the normative performance of the oldest group (20 year olds) from that sample. Compared to this group, the average standard score of our intervention teachers on the Elision subtest was 105, with a range from 90 to 110. The average standard score on Phonemic Decoding Efficiency was 97, with a range from 79 to 120. The average standard score on the latter measure for each instructional condition was Corrective Reading = 106, Spell Read = 100, and Wilson Reading and Failure Free Reading = 93. Thus, almost all of the teachers fell within the average range on these measures of phonemic awareness and phonemic decoding fluency, but a few in several of the conditions performed substantially below average for adults.

2. Teacher Training and Support

Representatives of the four reading programs used in the interventions trained the intervention teachers. Initial training was provided in a week-long session before school began. Following this initial training, teachers practiced delivering the interventions for about seven weeks with groups of fourth grade students from participating schools. During this practice period, trainers provided weekly training and observation contacts with the teachers. During the implementation phase with third and fifth graders, program providers made at least monthly follow-up visits with the teachers. Providers could, however, increase their follow-up support at their discretion in order to model more closely the typical support given to teachers involved in their programs. In fact, all four interventions chose to increase their support such that each teacher received an average 38.3 hours of professional development during approximately nine months of the practice and implementation period, with nearly 24 of the hours concentrated in the six- to eight-week practice period.

The initial training was conducted over five days. All of the teachers (including substitutes) convened in one setting but spent most of the training time working with trainers from the specific reading intervention to which they were assigned. During the week, a few training hours were devoted to explaining the purposes of the study and the logistics of student selection, formation of reading groups, student assessments, and record keeping. We estimate that, on average, teachers received training related to the delivery of their reading interventions for about 6.5 hours per day, or 32.5 hours for five days.

Table III.4

Raw Scores for Teachers on Measures of Phonemic Decoding Efficiency and Phonemic Awareness

	Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Metric	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Phoneme Elision	19.3	0.82	18.9	1.10	18.8	1.20	18.7	0.82
Phonemic Decoding Efficiency	50.2	6.99	55.5	6.98	50.3	8.56	57.9	4.84

In general, the August training was structured to allow the teachers to experience their program from the perspective of a student. The teachers gradually took on more of the teaching behaviors as they practiced with their peers and with the trainer. The providers of Wilson Reading and Spell Read, which include detailed phonemic training as part of their programs, spent proportionally more time shaping teachers' skills in recognizing and reproducing phonemic patterns. The provider of Failure Free Reading spent nearly all of the initial training week on the computer-based aspect of the program. The provider of Corrective Reading, which contains a substantial number of specific, scripted teaching routines, worked with teachers to help them master the small-group instruction routines and gain familiarity with lesson formats.

For several reasons, there was some variation in the modes and amount of training that teachers received during the study. Several teachers attended only some of the initial five days of training or missed them altogether because of either personal circumstances or the fact that they had not been hired when the training took place. Trainers returned to deliver make-up training in late August and early September.

Another source of variation was differences in the amount and type of follow-up support that programs typically provided to teachers. While the study team agreed that the interventions could follow their typical practice after the initial training, the study team put in place procedures for documenting follow-up training and coaching activities. In this way, we were able to analyze and report on differences in training/coaching activities, something of potential interest if decision makers consider adopting the interventions in the future.

To document the amount and type of professional development that teachers received subsequent to the initial August training, both teachers and trainers maintained logs of their training-related activities. The forms provided space to record the date and duration of each activity and the number of participating teachers and trainers. In addition, the logs provided a series of check-off boxes to characterize the type of activity. Professional development activities through which teachers received guidance or support from the reading program providers included the following:

- Group instruction delivered by a reading program trainer (a meeting of all or most teachers delivering a particular reading program, during which the trainer presented new material and/or teachers discussed issues that had arisen as they worked with students)
- Coaching (the trainer worked with teachers in a classroom setting either individually or with other teachers observing)
- Telephone consultations of at least five minutes' duration focusing on instructional issues

- Independent study (Wilson Reading teachers were encouraged to work through a self-paced online course that reinforced information provided during the August training; other teachers reviewed training materials or pursued additional reading on their own)

Comparisons between activities recorded in the teacher and trainer logs revealed some inconsistencies in individuals' reports of the occurrence and duration of specific training events. Consequently, we based our estimates of the total hours of support provided to teachers on the combined reports of teachers and their trainers. (When a teacher and trainer reported different durations for a single event, we used the average of the two reports as the event duration.) In view of the incompleteness of the data submitted by individual teachers and trainers and the differences in details provided for the same events by different reporters, the following summary should be understood as an approximation rather than as a precise accounting of the professional development in which teachers participated.

Table III.5 summarizes the average hours of instruction and support that teachers received during the initial training, practice, and implementation phases of the study. The phases are defined as follows:

- The training phase, including the intensive training received before school began and make-up training provided in August and September
- The practice phase (with fourth-grade students between September and the beginning of the implementation phase)
- The implementation phase (with third- and fifth-grade students who were the subjects of the experimental study, beginning in November)

On average, teachers received almost 69 hours of professional development during the study—over 30 hours during the intensive training phase, 24 during the practice phase, and 14 during the implementation phase. With training, coaching, independent study, and telephone consultations considered together, we observed statistically significant differences between programs in the number of hours of professional development received by teachers during the implementation phase [$F(3,34) = 22.66$, $p<.0001$] and overall at the .05 level [$F(3,34) = 3.92$, $p =.0165$], but not during the intensive training or practice phases.

The interventions also varied somewhat in the mix of supports each provided (see Tables III.6 through III.8). The vast majority of professional development hours (64) took the form of training or coaching. However, the two interventions for which fewer training and coaching hours were reported, Wilson Reading and Failure Free Reading, used additional methods to support their teachers. Wilson Reading augmented its face-to-face training and coaching with an online course that included video clips of Wilson Reading training sessions, comments on the content covered in each part of the curriculum, and demonstrations of instructional techniques. Wilson Reading teachers reported that they spent an average 11 hours in independent study in contrast to teachers in the other interventions, who averaged about 20 minutes of independent study during the year. Follow-up support for Failure Free Reading teachers included periodic voluntary telephone conferences with program providers. Failure Free Reading teachers reported that they participated in about 5.9 hours of telephone conferencing over the year in contrast to teachers in other interventions, who averaged about 25 minutes of telephone conversations.

Table III.5

Average Hours of Professional Development Received by Teachers, by Intervention[a]

	All Interventions (N = 38)	Failure Free Reading (N = 10)	Spell Read (N = 10)	Wilson Reading (N = 9)	Corrective Reading (N = 9)	
Intensive training phase	30.5	29.6	30.1	29.4	32.8	
Practice phase	23.9	25.2	24.9	18.9	26.4	
Implementation phase	14.4	8.7	23.1	14.2	11.6	*
Overall	68.8	63.5	78.1	62.5	70.8	*

[a]Professional development includes training and coaching by reading program staff, independent study of program materials, and telephone conferences.
* Overall difference between groups is statistically significant at the 0.05 level.

Table III.6

Average Hours of Training and Coaching Received by Teachers from Reading Program Staff

	All Interventions (N = 38)	Failure Free Reading (N = 10)	Spell Read (N = 10)	Wilson Reading (N = 9)	Corrective Reading (N = 9)	
Intensive training phase	30.5	29.6	30.1	29.4	32.8	
Practice phase	21.0	21.3	24.6	11.4	26.2	*
Implementation phase	12.6	6.1	22.6	9.9	11.5	*
Overall	64.1	57.0	77.3	50.8	70.6	*

* Overall difference between groups is statistically significant at the 0.05 level.

Table III.7

Average Hours of Independent Study Reported by Teachers

	All Interventions (N = 38)	Failure Free Reading (N = 10)	Spell Read (N = 10)	Wilson Reading (N = 9)	Corrective Reading (N = 9)	
Practice phase	2.0	0.6	0.2	7.5	0.1	
Implementation phase	0.9	0.0	0.0	3.7	0.0	*
Overall	2.9	0.6	0.2	11.2	0.1	*

* Overall difference between groups is statistically significant at the 0.05 level.

Table III.8

Average Hours of Telephone Consultations Reported by Teachers

	All Interventions (N = 38)	Failure Free Reading (N = 10)	Spell Read (N = 10)	Wilson Reading (N = 9)	Corrective Reading (N = 9)	
Practice phase	0.9	3.3	0.09	0.05	0.07	*
Implementation phase	0.9	2.6	0.50	0.49	0.04	*
Overall	1.8	5.9	0.58	0.54	0.11	*

* Overall difference between groups is statistically significant at the 0.05 level.

In summary, over the course of the study, the reading program providers delivered nearly 70 hours of training and professional development to intervention teachers. The total amount of professional development and the amount of face-to-face coaching and instruction offered by the various programs differed significantly from intervention to intervention. However, all the program providers agreed that the amount of training and professional development equaled or exceeded what they would typically deliver to new teachers in a school setting.

In addition to the support provided by the program providers, the study coordinators from the AIU assisted teachers in dealing with issues related to scheduling instructional sessions, obtaining permission forms from parents, rescheduling instructional sessions, and behavior management that arose in the course of instruction.

E. TEACHER QUALITY AND FIDELITY OF INSTRUCTIONAL IMPLEMENTATION

The study evaluated the performance of the intervention teachers along two dimensions: (1) the fidelity with which they implemented the specific requirements of the reading program to which they were assigned and (2) the extent to which they exhibited more general behaviors, such as good organization, that are consistent with good-quality teaching.

Two sources of data contributed to the fidelity evaluation while a third source was available for the evaluation of general teacher quality. For the fidelity evaluation, we obtained two rounds of ratings from the reading program trainers and coded two videotapes of each teacher. For the more general teacher quality evaluation, we used data from these same two sources and obtained ratings for an average of three sessions per teacher observed by the AIU coordinators. The value of the videotape analysis was that it allowed for an independent and fine-grained analysis of instructor behavior. However, resource constraints dictated that such an analysis could cover only a small sample of the instructors' total performance. Moreover, there were significant aspects of the program implementations that did not lend themselves to evaluation through this type of time-sampling methodology. In particular, all of the programs had some expectation that instructors would pace the instruction and individualize the intervention in relation to each student's progress, and this is not readily observed in an analysis of a single instructional session. (The extent to which instructors were expected to tailor the instruction varied from program to program, however, with Corrective Reading making the fewest demands in this respect and Wilson Reading making the greatest.)

The ratings by the program providers, who worked with the instructors on an ongoing basis, offered the opportunity to capture this missing information on pacing, as well as other aspects of instructor

performance. In addition, the providers were clearly expert in the fidelity requirements of their specific programs, so their ratings could not be criticized for missing critical aspects of instructor behaviors. On the other hand, however, the providers had a stake in the outcomes of the study and thus could not be classed as independent observers. To balance concerns about the provider's stakeholder status, all ratings of the fidelity of the intervention were collected before the providers were given any information about the impact on student performance. In fact, information on student outcomes was also withheld from the study staff responsible for the fidelity analysis until after that analysis was complete.

All of the teacher quality and fidelity evaluations focus on the regular teachers, not on the substitutes. As shown in the section on hours of intervention, the regular teachers delivered a high percentage of the total intervention hours. The following discussion considers the two types of rating data and the videotape analysis.

1. Trainer Ratings of Fidelity and Teacher Quality

Trainers rated teachers twice: in the fall (at the end of the practice period) and in the spring (near the end of the intervention period). The trainers provided two types of ratings: (1) a global estimate of how a teacher's performance compared with the performance of all teachers with similar amounts of training and teaching experience that the trainer had ever observed, and (2) ratings on eight dimensions of the teacher's delivery of the program. The first five dimensions specifically address intervention fidelity while the remainder deal with general teacher quality.

Table III.9 shows the average global ratings assigned by each program, based on a six-point scale that locates the teacher within percentile ranges (1 = lowest 10 percent, 2 = lowest quarter but not lowest 10 percent, 3 = lower half but not lowest quarter, and so on). The table shows that, on average and despite significant differences among programs, trainers judged teachers to fall somewhere in the top half among similarly experienced teachers whom they had observed. In the fall, the average ratings earned by the Spell Read teachers were lower and significantly different [Tukey's HSD (Alpha: .05, Error: 34) = 1.006] than the ratings earned by the Failure Free Reading or Corrective Reading teachers. (In the spring, the ratings of the Wilson Reading teachers were significantly lower than those of the Corrective Reading teachers [Tukey's HSD (Alpha: .05, Error: 34) = 1.90]. However, given that trainers rated only those teachers trained in their given intervention, it is not possible to determine the extent to which the observed differences across programs may reflect rater bias rather than actual differences in teacher quality.

Table III.10 summarizes the ratings on eight dimensions of program delivery. The ratings used a seven-point scale ranging from 1 = unsatisfactory performance through 3 = satisfactory performance to 7 = expert performance. The average ratings on all eight dimensions in both fall and spring generally ranged from about 4.0 to 6.8—well above the satisfactory (3) level. We thus see that the program providers did not have any serious reservations about the quality and fidelity of the instruction delivered in this study.

2. Ratings of Instructional Sessions by AIU Staff

AIU staff observed each intervention teacher about three times during the year, at roughly two-month intervals. Observations lasted for approximately a half hour, with the teachers' performance during the period rated on seven dimensions in accordance with a three-point scale (1 = significant problems, 2 = minor problems, 3 = satisfactory performance). We used the sum of the ratings to construct an overall session rating as well. The range for the summary scale was 7 to 21, although no session received a summary score lower than 13.

Table III.9

Trainers' Global Ratings of Program Implementation

	Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Global implementation rating (1–6 scale)	Mean	N	Mean	N	Mean	N	Mean	N
Fall 2003	6.00	10	4.00	10	5.00	9	5.56	9
Spring 2004	5.30	10	4.10	10	3.61	9	5.67	9

Overall, the ratings suggest that, on average, AIU staff saw consistent instruction and classroom management during their visits to the instructional groups. Average session ratings for the four programs ranged from 19.6 to 20.6 (see Table III.11), and were not significantly different at the .05 level [$F(3,107) = 2.38$, $p = .0732$]. All of the average dimensional ratings were at least 2.5 points, and most were over 2.8. Variations among programs were significant at the .05 level in only two instances: mean ratings for the Wilson Reading sessions were lower than those for other programs on the teachers' management of student behavior [$F(3,107) = .0042$, $p = .0042$] and on the provision of feedback in a positive manner [$F(3,107) = 7.06$, $p = .0002$]. (Wilson Reading ratings on these two dimensions were 2.5 and 2.7, respectively.)

3. Videotape Analysis

The intervention period provided opportunities to complete two videotaped observations of each intervention teacher, one videotape of a third-grade instructional group and the other of a fifth-grade instructional group. A total of 38 teachers were videotaped, 9 each from Corrective Reading and Wilson Reading and 10 each from Spell Read and Failure Free Reading. Each videotape covered an entire instructional session. The study made every effort to complete the first videotaping of each teacher during the first half of the intervention period and the second during the second half, although the logistics of developing a workable videotaping schedule sometimes necessitated a shorter-than-desired period between the two sessions.

Trained coders analyzed the videotapes with respect to the core instructional elements of each of the four interventions and the manner in which the elements might be expected to interact in order to achieve desired outcomes. The output of the analysis took the form of a "running record" for each videotape. This running record provided a running summary of the activities taking place in the classroom, on a minute-by-minute basis, and was the basis for both the ratings of fidelity/general teacher quality and the time-by-activity analysis discussed later in this section. Appendix I presents the coding procedures used to analyze the videotapes.

The study hired and trained seven coders, all educators with experience teaching reading in the primary grades, to assist with the construction of the running records. The coders analyzed each of the 76 recorded sessions, and a sample of 18 sessions, distributed across the four reading programs, was reanalyzed by a second coder who constructed a second running record.

The videotape analysis of the interventions followed a general procedure for all four interventions but also focused on various features specific to each intervention. Coders noted the beginning and ending times for each activity within a session and were directed to note and time stamp significant events,

Table III.10

Trainers' Ratings of Eight Dimensions of Program Implementation

Rating Dimensions	Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
	Mean	N	Mean	N	Mean	N	Mean	N
Fall 2003 Ratings								
1. Lessons include all prescribed program elements, appropriate sequence, and time frame	3.60	10	4.90	10	4.22	9	5.78	9
2. Mastery of program techniques, materials, and technology	3.10	10	4.50	10	4.44	9	5.67	9
3. Program's prompting, correction, and questioning strategies used	3.30	10	4.60	10	4.78	9	6.00	9
4. Effective lesson delivery, attention to pacing and transitions	3.30	10	4.90	10	4.78	9	5.89	9
5. Lesson plans and program record keeping completed	3.70	10	4.90	10	5.00	8	6.11	9
6. Student performance monitored, attention divided equally among students	3.50	10	4.50	10	5.00	9	6.00	9
7. Intervention as necessary to maintain students' attention and appropriate behavior	3.90	10	4.80	10	5.00	9	5.89	9
8. Good rapport and use of positive reinforcement	3.90	10	5.40	10	5.11	9	6.11	9
Spring 2004 Ratings								
1. Lessons include all prescribed program elements, appropriate sequence, and time frame	5.50	10	5.20	10	5.00	9	6.56	9
2. Mastery of program techniques, materials, and technology	5.50	10	5.00	10	5.33	9	6.44	9
3. Program's prompting, correction, and questioning strategies used	5.90	10	4.90	10	5.56	9	6.56	9
4. Effective lesson delivery, attention to pacing and transitions	5.70	10	5.00	10	5.00	9	6.44	9
5. Lesson plans and program record keeping completed	6.10	10	5.10	10	5.56	9	6.78	9
6. Student performance monitored, attention divided equally among students	6.00	10	4.60	10	5.11	9	6.44	9
7. Intervention as necessary to maintain students' attention and appropriate behavior	6.80	10	4.80	10	5.33	9	6.11	9
8. Good rapport and use of positive reinforcement	6.70	10	5.10	10	5.56	9	6.44	9

Table III.11

AIU Staff Ratings of General Teacher Quality

	Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Rating Dimension	Mean	N	Mean	N	Mean	N	Mean	N
1. Managed time appropriately	2.90	29	2.83	30	2.88	25	2.81	27
2. Was well prepared	2.86	29	2.97	29	2.88	25	2.93	27
3. Followed effective instructional procedures	2.97	29	2.90	30	2.92	25	2.85	27
4. Managed student behavior effectively	2.93	29	2.73	30	2.52	25	2.85	27
5. Monitored student behavior effectively	2.93	29	2.97	30	2.92	25	2.85	27
6. Provided feedback in a positive manner	2.97	29	2.93	30	2.68	25	3.00	27
7. Had good rapport with students	3.00	29	2.93	30	2.80	25	2.93	27
Overall session rating	20.55	29	20.27	30	19.60	25	20.22	27

N = number of sessions rated.

depending on the intervention, that occurred within each activity. In Corrective Reading sessions, for example, coders made note of the teacher's use of correcting procedures while, in Spell Read sessions, coders noted the teacher's monitoring of hand motions. Coders noted the extent to which teachers "wove" previously learned concepts into new instruction in Wilson Reading sessions. As a more individualized program, Failure Free Reading required separate analysis of the instructional experiences of each student, with the most attention devoted to capturing teacher-student interactions and somewhat less attention directed to noting time either on the computer or engaged in individual written work.

Coders wrote brief notes describing types of motivators (e.g., candy, stickers, bonus points, and so forth.), evidence of homework, the nature of the instructional space (e.g., size of room, noise level, and so forth), and their impressions of the affective environment of the lesson. In addition, coders filled out a sheet that summarized key components of the observation. Although some components addressed by the summary sheets were intervention-specific, all addressed teacher organization and preparation, classroom management, and positive reinforcement and praise. Program providers reviewed the coding conventions for the analysis of each intervention and modified them before use by the coders.

After completion of the running records, two study staff members undertook the fidelity/teacher quality analysis by using a set of dimensions that were as comparable as possible across programs. The dimensions included (1) coverage of program content, (2) use of program techniques, (3) management of instruction, (4) appropriate use of positive reinforcement, (5) general affective environment, and (6) total teaching time. In addition, appropriate allocation of time across session components was a factor for every reading program except Corrective Reading. (The highly constrained session script used in Corrective Reading ensures an appropriate allocation of time across components.)

In some cases, the dimensions required further refinement in order to capture potential differences in the teacher's fidelity across disparate program elements. For example, in Spell Read, content coverage, time allocation, and technique needed to be rated separately for the phonemic portion of the lesson and for the story reading portion of the lesson.

The two study staff members coded each dimension on a three-point scale. A code of 3 indicated that performance on that dimension met criterion. (Meeting criterion did not necessarily signify that performance was highly expert but rather that it was faithful to the basic requirements of the program.)

A code of 2 indicated minor deviations from the criterion, and a code of 1 indicated moderate deviations. There were no instances of extreme deviations.

The specific coding systems were submitted to the reading program providers for comment and approval. All of the providers were satisfied that the specified dimensions and criteria would capture fidelity within the context of a single session. However, the study staff and program providers agreed that some important features of program implementation did not lend themselves to evaluation in the context of a single session. For example, the session analysis was not suited to evaluating the extent to which teachers were able to judge the specific strengths and weaknesses of individual students over time and thus adjust the pacing or choice of discretionary exercises accordingly. In Wilson Reading, in particular, which accords teachers considerable latitude in constructing sessions out of a variety of available lesson materials, appropriate session planning is an important skill.

The same two study staff members rated each running record. In the case of more than one running record for the same videotape, they rated each running record separately. The Kappa statistics for inter-rater reliability—across raters and across ratings made from different running records—were: Corrective Reading = .89, Spell Read = .80, Wilson Reading = .90, and Failure Free Reading = .84. These levels of agreement were high, but not unexpected given that the two raters had both been involved in the development of the rating scheme and had detailed discussions about the kinds of evidence that would be used to support the ratings before they began.

Tables III.12 through III.15 present the average ratings on the fidelity dimensions coded for each program. As seen in Table III.12, average scores were above 2.75 on most dimensions, indicating that most Corrective Reading sessions met criterion on these dimensions. However, average scores were lower for proper use of program techniques and total teaching time. With respect to program techniques, the problems reflect the fact that Corrective Reading operates with a highly prescriptive formula for student corrections; many teachers did not strictly adhere to that formula. (Other shortcomings in technique were also observed, but the infractions affecting the correction routine were the most common.) With respect to total teaching time, criterion was set at 55 minutes or more time. Even though program providers and project staff generally agreed that 55 minutes or more was an appropriate criterion, a high proportion of sessions in all programs failed to meet the criterion. In the case of Corrective Reading, most sessions were between 45 and 55 minutes in length, which resulted in ratings of "minor problems" on the total teaching time dimension.

Table III.13 shows that, for Spell Read, average scores were 2.50 or higher on most dimensions. The exceptions were coverage of lesson content—reading and writing (2.37), proper use of program techniques—reading and writing (2.47), and total teaching time (2.35).

Table III.14 presents the Wilson Reading ratings. Given the program's greater variability in session structure (different activities occur on different days), the average ratings for some dimensions are based on fewer than 18 sessions. However we once again see that most dimensions have average scores above 2.50. As with Spell Read, the lower-rated dimensions are concentrated in the areas of passage reading Wilson Reading than for Corrective Reading or Spell Read (although not more pronounced than for and total teaching time. In fact, deficiencies with regard to total teaching time were more pronounced for Failure Free Reading, as discussed below). Of the 17 Wilson Reading sessions evaluated for total teaching time, only 3 sessions met the 55-minute criterion, 8 sessions lasted between 45 and 55 minutes and demonstrated minor time criterion problems, and 6 sessions had moderate problems such that total session length was less than 45 minutes. One Wilson Reading session could not be rated on the time dimension because the videotape stopped before the session concluded.

Table III.12

Scores on Fidelity Dimensions Coded from Videotapes: Corrective Reading

	Average Score[a]
Coverage of lesson content	2.78
Proper use of program techniques	1.83
Management of instruction	2.94
Positive reinforcement	2.89
Affective environment	2.83
Total teaching time	2.22

[a] Scale: 3=meets criterion; 2=minor problems; 1=moderate problems

Table III.13

Scores on Fidelity Dimensions Coded from Videotapes: Spell Read P.A.T.

	Average Score[a]
Coverage of lesson content—phonics	2.60
Duration of lesson content—phonics	2.90
Coverage of lesson content—reading and writing	2.37
Duration of lesson content—reading and writing	2.50
Proper use of program techniques—phonics	2.50
Proper use of program techniques—reading and writing	2.47
Management of instruction	2.85
Positive reinforcement	2.90
Affective environment	2.85
Total teaching time	2.35

[a] Scale: 3=meets criterion; 2=minor problems; 1=moderate problems

Table III.14

Scores on Fidelity Dimensions Coded from Videotapes: Wilson Reading

	Average Score[a]
Coverage of lesson content—decoding	2.78
Duration of lesson content—decoding	2.50
Coverage of lesson content—encoding	2.88
Duration of lesson content—encoding	2.87
Coverage of lesson content—passage reading	2.69
Duration of lesson content—passage reading	2.43
Proper use of program techniques—decoding and encoding	2.56
Proper use of program techniques—passage reading	2.46
Management of instruction	2.78
Positive reinforcement	2.56
Affective environment	2.72
Total teaching time	1.82

[a] Scale: 3=meets criterion; 2=minor problems; 1=moderate problems

Finally, Table III.15 provides the ratings for Failure Free Reading. Even more than with the other programs, Failure Free Reading exhibited deficiencies in adherence to the criterion for total teaching time. Only 2 of the 20 videotaped sessions met the criterion of a 55-minute session, and 6 received a rating of "moderate problems" on the time dimension, resulting in an average score of 1.80 on this dimension. The three dimensions that measured the allocation of time across teaching modalities (teacher-directed, independent student, and computer activities) also earned relatively low average scores (2.0 to 2.10). According to program guidelines, students are expected to spend 20 minutes in each modality. To meet criterion for a particular modality, each student had to spend between 15 and 25 minutes working in that modality during a given session.

Failure Free Reading offers teachers considerable flexibility in meeting program goals. However, a central tenet of the program is that teachers should provide extensive scaffolding so that students do not experience reading failures. The average score of 2.40 on the program techniques dimension reflected instances in which the scaffolding was somewhat inadequate.

In summary, there were relatively few instances of moderate fidelity problems, and no instances of severe fidelity problems, across programs and dimensions. Such problems as did occur tended to be concentrated in the fine points of program techniques and total session time. With many sessions in all four programs running shorter than intended, it was also the case that activities at the ends of the sessions tended to get short changed more often than activities occurring earlier. This was particularly evident in Spell Read, where nearly all of the sessions met criterion for the duration of the phonics portion of the lesson, but only about half met the criterion for the duration of the reading and writing activity that came at the end of the session. This had implications for the time-by-activity described in the next section.

4. Cross-Program Comparisons on Videotape Ratings

To compare videotape ratings across programs we collapsed the ratings for each program into a common set of dimensions and then constructed two superordinate ratings. We considered the first, which captured the coverage, time allocation, and program technique dimensions, as representing program fidelity. We classified the second, which encompassed management of instruction, positive reinforcement, affective environment, and, in the case of Failure Free Reading, monitoring student activity, as representing general teaching quality. The superordinate ratings were based on the average scores for the contributing dimensions, after setting aside the "not applicable" ratings.

The mean scores for the overall fidelity rating, by program, were as follows: Corrective Reading = 2.38, Spell Read = 2.61, Wilson Reading = 2.7, and Failure Free Reading = 2.29. These scores were significantly different across the four groups [$F(3, 956) = 23.26$, $p<.001$]. Mean scores for the overall teaching quality rating were as follows: Corrective Reading = 2.91, Spell Read = 2.91, Wilson Reading = 2.76, and Failure Free Reading = 2.86. These ratings were also significantly different across the four groups [$F(3,622) = 5.10$, $p<.01$].

5. Summary of Fidelity and Teacher Quality Ratings

In summary, the several sources of ratings for intervention teachers on both implementation fidelity and general teacher quality included ratings by the reading program trainers who observed the teachers and coached them over a period of months, ratings by the AIU project coordinators who observed a sample of instructional sessions, and ratings based on a sample of videotaped sessions. On all measures, the

Table III.15

Scores on Fidelity Dimensions Coded from Videotape: Failure Free Reading

	Average Score[a]
Coverage of lesson content	2.65
Duration by modality	
Teacher-directed activity	2.10
Independent student activities	2.00
Computer activities	2.05
Proper use of program techniques	2.40
Management of instruction	2.75
Monitoring student activity	2.75
Positive reinforcement	2.85
Affective environment	2.90
Total teaching time	1.80

[a] Scale: 3=meets criterion; 2=minor problems; 1=moderate problems

average scores fell well within the acceptable range for every program. The videotape analysis, however, made it clear that initial expectations for average session length were overly optimistic. Like the proverbial 50-minute therapy hour, the majority of one-hour sessions lasted between 50 and 55 minutes, probably reflecting the realities of elementary school life in which the time required for students to transition from one instructional setting to another is subtracted from the time allocated for instruction.

F. TIME-BY-INSTRUCTIONAL-ACTIVITY ANALYSES

Knowledge of the instruction actually received by students in the study can assist with interpretation of the impacts and assess how closely the program models were followed. As part of our implementation analysis, we conducted two examinations intended to provide more detail on the instruction received by students in each instructional condition. First, we noted how far each instructional group progressed through the available program materials and compared group progress against the scope and sequence for each reading program. Second, we conducted a detailed time-by-activity analysis of the sample of videotaped sessions that were also used in the fidelity analysis. In the latter case, we constructed a set of coding categories for application across programs in order to compare the distribution of activities in each program against each other (see Appendix I). The comparison allowed us to highlight similarities and differences among programs and provided evidence regarding the suitability of the initial planned contrast, which grouped two program interventions as "word level" and the other two as "word level plus comprehension."

1. Progression Through Program Materials

All of the programs provided for flexible pacing through the program materials in order to accommodate the entry skills of the students and the speed with which they master new content. For Corrective Reading, Wilson Reading, and Spell Read, the average capabilities of each three-student instructional group determined the pace. As noted, some of the groups were heterogeneous with regard to students' basic reading skills, leading to difficulties in matching instructional pace to individual student needs. For Failure Free Reading, on the other hand, each student progressed at his or her own pace. Wilson Reading and Spell Read used a common starting point for all instructional groups, whereas Corrective Reading and Failure Free Reading started the groups or students at different points,

depending on pretest results. As a result of such individualized starting points and/or pacing, students were exposed to different portions of their assigned program during the study's instructional period. Appendix I provides a summary of the scope and sequence for each reading program as related to the modal end points for the instructional groups.

Notably, Corrective Reading, Wilson Reading, and Spell Read all provide systematic and explicit instruction in phonemic decoding strategies for reading new words in text. Consequently, progress through the curricula implies exposure to an increasingly broad range of letter/sound combinations and syllable types as well as to an increasing number of irregular words. Passage reading also progresses in complexity as students master additional decoding rules, particularly in programs that base passage reading on controlled text (Corrective Reading and Wilson Reading).

Although Failure Free Reading does not use explicit phonemic instruction, students encounter increasingly complex words and passages as they advance through the program. However, at a certain point, Failure Free Reading students who completed the available story sequences moved into an entirely different type of instruction called Verbal Master, which focuses on learning vocabulary words and practicing writing skills rather than on passage reading.

Given that the Failure Free Reading modules target students with very low reading levels and that many students in this study were near or just below average, a substantial portion of the fifth-grade students assigned to this intervention progressed to Verbal Master at some point during the instructional sequence. Sixty-five percent of fifth-grade students spent half or more of their instructional time in Verbal Master. Fewer than three third-grade students spent half or more of their instructional time in Verbal Master.[25]

2. Time-by-Activity Analysis

Using the same running records that were constructed for the fidelity coding, we undertook a time-by-activity analysis of 18 Corrective Reading, 18 Wilson Reading, 20 Spell Read, and 20 Failure Free Reading sessions that had been videotaped over the course of the intervention period. Based on the running records, we noted beginning and ending times for each activity observed within the session and used coders' notes in conjunction with the relevant instructional materials (such as students' workbooks and instructors' manuals) to analyze each activity along three dimensions: (1) language level; (2) instructional process; and (3) format. Appendix I details the coding structure used for the analysis, and it also provides a detailed treatment of differences among the programs in the three coding categories.

3. Comparison of Interventions

It was our intention in this study to pair instructional interventions into two categories. We paired Corrective Reading and Wilson Reading as word-level interventions, in which instruction would focus primarily on the development of word-recognition skills and emphasize learning to read with accuracy and fluency. We paired Spell Read and Failure Free Reading as word-level plus comprehension interventions that would strike an even balance between word-level skill development and activities intended to develop vocabulary and comprehension. However, as is documented by the analyses reported in Appendix I, the interventions originally conceived of as pairs sometimes were significantly

[25] We cannot disclose the actual number of third-grade students in this category due to Institute for Education Sciences confidentiality standards.

different from one another along one or more dimensions while programs in opposite pairs were sometimes more similar to one another.

The most important point from the detailed time by activity analysis reported in Appendix I is that, across the instructional programs, the distribution of time in word-level versus vocabulary/comprehension activities did not conform to the categorization of the interventions based on the description of instructional activities from the program providers. As implemented, one of the programs in the WL category, Corrective Reading, spent more time on comprehension-oriented activities than one of the interventions in the WL+C condition, Spell Read. Failure Free Reading was the only intervention that had a relatively even balance between word-level and comprehension/vocabulary instruction. An estimate of total time spent during each instructional session on activities to increase reading accuracy/fluency was obtained by adding together the times in the decoding and encoding categories reported in Appendix I. Conversely, an estimate of the time spent on vocabulary/comprehension activities was obtained by adding together the times in the vocabulary and comprehension categories. Table III.16 displays the resultant distribution of instructional time (in minutes) spent by each program in each of these two major areas. Overall instructional time is reduced to the extent that sessions tended to run shorter than the anticipated hour as well as by time transitions between activities and by any time spent off task. The lower overall amount of instructional time noted for the Wilson program in Table III.16 is at least partially a function of how the instructional times were coded from the videotapes. We entered the start and stop times for each instructional activity, and did not include set up or transition time in these counts. If students were getting out letter tiles, or putting away journals, etc., we did not count that as part of the instructional time, even if the teacher was beginning to talk about the next task. The Wilson Reading program was more affected by this rule than the other programs, although the overall session time for the Wilson teachers was also slightly shorter than for the other programs (Wilson = 51.9 minutes, Failure Free = 53.9 minutes, Corrective = 54.5 minutes, and Spell Read = 50.4 minutes).

Table III.16

Minutes per Session Spent on Instruction to Improve Word Reading Accuracy/Fluency versus Time Spent on Activities Related to Vocabulary/Comprehension

	Failure Free Reading	Spell Read	Wilson Reading	Corrective Reading
Word level	23.6	43.6	40.8	39.1
Comprehension/vocabulary	25.6	9.0	2.8	13.4
Total time	49.2	52.6	43.6	52.5

This time by instructional activity analysis suggests that, in terms of the distribution of activities focused on word-level versus vocabulary/comprehension skills, Corrective Reading was more similar to Failure Free Reading than was Spell Read. However, Corrective Reading is much more similar to Wilson Reading and Spell Read in its approach to increasing accuracy and fluency of word reading than it is to Failure Free Reading, which does not teach phonemic decoding strategies at all. Furthermore, in terms of total amount of time per session devoted to instruction focused on building word reading accuracy and fluency, Corrective Reading is very similar to the other two phonemically oriented programs. Because the distribution of instructional activities across programs differed from expectations and because of the particularly large differences between the three phonemically oriented programs and Failure Free Reading with respect to the method used to increase reading accuracy, we could not justify reliance on the original instructional categorization scheme in the analysis of instructional impacts. Where programs are grouped together for the sake of statistical power to examine differences between

intervention and control conditions, we group together the three phonemically oriented approaches. We also do not make direct contrasts between instructional conditions because, without being able to group programs together in a two-by-two categorization, the present experiment does not have sufficient statistical power to warrant such comparisons.

In retrospect, it is unfortunate that we were not able to complete a pilot time by activity analysis of the instructional conditions before the study was implemented. However, given the time frame under which this study was designed, funded, and implemented, it was simply not possible to do the kind of analysis that could have helped us determine the lack of fit between our classifications and the programs that were chosen for the study. In the case of Spell Read, moreover, even a detailed review of program documents would not have revealed the extent of the misfit, which arose in this specific implementation, apparently as a consequence of compressing program delivery into a 60 minute time period, rather than the 70 minute session preferred by the program. Within the instructor's guide, as well as within the training provided to the teachers, the word level elements are highly structured and take a considerable amount of time to complete in a way that is true to the design of the program. As a consequence, the less highly structured reading and writing component, which also comes at the end of the session, was not given as much instructional time as planned.

G. TEACHER REPORTS OF STUDENTS' HOURS OF READING INSTRUCTION

The survey forms filled out by the classroom teachers asked questions intended to elicit information about and quantify the reading instruction delivered to each student in the intervention and control groups. Some of the questions pertaining to the reading mix asked how much reading instruction each student received in large groups, small groups, and one-on-one settings. The questions also allowed us to categorize the instructors providing the reading instruction as either "General education teachers" or "Specialist teachers," the latter defined here as a special education teacher, a Title I teacher, an ESL teacher, a reading specialist, or other instructor.

We analyzed data from the classroom teacher surveys using a weighted conditional two-level hierarchical linear model, with students and school units making up the two levels. These data were analyzed in a manner similar to the outcomes for reading performance, and a more detailed explanation of those procedures is provided in Chapter IV.

During data cleaning, we discovered that the total number of hours of reading instruction reported for some students (i.e., the sum of large-group, small-group, and one-on-one instructional hours) were implausibly high (e.g., 45 hours per week) and thus erroneous. Given that we were unable to find a pattern in these erroneous reports that would allow us to make corrections, we decided to limit the analysis to students whose reported reading instruction totaled no more than 20 hours per week. We chose the 20-hour cut-off criterion (4 hours per day) because it is a high but not implausible number of instructional hours for struggling readers to receive. The analysis included information on 701 students (412 intervention students and 289 control students) with plausible values for total instructional hours.

We created a measure of total weekly hours of reading instruction for each student by summing responses reported by the classroom teachers for weekly hours of reading instruction (other than instruction provided by our intervention teachers) in the following six modalities: large group, generalist teacher; large group, specialist teacher; small group, generalist teacher; small group, specialist teacher; one-on-one, generalist teacher; one-on-one, specialist teacher. For students in the intervention group, we added a constant amount of 4.5 hours of small-group instruction per week. The results of our comparisons of instructional hours received by intervention and control students are summarized below. Overall, we found no significant difference [$t(650) = 1.47$, $p = .1415$] between the combined intervention groups' mean 9.3 average weekly hours of reading instruction and the combined control

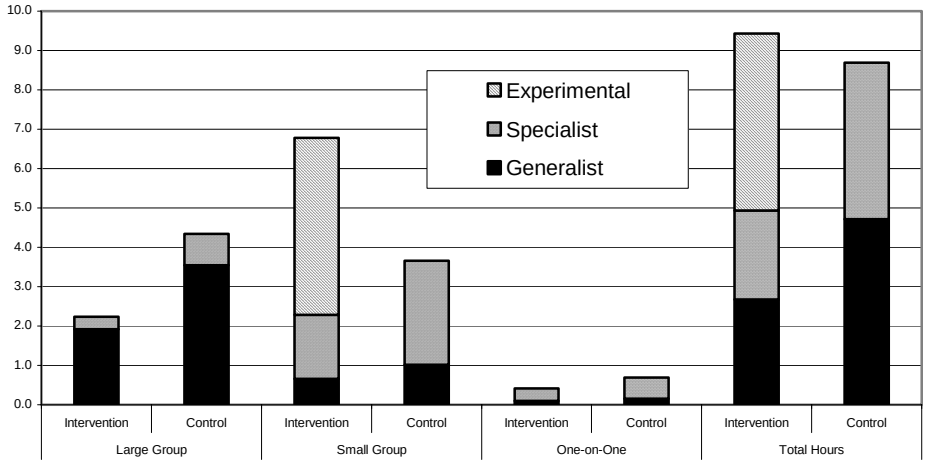
groups' mean 8.7 hours. We also found no significant differences in total hours between grades three and five [t(650) = 1.81, p = .0711].

Figure III.1 presents the average number of weekly reading instructional hours for the intervention and control groups disaggregated by size of the instructional group (large, small, one-on-one), and by who provided the instruction (generalist, specialist, teachers from this study) together with the total hours in each of these categories. Generalist teachers delivered most of the large-group reading instruction, and the control group as a whole received significantly more large-group generalist hours than the intervention group [t(650) = -4.08, p < .0001].

The intervention group received more small-group intervention hours than the control group, with significant differences observed in generalist [t(649) = -2.22, p = .0267], specialist [t(649) = -3.57, p = .0004], and intervention teacher (provided by this study) small-group hours. Most of the small-group reading instruction hours received by the intervention groups came from the 4.5 hours of pull out instruction provided by the study. One-on-one average weekly reading instructional hours were uniformly small (less than 1 hour), with the differences between the combined intervention and control groups not significant [t(649) = -1.46, p = .1453]. Differences in specialist one-on-one hours were not significant between the combined intervention groups and the combined control groups [t(649) = -1.52, p = .1282].

Figure III.1

Average Hours of Reading Instruction per Week in Groups with Different Types of Instructors and of Differing Instructional Size, for Combined Intervention and Combined Control Groups



1. Average Weekly Hours of Reading Instruction by Intervention

Regardless of instructional group size, we also analyzed average weekly hours of total reading instruction for each of the four interventions, comparing them against their individual controls (see Figure III.2). Overall, when looking at the sum of the average weekly hours provided by generalist teachers, specialist teachers, and intervention teachers, we found, at the program level, that only the Corrective Reading intervention differed significantly from its control at the .05 level [t(650) = 1.98, p = .0482]. We did, however, find more significant differences between each intervention program and its control with regard to the mix of small, large, and one-on-one specialist and generalist hours of weekly reading instruction.

Figure III.3 presents the same average hours of reading instruction data, but in groups of different instructional size: large, small, and one-on-one instructional settings. Most noticeable in Figure III.3 is the large magnitude of the small-group reading instructional hours, whereby students in each of the four program interventions received, on average, more than 6 hours of small-group reading instruction per week. Indeed, small-group reading instruction hours represent the large share of the reading instruction hours received by the students in the four intervention programs. This is in contrast to the four intervention controls, which had uniformly fewer small-group hours. Predictably, most of the small-group reading instruction came from the 4.5 hours of pull out instruction provided by the study.

Differences between individual intervention groups and their individual controls with regard to small-group generalist hours were significant at the .01 level for Failure Free Reading [$t(650) = -2.57$, $p = .0104$] and at the .05 level for Corrective Reading [$t(650) = -1.96$, $p = .05$] while differences with regard to small-group specialist hours were significant at the .01 level for Failure Free Reading [$t(650) = -3.46$, $p = .0006$] and Spell Reading P.A.T. [$t(650) = -3.31$, $p = .001$].

Figure III.3 also illustrates that a general education teacher delivered most of the weekly large-group reading instruction. Here, we found significant differences between the intervention groups and their controls at the .05 level for Failure Free Reading [$t(650) = -2.24$, $p = .0257$], at the .1 level for Spell Read [$t(650) = -1.7$, $p = .0903$], and the .01 level for Wilson Reading [$t(650) = -3.07$, $p = .0022$.]. We also found significant differences in the number of hours of large-group reading instruction provided by a specialist for Failure Free Reading [$t(650) = -4.47$, $p = <.0001$].

In addition, Figure III.3 shows that treatment students received less than one hour a week of one-on-one reading instruction. There were significant differences identified between the program interventions and their controls with regard to generalist one-on-one hours at the .01 level for Failure Free: [$t(650) = -3.39$, $p=.0007$] and at the .05 level for Spell Read [$t(650) = 2.14$, $p=.0328$]. In terms of specialist one-on-one hours, there were no significant differences observed between the programs (see details in Appendix K).

2. Tutoring Outside Normal School Hours

The classroom teachers answered questions about any private tutoring in reading that each of their students might be receiving outside normal school hours. When teachers did not know if a particular student was receiving private tutoring, we excluded the student from the tutoring analysis. As a result, only 627 out of 772 observations were available. Overall, we found no significant differences in average weekly hours of private tutoring by treatment/control status [$F(1,627)=.97$, $p=.3254$], treatment program [$F(4,624)=.51$, $p=.7299$], or grade [$F(1,626)=.99$, $p=.3205$]. On average, the control group received .1 hour of weekly tutoring and the treatment group overall .06 hour of average weekly private tutoring outside normal school hours.

Figure III.2

Average Hours of Reading Instruction per Week, in Groups with Different Types of Instructors, for Program Intervention Groups and Program Controls

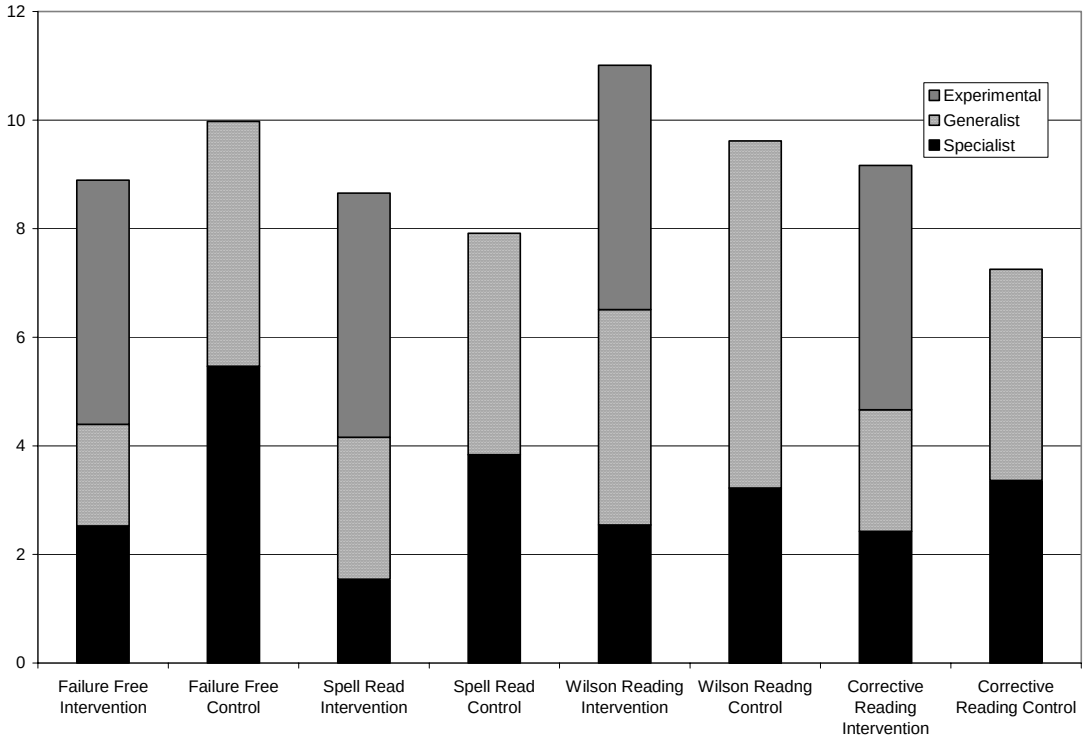
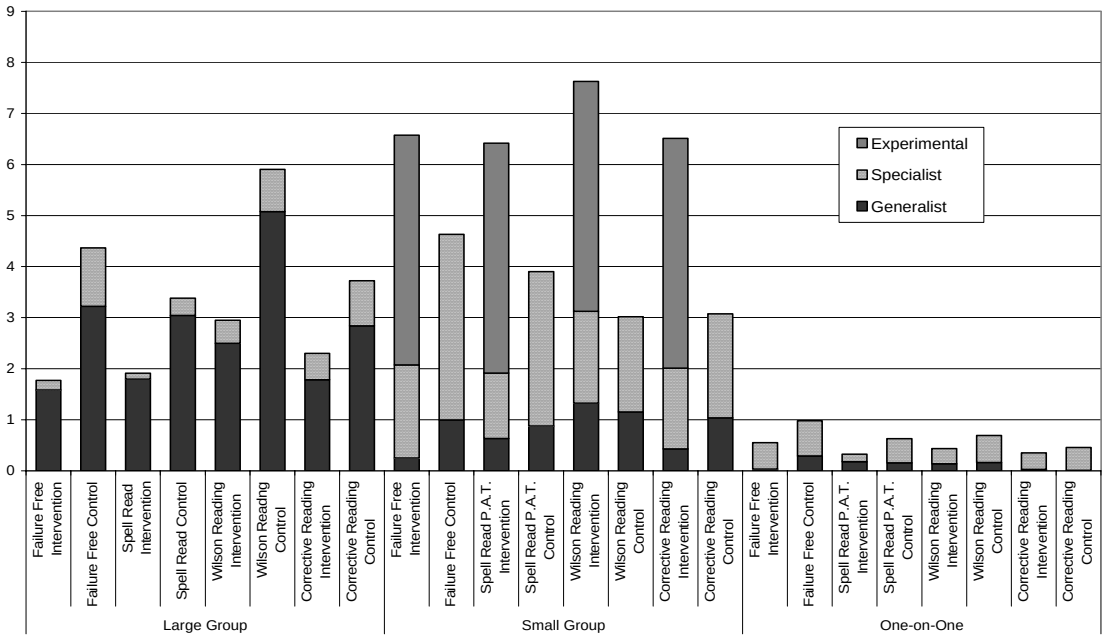


Figure III.3

Average Hours of Reading Instruction per Week in Groups with Different Types of Instructors and of Different Size for Program Intervention Groups and Program Controls



IV. IMPACT ANALYSIS

The main objective of this evaluation is to estimate the impacts of the four interventions on students' reading skills. Specifically, we estimate the impacts of the four interventions combined, the three word-level interventions combined, and each of the four individual interventions for not only all third-grade and all fifth-grade students eligible for the interventions, but also several key subgroups of students. In this chapter, we present the findings of our impact analysis after describing our estimation methods and technical and contextual issues pertaining to the interpretation of the impact estimates.

A. ESTIMATION METHOD

The experimental design can be described as a randomized blocks design with random assignment carried out at two levels. First, as discussed in Chapter II, we randomly assigned 32 school units to the four interventions within blocking strata determined by the percentage of students eligible for free or reduced-price school lunch.[26] Next, within schools, we randomly assigned eligible students within grade levels (third or fifth) to the treatment or control group. The resultant data have a hierarchical structure of students nested within school units.

To reflect the fact that students within a school unit are not independent, in estimating intervention impacts and standard errors we used a weighted two-level hierarchical linear model (HLM) that allows for nested data.[27] The first level corresponds to students within school units and the second to the school units, accounting for the clustering (nonindependence) of students in school units.

Research has shown that the impacts of interventions may vary by age and that older students experience more difficulty in improving their reading skills (Torgesen 2005). To test for differential impacts, we estimated impacts separately for third and fifth graders. The model is:

Level One: Student (i) within school unit (j)

$$y_{1ij} = \beta_{0j} + \beta_{1j}T_{ij} + \beta_{2j}y^{*}_{oij} + \beta_{3j}G^{3}_{ij} + \beta_{4j}T_{ij}G^{3}_{ij} + r_{ij} \qquad \text{(IV.1)}$$

[26] The sample includes 31 school units with about 740 students; one school unit dropped out of the study after random assignment, but before learning the intervention to which it had been assigned.

[27] We also investigated a three-level model that includes a level for the clustering of students in instructional groups. The results are similar when using the three-level model; see Appendix F for details of that model and the results. In most cases, standard errors of the impacts are smaller in the three-level model, but not enough to change our conclusions about impacts.

Level Two: School unit (*j*)

$$\begin{aligned}
\beta_{0j} &= \gamma_{00} + \gamma_{01}A_j + \gamma_{02}B_j + \gamma_{03}C_j + \sum_{l=1}^{3}\xi_{0l}P_{lj} + \mu_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}A_j + \gamma_{12}B_j + \gamma_{13}C_j + \sum_{l=1}^{3}\xi_{1l}P_{lj} + \mu_{1j} \\
\beta_{2j} &= \gamma_{20} + \gamma_{21}A_j + \gamma_{22}B_j + \gamma_{23}C_j + \sum_{l=1}^{3}\xi_{2l}P_{lj} + \mu_{2j} \\
\beta_{3j} &= \gamma_{30} + \gamma_{31}A_j + \gamma_{32}B_j + \gamma_{33}C_j + \sum_{l=1}^{3}\xi_{3l}P_{lj} + \mu_{3j} \\
\beta_{4j} &= \gamma_{40} + \gamma_{41}A_j + \gamma_{42}B_j + \gamma_{43}C_j + \sum_{l=1}^{3}\xi_{4l}P_{lj} + \mu_{4j}
\end{aligned} \qquad \text{(IV.2)}$$

where

$T_{ij} = 1$ if student i in school unit j was randomly assigned to the treatment group (intervention), and $T_{ij} = 0$ if student i in school unit j is in the control group;

$G_{ij}^{3} = 1$ if student i in school unit j is in third grade, and $G_{ij}^{3} = 0$ if student i in school unit j is in fifth grade;

$A_j = 1$ if school unit j was randomly assigned to the Failure Free Reading intervention, and $A_j = 0$ otherwise;

$B_j = 1$ if school unit j was randomly assigned to the Spell Read P.A.T. intervention, and $B_j = 0$ otherwise;

$C_j = 1$ if school unit j was randomly assigned to the Wilson Language Training intervention, and $C_j = 0$ otherwise;

$P_{1j} = 1$ if school unit j is in blocking stratum 1, and $P_{1j} = 0$ otherwise;

$P_{2j} = 1$ if school unit j is in blocking stratum 2, and $P_{2j} = 0$ otherwise;

$P_{3j} = 1$ if school unit j is in blocking stratum 3, and $P_{3j} = 0$ otherwise;

y_{1ij} = post-test score;

y_{0ij}^{*} = centered pretest score.

For our analyses, we use a centered pretest score:

$$y_{0ij}^{*} = y_{0ij} - \bar{y}_{..}, \qquad \text{(IV.3)}$$

where $\bar{y}_{..}$ is the weighted mean of the pretest score across all students in the evaluation sample. By mean-centering the pretest score (the score at the beginning of the school year), we can interpret parameters and combinations of parameters in the level-one model as means for students with the average baseline test score. For example, the impacts, estimated as described below, are interpreted as the impact for a student in a given grade (third or fifth) with a baseline test score equal to the average baseline test score.

The level-one model (Equation (IV.1)) relates a student's post-intervention test scores to a treatment indicator, an indicator for being in third grade, the student's pretest score, and a residual term (unexplained variation). The level-two model (Equation (IV.2)) relates the level-one parameters (coefficients $\beta_{0j}, \beta_{1j}, \beta_{2j}, \beta_{3j}$, and β_{4j}) to indicators showing the interventions to which the school units were randomly assigned as well as the blocking strata. The interventions Failure Free Reading, Spell Read, Wilson Reading, and Corrective Reading are denoted as A, B, C, and D, respectively.[28] The blocking strata grouped school units into four approximately equal-sized groups based on the percentage of students eligible for free or reduced-price school lunch (FRPL). We represent the four blocking strata with three dummy variables, where each dummy variable equals 1 for school units that belong to that blocking stratum, and zero otherwise.[29]

1. Estimation of Impacts

The main parameters of interest are those from which we estimate the impacts of the interventions on students' reading skills, where an impact is defined as the regression-adjusted difference in the average achievement scores for the treatment and control groups.[30,31] As we describe later, we compute three sets of impacts. The first set describes the impact of the interventions on all students. This set of impacts shows how much difference an intervention will make if it is made available to students with characteristics similar to those of the students in the evaluation sample. This is also the most robust estimate of program impacts because it involves the fewest assumptions when estimating the impacts. The second and third sets of impacts describe the intervention impacts on those who participated in the interventions and on those who received at least 80 hours of instruction, respectively. Given that almost all students in the treatment group received some of the treatment and that a very large percentage received 80 or more hours of instruction, the results are similar regardless of the definition of impacts.

[28] The listed order of the interventions and labels A, B, C, and D are arbitrary and not related to the performance of the interventions. In the hierarchical model, we can represent the four interventions with three dummy variables: A, B, and C. Intervention D is represented when the dummy variables for interventions A, B, and C all equal zero (i.e., A=B=C=0).

[29] When estimating impacts, we weight the blocking strata effects equally.

[30] Our analyses compare the treatment students in each intervention to control students in the same schools, which require minimal assumptions about how the controls differ across interventions compared with an analysis that pools all of the controls. The impacts refer to the average impacts across school units and to students with the average baseline test score.

[31] Appendix C provides details on deriving the impact equations.

From the HLM model, we estimate impacts for each of the four interventions.[32] We also estimate the impact of assignment to any of the interventions—denoted as the combined intervention impact (ABCD) —as the average of the four intervention impacts.

As explained earlier in this report, we had originally intended to group the four intervention programs into two intervention classes, word-level interventions and word-level–plus comprehension/vocabulary interventions. However, the time-by-activity analysis indicated that such a categorization was not accurate. In actuality, three of the interventions, Corrective Reading, Spell Read, and Wilson Reading, were appropriately grouped as phonemically oriented word-level interventions while the fourth, Failure Free Reading, provided non-phonemically oriented support for reading accuracy and fluency along with instruction in comprehension and vocabulary. For the analyses reported here, we consider impacts for:

1. All interventions combined (ABCD)
2. The three word-level interventions combined: Spell Read, Wilson Reading, and Corrective Reading (BCD)
3. The four individual interventions (A,B,C,D)

In addition to estimating impacts for all third or all fifth graders, we estimated impacts for subgroups of students within each grade. The ability to estimate impacts for subgroups and to test for differences in impacts between subgroups is important in that it allows for potentially better targeting of interventions, for example, to students with especially low phonemic decoding skills. To estimate subgroup impacts, we modified the model specification in Equation (IV.1) to allow for different impacts (within each grade) for a subgroup (see Appendix C).[33]

We define the impacts when grouping interventions as:

$$\begin{aligned} &\text{Impact of being in any intervention (ABCD)} = (I_A^g + I_B^g + I_C^g + I_D^g)/4 \\ &\text{Impact of being in a word-level intervention (BCD)} = (I_B^g + I_C^g + I_D^g)/3, \end{aligned} \tag{IV.4}$$

where the intervention impacts for Failure Free Reading (A), Spell Read (B), Wilson Reading (C), and Corrective Reading (D), respectively, are:

[32] We used HLM 5 ® software published by Scientific Software International, Inc., to obtain the HLM estimates. We obtained parameter estimates using the restricted maximum likelihood (REML) procedure, as discussed in Raudenbush and Bryk (2002).

[33] Preliminary analyses showed substantial differences in impacts by grade. Because of the differences in impacts, we allowed subgroup impacts to vary by grade level. When we designed the study, our power analyses assumed that we could combine grades when conducting subgroup analyses. Because we cannot, our ability to detect significant impacts for subgroups is diminished. The probability of detecting differences between subgroups is particularly low. See Chapter II for estimates of minimum detectable impacts.

$$
\begin{aligned}
I_A^g &= \hat{\gamma}_{10} + \hat{\gamma}_{11} + (1/4)(\hat{\xi}_{11} + \hat{\xi}_{12} + \hat{\xi}_{13}) + (g)\left(\hat{\gamma}_{40} + \hat{\gamma}_{41} + (1/4)(\hat{\xi}_{41} + \hat{\xi}_{42} + \hat{\xi}_{43})\right) \\
I_B^g &= \hat{\gamma}_{10} + \hat{\gamma}_{12} + +(1/4)(\hat{\xi}_{11} + \hat{\xi}_{12} + \hat{\xi}_{13}) + (g)\left(\hat{\gamma}_{40} + \hat{\gamma}_{42} + (1/4)(\hat{\xi}_{41} + \hat{\xi}_{42} + \hat{\xi}_{43})\right) \\
I_C^g &= \hat{\gamma}_{10} + \hat{\gamma}_{13} + (1/4)(\hat{\xi}_{11} + \hat{\xi}_{12} + \hat{\xi}_{13}) + (g)\left(\hat{\gamma}_{40} + \hat{\gamma}_{42} + (1/4)(\hat{\xi}_{41} + \hat{\xi}_{42} + \hat{\xi}_{43})\right) \\
I_D^g &= \hat{\gamma}_{10} + (1/4)(\hat{\xi}_{11} + \hat{\xi}_{12} + \hat{\xi}_{13}) + (g)\left(\hat{\gamma}_{40} + \hat{\gamma}_{42} + (1/4)(\hat{\xi}_{41} + \hat{\xi}_{42} + \hat{\xi}_{43})\right),
\end{aligned}
\tag{IV.5}
$$

where $g = 1$ for third graders and $g = 0$ for fifth graders.[34] When the interventions are grouped, each intervention in the group receives equal weight.

2. Effect of Treatment on the Treated

The impacts described in the previous section are known as intent-to-treat (ITT) impacts because they estimate the impact of random assignment to one of the interventions (the treatment group), without taking into account whether students actually receive the treatment. In this study, a few students assigned to the treatment group did not participate in one of the interventions. To adjust for students not participating in the intervention or for participating for substantially fewer hours than planned, we provide additional estimates of intervention impacts. We refer to these estimates as the impact of the treatment on the treated (TOT).[35] A TOT impact takes into account the treatment received by students in the study but requires additional assumptions that are untestable.[36] In this evaluation, a small number of students assigned to the treatment group (13 students, or less than 1 percent) did not receive any instruction and are labeled as no-shows. (Students' reasons for dropping out of the treatment group are described in Chapter II.) In addition, approximately 7 percent of treatment group members received fewer than 80 hours of instruction, which we defined as the threshold for receiving a "full dose" of the intervention. When estimating the effect of the treatment on the treated, we considered both definitions of "the treated."[37]

---

[34] The sum of the three blocking strata parameters ($\hat{\xi}_{11} + \hat{\xi}_{12} + \hat{\xi}_{13}$) is multiplied by ¼ because of the fourth blocking stratum, which is the excluded category. The term could also be written as $\frac{1}{4}(\hat{\xi}_{11} + \hat{\xi}_{12} + \hat{\xi}_{13} + 0)$.

[35] This is also sometimes referred to as the Complier Average Causal Effect (CACE) or the Instrumental Variables (IV) estimate.

[36] Two major assumptions are involved in estimating the TOT impacts. The first is that assignment to the treatment group has no impact on students who do not participate in one of the interventions (Rubin's exclusion restriction, see Angrist, Imbens, and Rubin 1996). For treatment group students who did not show up for any instruction, the assumption is reasonable. However, for those with between 1 and 79 hours of instruction, the assumption probably is not reasonable, and we must use caution in interpreting the TOT estimate for students with the "full dose" (> 79 hours). The second major assumption is that some individuals participate in one of the interventions only when assigned to the treatment group (compliers). The assumption is reasonable here, as most members of the treatment group do participate in one of the interventions, and individuals assigned to the control group do not have access to the interventions. Both of these assumptions are untestable because we observe each individual's behavior and outcomes only under the treatment to which they were assigned; it is impossible to observe the behavior and outcomes of individuals as if they had been assigned to another group. Thus there are no data available on which to test these assumptions.

[37] See Bloom (1984); Angrist, Imbens, and Rubin (1996); or Little and Rubin (2000) for general background information on computing TOT estimates.

Because the TOT impacts rely on untestable assumptions, we present the ITT impacts as the main results and present the TOT impacts in supplementary tables. In this setting, with no control group students who receive the intervention, the TOT impact estimates will always be equal to or greater than the ITT impact estimates. The TOT impacts in this study are similar to the ITT impacts because the percentage of treatment students who received the intervention is very high (0.99 for any treatment received and 0.93 for those with at least 80 hours of treatment). Therefore, the adjustment for no-shows increases impacts by about 1 percent while the adjustment for those who do not receive at least 80 hours of intervention increases impacts by about 8 percent. For example, for an ITT impact of 4 standard score points, the TOT impact adjusted for no-shows is about 4.04 points, and the TOT impact adjusted for those receiving fewer than 80 hours of interventions is 4.28 points.

B. INTERPRETATION OF IMPACTS

In this study we are interested in estimating the impact of the four remedial reading interventions relative to the instruction that students ordinarily receive. When interpreting the impacts of the four interventions on students' reading skills it is important to consider three elements of the broader context in which the interventions were operating: (1) where the students began in terms of reading ability at the beginning of the school year, (2) how much improvement the students would have had in the absence of the interventions, and (3) the amount of the intervention that treatment and control students actually received.

We illustrate the first two elements using a hypothetical example, in Figure IV.1.[38] At the beginning of the school year, all students in the intervention (represented by "T") and control (represented by "C") groups started out at approximately the same point—due to randomization—with an average baseline test score of 85 (16th percentile).[39] This is similar to the actual baseline test scores seen for students in this study (see Tables II.2 through II.7).

The improvement that students would have made in the absence of the interventions is indicated by the gain that the students in the control group experienced between the beginning and end of the school year. In Figure IV.1, this gain is 4 standard score points, as shown by the dashed line.

Because standard scores show students' relative standings in a national population of students at a given grade level, we would expect the average gain to be 0 if we had a national sample of students at all levels of reading ability. However, the students in this example (and in the actual study) began the year reading below grade level, indicated by standard scores less than 100. For such students, positive gains indicate the amount by which the students at least partially "caught up" to the average student in their grade. Negative gains indicate the amount by which the students fell further behind.

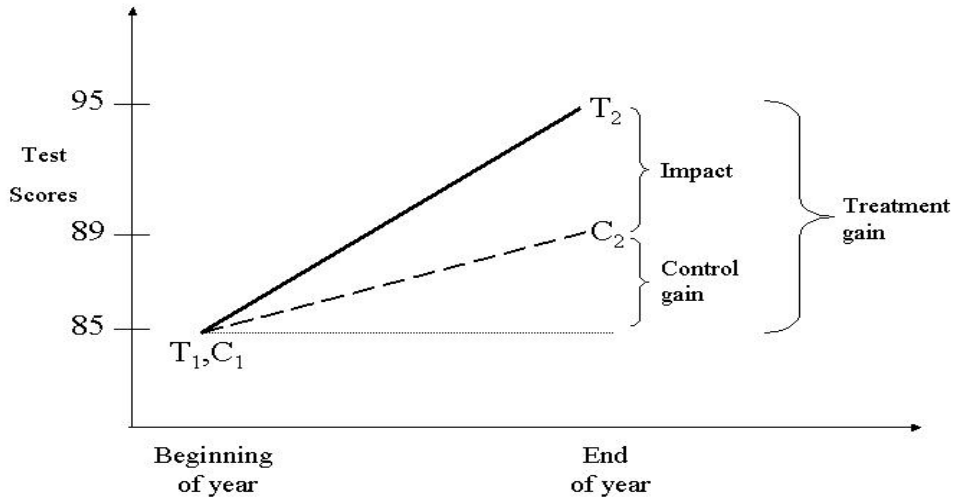
The impact shows the value added by the intervention; that is, the gain above that achieved by the control group. In other words, the impact is the amount that the interventions increased students' test scores relative to the control group. Because of random assignment, the intervention and control groups started out at the same place (85, in this example), and thus the impact can be calculated by comparing

[38] The third element is discussed in the next section.

[39] Randomization ensures that the treatment and control students start out with similar reading ability (similar test scores). However, there may still be small differences between the groups that are attributable to chance, unless the samples are very large. The HLM model in this analysis adjusts for the small differences that may exist between the groups.

Figure IV.1

Hypothetical Example of Gains and Impact



either the end of year test scores for the intervention and control groups or the test score gains for the intervention and control groups. Using the end of the year test scores, the impact in Figure IV.1 is 95-89=6 (T_2-C_2). Alternatively, using gain scores, the impact in Figure IV.1 is (95-85)-(89-85)=10-4=6 ((T_2-T_1)-(C_2-C_1)). Thus, the intervention in this example raised students' test scores 6 points higher than they would have been without the intervention.

The change ("gain") in the intervention group students' average test scores between the beginning and end of the school year can be calculated by adding the control group gain and the impact, as illustrated in Figure IV.1. If the control group students' average score increased between the beginning and end of the school year and there is a positive impact, then the treatment group gain will also be positive, as in Figure IV.1, where the treatment group gain is 10 points. However, if the control group students' scores decreased between the beginning and end of the school year, then the intervention group may also experience a negative gain, even if the impact is positive. Depending on the relative magnitudes of the control group gain and the impact, a negative control group gain combined with a positive impact may imply that the intervention group students held their ground (or improved) while the control group declined, or may imply that the intervention group experienced a negative gain as well.

C. CONTEXT OF THE IMPACTS

We now consider our empirical findings pertaining to the three elements of the broader context for this evaluation: (1) where the students began in terms of reading ability at the beginning of the school year, (2) how much improvement the students would have had in the absence of the interventions, and (3) the amount of the intervention that treatment and control students actually received. Indicating where students began, the first column of Table IV.1 shows the baseline test scores of students in third and fifth grades. (All tables appear at the end of this chapter.) The average baseline test scores are all below

average (less than 100)—ranging from a low of 81 (10th percentile) for the Phonemic Decoding Efficiency test in the fifth grade to a high of 93 (32nd percentile) for Word Attack in third and fifth grades, and Passage Comprehension in fifth grade. Although not as severely impaired as many of the students studied in previous small-scale assessments of intensive reading interventions (see the review by Torgesen, 2005), the typical student in our evaluation is struggling with basic reading skills. That student along with a substantial fraction of the broad range of students included in our sample are among those often targeted by providers and school districts for the types of interventions that we are evaluating. Such targeting is a response to both the needs of these students and the fact that except perhaps in the largest urban school districts, most schools would have only a small number of students in each grade who are as severely impaired as the students included in some previous studies. While it is important to assess the effectiveness of interventions for these more severely impaired students, the results obtained might not pertain to broader groups of struggling readers that include less severely impaired students. Hence, we have drawn our sample from regular elementary schools and included students with a relatively wide range of reading difficulties.

When we assess the improvement that students had in the absence of the interventions, we see mostly positive gains among the control students in both third and fifth grade, as presented in Table IV.1. In the third-grade pooled (ABCD) control group, students typically had positive gains between 0 and 3 standard score points, but there were some negative gains, particularly on the reading comprehension subtest from the Group Reading Assessment and Diagnostic Evaluation (GRADE). The negative gain on the GRADE test suggests that the average student in the study lost ground relative to other students on this reading comprehension test. That is, if third-grade students selected for the study had not participated in an intervention, we would expect them, on average, to lose ground in their ability to extract meaning from text, as measured by the GRADE test. Among fifth graders, the gains were nearly all positive, generally between 1 and 6 standard score points. The exception for fifth-grade control students is the Passage Comprehension test, which had a small negative gain.

The generally positive gains experienced by the control students indicate that these students' reading ability improved between the beginning and end of the school year relative to the normal growth expected during this time. A positive control gain may be due to students' usual classroom instruction, additional instruction received in or out of school, or a statistical phenomenon known as "regression to the mean." Regression to the mean can occur when students are selected for a study because of low scores on a test because students are more likely to be selected when testing error was negative. The next test is more likely to have a positive testing error or a smaller negative testing error, which appears to be a gain but is instead an artifact. In the case of the present study, students were selected on the basis of their screening—not baseline—scores. Thus, for the sample of all students, the regression to the mean effect should have occurred between screening and baseline testing, not between baseline and follow up. Thus, the phenomenon of regression to the mean is not likely to play a significant role in explaining the reading gains of students in either the intervention or control groups in the study sample. However, in subgroup comparisons that select students because of either low or high scores on a given measure within the total sample, regression to the mean could certainly explain some of the improvement (or some of the decline) in scores between the baseline and follow-up testing.

The final contextual element to consider when interpreting impacts is the amount and type of reading instruction that the students in the study actually received. During the school year, each student in the intervention group was supposed to receive approximately 60 minutes of reading instruction per school day. However, as reported in Chapter III, we found that when the interventions were implemented, students received 54.1 minutes of instruction per day on average, and the amount of instruction received was similar across the interventions. By design, none of the control students received the intervention. Instead, the students in the control group received their typical instruction, which included regular classroom instruction and often included other services such as another pull out program.

Thus, the impacts presented in Tables IV.1 to IV.10—the ITT impacts—show the effects of students' being given the opportunity to receive a little less than one hour of intensive reading instruction per day, implemented as a pull out program from their usual classrooms, where they might have received some additional reading instruction if they had not been assigned to the intervention group. Tables IV.21 to IV.30 provide parallel estimates of the effect of actually receiving the intervention, which take into account the percentage of intervention students who did not participate (TOT impacts). We define participation in two ways: (1) receiving any intervention instruction and (2) receiving more than 80 hours of instruction. Because nearly all students in the intervention group participated for at least one hour (99 percent), and most received over 80 hours of instruction (93 percent), the TOT estimates are very similar to the ITT estimates discussed in the text.

Preliminary analyses showed substantially different patterns of impacts by grade. Although the impacts are not significantly different at the 0.05 level between the third and fifth graders, the point estimates of the impacts for the two grades often appear quite different (see Table IV.1). Furthermore, more significant impacts—that is, impacts that are different from zero—are found for third graders than for fifth graders. In light of these findings, we present results separately by grade in the following sections.

As discussed in Chapter II, we present impacts on seven measures of reading ability that fall into three categories. Two tests measure phonemic decoding ability: the Woodcock Reading Mastery Test-R Word Attack test and the Phonemic Decoding Efficiency subtest of the Test of Word Reading Efficiency. Three tests measure word reading accuracy and fluency: the Woodcock Reading Mastery Test-R Word Identification test, the Sight Word Efficiency subtest of the Test of Word Reading Efficiency, and the Oral Reading Fluency (Aimsweb) test. The third category, reading comprehension, is assessed using the Woodcock Reading Master Test-R Passage Comprehension test and the Group Reading Assessment and Diagnostic Evaluation (GRADE) Passage Comprehension subtest.

When estimating impacts for multiple outcomes—such as these seven measures of reading ability—and testing multiple interventions, there is a concern that some estimated impacts will be found to be significantly different from zero, even if there is actually no impact of the interventions (a "Type 1" error). In fact, even if there were no differences between the treatment and control groups, five percent of test statistics comparing the outcomes of the two groups would be expected to be significant at the five percent level just by chance. A variety of procedures have been developed to address the concerns around this, with varying levels of complexity. To maintain a straightforward presentation of results, without introducing the complexities of and debate surrounding the details of the implementation of multiple comparisons adjustments, the impacts presented here in the main text do not include an adjustment for multiple comparisons. However, we present in Appendix D the results using two methods that adjust the significance levels of tests to account for the number of tests being performed: the Bonferroni correction, and a more powerful adjustment developed by Benjamini and Hochberg (1995) that is particularly relevant for this study, where interest is in assessing the impact of an intervention on multiple outcomes. The results in Appendix D show that adjustments for multiple comparisons do not affect the general conclusions of this report.

D. IMPACTS FOR THIRD-GRADE STUDENTS

Combined, the four interventions improved the phonemic decoding skills of third graders, raising Word Attack scores by approximately 5 standard score points (effect size 0.33)[40] and Phonemic Decoding

[40] The impacts presented in this report are generally in terms of standard scores; however, they can also be expressed as effect sizes, which divide the impact by the standard deviation of the standard score. The effect sizes corresponding to the impacts in Tables IV.1 through IV.10 are shown in Tables IV.11 through IV.20. Because an

Efficiency scores by approximately 3 points (effect size 0.20), as seen in Tables IV.1 and IV.11. These impacts for the pooled interventions (ABCD) suggest that being assigned to one of the reading interventions moved students in the interventions up the distribution of phonemic decoding ability approximately 5 to 10 percentile points more than they would have gained had they not been in one of the interventions.[41] When assessing the impacts of the three word-level interventions (BCD)—Spell Read, Wilson Reading, and Corrective Reading—we also found impacts on both of these measures of phonemic decoding ability. However, individually, not all of the interventions had impacts on the accuracy and fluency of phonemic decoding. Failure Free Reading had no impacts on these measures, and Corrective Reading had an impact only on Word Attack test scores and not on Phonemic Decoding Efficiency test scores. In contrast, Spell Read and Wilson Reading improved scores on both tests, with effect sizes of approximately 0.4 to 0.6, corresponding to moving students in those interventions up the distribution of reading ability approximately 12 to 19 percentile points more than they would have gained had they not been in one of the interventions.

The four interventions combined and the three word-level interventions combined improved reading accuracy and fluency. This is primarily due to impacts of Corrective Reading, as Failure Free Reading, Spell Read, and Wilson Reading had no impacts on fluency. Corrective Reading improved scores on the Word Identification test by about 3 standard score points (effect size 0.22), scores on the Sight Word Efficiency test by about 5 points (effect size 0.30), and the number of correct words per minute read on the oral reading passages (Aimsweb) by about 11 words (effect size 0.27). These impacts correspond to moving students up the distribution of reading ability by approximately 5 to 10 percentile points more than they would have gained had they not been in one of the interventions.

Together, the four interventions had an impact of about 5 standard score points on third graders' reading comprehension (effect size 0.31) as measured by the GRADE test, but not as measured by the Passage Comprehension test. Although the impact is substantial for the GRADE, it is important to consider the experience of the control group. For the controls, there was a decline in comprehension scores of about 4 points between the fall baseline test and the spring follow-up test. Thus, the impact of 5 standard score points on comprehension for the combined interventions was obtained mostly because of this decline in scores in the control group. Students in the intervention groups actually gained only 1 standard score point, in absolute terms, between the baseline and follow up testing. In addition, despite the combined impact on GRADE test scores, neither the three word-level interventions combined nor any of the individual interventions had a statistically significant impact on either measure of reading comprehension.

(continued)

objective of the study is to measure the extent to which struggling readers catch up with students in the full population, we use the population standard deviation of each test to calculate effect sizes. This standard deviation is 15 for all tests, with the exception of the Aimsweb, which has a standard deviation of 39 for third graders and 47 for fifth graders. An effect size of 1 means that the intervention increased test scores by 1 standard deviation.

[41] Effect sizes can be converted into the number of percentile points by which the intervention moved students up in the distribution of reading ability. For example, for students who started out at approximately the 16th percentile on most tests, an effect size of 0.3 means that the interventions moved students up 8 percentile points more than they would have risen had they not received the intervention. Therefore, if control group students move from the 16th to the 18th percentile, the treatment group students would move from the 16th to the 26th percentile. Appendix K gives approximate percentile increases for other effect sizes, for the students in this study.

E. IMPACTS FOR FIFTH-GRADE STUDENTS

The interventions had fewer impacts for fifth graders than for third graders (see Table IV.1 for impacts and Table IV.11 for effect sizes). Combined, the four interventions improved fifth graders' phonemic decoding skills by approximately 3 points (effect size 0.18) on the Word Attack test, but they did not have a statistically significant impact on Phonemic Decoding Efficiency test scores. At the end of fifth grade, students in the control group had an average Word Attack score of approximately 95 (37th percentile), while the average score among students in the interventions was approximately 98 (45th percentile). The three word-level interventions also improved scores on the Word Attack test, with an impact of about 4 points (effect size 0.26), but they did not have a statistically significant impact on scores on the Phonemic Decoding Efficiency test. Across the individual interventions, only Spell Read and Wilson Reading had significant impacts on Word Attack test scores, and only Spell Read had a significant impact on Phonemic Decoding Efficiency test scores. Spell Read and Wilson Reading increased Word Attack test scores by about 5 and 4 standard score points, respectively, corresponding to effect sizes of 0.35 and 0.29.

For fifth graders, the four interventions combined had an impact on only one of the three measures of reading accuracy and fluency: an impact of approximately 1 point (effect size 0.09) on Sight Word Efficiency test scores. Neither the three word-level interventions combined nor any of the individual interventions had an impact on any of the measures of reading accuracy and fluency.

The four interventions, combined, did not affect fifth graders' reading comprehension skills. Similarly, neither the three word-level interventions combined nor any of the individual interventions improved fifth graders' reading comprehension by either measure.

F. IMPACTS FOR SUBGROUPS OF THIRD AND FIFTH GRADERS

Three of the four interventions—Spell Read, Wilson Reading, and Corrective Reading—focus on improving students' word-level reading skills. In order to examine whether the impacts of these interventions and the fourth intervention—Failure Free—were greater for students who began the interventions with more significant impairments in their word-level reading skills (specifically their phonemic decoding skills), we formed subgroups of students based on their entering scores on the Word Attack subtest. Students who began the study with lower scores on Word Attack were further subdivided into those who entered the study with lower or higher scores on the Peabody Picture Vocabulary Test. Since broad vocabulary is one of the significant factors that contribute to performance on measures of reading comprehension (Stahl, 1998), it is of interest to determine whether the impact of the interventions varied among students with different entering scores on this dimension. In addition, because the No Child Left Behind legislation has increased funding for and attention on Title 1 schools, which by definition have high proportions of low-income students, we also examined the impacts of the interventions on students who qualified for free or reduced-price school lunch to determine if the interventions were particularly effective for that group.

The study was not designed to estimate the impacts of the individual interventions on subgroups of students and thus did not enroll sufficient numbers of students to obtain precise estimates of such impacts. For this reason, we focus on the impacts of the four interventions combined and the three word-level interventions combined. The full subgroup results—including the estimated impacts of the individual interventions on subgroups of students—are presented in Tables IV.2 through IV.10, with effect sizes shown in Tables IV.12 through IV.20.

All of the tables of subgroup results contain two types of significance tests. One significance test is used to assess whether the impact for that subgroup is statistically different from 0, as indicated by an asterisk.

That is, within a subgroup—for example, third graders with Word Attack scores below the 30th percentile at the beginning of the school year—an asterisk indicates that the interventions improved reading ability, as measured by that particular test, as compared with the control group. The other significance test is whether the impact for the subgroup is different from the overall impact (within grade levels), as indicated by a pound sign (#). In the example above, a pound sign would indicate that the impact for third graders with low Word Attack scores at the beginning of the year was significantly different from that for all third graders. Comparing third graders with low Word Attack scores to all third graders is algebraically equivalent to comparing third graders with low Word Attack scores to third graders with high Word Attack scores. With the exception of comparisons between impacts for third and fifth graders, in the text we describe the tests as that of a comparison between students with low Word Attack scores and all students because we are interested in determining whether the impacts would be different had we enrolled only students with low Word Attack scores, as compared to the full range of scores found in the study.[42]

1. Students with Relatively Low or High Word Attack Scores at Baseline

The first subgroup examined is students who entered the study with relatively low scores in phonemic decoding—specifically, Word Attack test scores below the 30th percentile. Although the overall average score on the Word Attack test for this subgroup is still substantially higher than has been reported in many earlier intervention studies of substantially more impaired students of this age, there were no students in this group with average or above average scores in phonemic decoding before the intervention began.

The impacts for students with low Word Attack scores were generally similar to those for the full sample of students (see Table IV.2). Among third graders with low Word Attack scores, the four interventions combined and the three word-level interventions combined had positive impacts on both measures of phonemic decoding, as was seen for all third graders. Likewise, the four interventions combined and the three word-level interventions combined improved scores on the measure of reading accuracy (Word Identification) for all third graders and for third graders with low Word Attack scores. However, while the four interventions combined and the three word-level interventions combined also improved scores on the Sight Word Efficiency and Aimsweb fluency tests for the sample of all third graders, they did not improve scores on these tests for third graders with low Word Attack scores. The impacts on reading

[42] The estimated impacts are model-based estimates, derived from the estimated parameters of the two-level hierarchical linear model specified earlier in this chapter. From those estimated parameters, we also derive standard errors for the estimated impacts and statistics for conducting significance tests pertaining to the impacts. These standard errors and test statistics are reported in Appendix M. Although model-based impact estimates are more precise than, for example, simple difference-of-means estimates, some of the reported impacts—especially those for small subgroups—are estimated much less precisely than other impacts that are presented, such as those for all third graders or all fifth graders. When the data do not enable us to have substantial confidence in an estimated impact because, for example, there is substantial variability in outcomes across a small sample of students, the standard error for the impact estimate will be large relative to the impact estimate. Furthermore, the test statistic for testing the hypothesis that the impact is zero will be relatively small, providing insufficient evidence to reject the hypothesis. Then, we conclude that the impact is "not significant." When assessing the potential implications of such a finding, however, it is important to keep in mind the power of the evaluation to detect significant impacts and, especially, the fact that the minimum detectable impact (MDI) of an individual intervention on a subgroup is fairly large—0.7, as noted in Chapter II. (The MDI on a subgroup is 0.35 for the four interventions combined.) As discussed above, the evaluation was not designed to estimate the impacts of the individual interventions on subgroups of students and, thus, did not enroll sufficiently large numbers of students to obtain precise estimates of such impacts. In fact, based on findings from previous studies, this evaluation was designed to detect fairly large impacts—even for all eligible students in a grade—and not to estimate small impacts precisely.

comprehension are similar to those for the full sample, with impacts on GRADE test scores but not on Passage Comprehension test scores. The three word-level interventions also had a statistically significant impact on GRADE test scores for these students.

Among fifth graders with low Word Attack scores, the impacts are similar to those seen for all fifth graders. For fifth graders with low Word Attack scores, the four interventions combined improved Word Attack test scores and one measure of reading accuracy and fluency, albeit a different measure than was seen for the full sample of fifth graders—Word Identification rather than Sight Word Efficiency. The three word-level interventions improved Word Attack scores among this group, as for all fifth-graders, but also improved scores on the Word Identification test. Although for some of the reading measures the size of the impact appears to be larger for the low Word Attack group than for the sample as a whole, the impacts for these two groups are not significantly different from each other in most cases. We thus cannot conclude that low scores on the Word Attack test at the beginning of the school year made a reliable and consistent difference in the size of impacts obtained.

Consistent with that conclusion, in general, the impacts for students with relatively high Word Attack scores at baseline are also similar to those for all students, among both third and fifth graders (see Table IV.3). Among third graders with Word Attack scores greater than 92, the four interventions combined had impacts on almost all of the same tests as was seen for all third graders. The Aimsweb and GRADE tests are the exception; impacts on these test scores are seen for the full sample but not for students with relatively high Word Attack scores. As was seen for the sample of all fifth-grade students, there are only scattered impacts among fifth-grade students with Word Attack scores above 92. In this group the four interventions combined and the three word-level interventions combined had impacts on scores of only one test: the Aimsweb test of reading fluency, a test on which no impacts were seen for the full sample of fifth graders.

2. Students with Relatively Low or High Vocabulary at Baseline

Because the impacts of the interventions may vary by students' broad vocabulary level, we also examined impacts for students with relatively high or relatively low verbal ability according to the Peabody Picture Vocabulary Test—Revised (selecting scores above or below the 30th percentile, respectively). The patterns of impacts for third and fifth graders in these two subgroups are fairly similar to those seen for all third- and fifth-grade students, respectively.

Slightly fewer impacts are seen for third graders with low Peabody Picture Vocabulary test scores than were seen for all third graders. However, none of the differences in impacts is statistically significant (see Table IV.4). It appears as though the four interventions had slightly more impacts on third-grade students who began the year with relatively high Peabody Picture Vocabulary test scores (see Table IV.5), as compared to all third graders, although again, none of the differences in impacts is statistically significant. The four interventions combined improved scores on all three measures of reading accuracy and fluency for third-grade students with high Peabody Picture Vocabulary test scores. For students with low Peabody Picture Vocabulary test scores, the four interventions improved only Sight Word Efficiency scores. The three word-level interventions improved Word Identification and Aimsweb scores for students with high verbal ability, and Sight Word Efficiency scores for students with low verbal ability.

Among the fifth graders with relatively high or low Peabody Picture Vocabulary test scores, the impacts of the four interventions combined, and the three word-level interventions combined are similar to those for all fifth graders. The exceptions for students with low Peabody Picture Vocabulary test scores are that the four interventions combined improved not only Word Attack and Sight Word Efficiency scores but also scores on the Phonemic Decoding Efficiency test. The three word-level interventions

combined improved scores on the Word Identification and Sight Word Efficiency tests in addition to the Word Attack test. The exceptions for students with high Peabody Picture Vocabulary test scores are that the four interventions combined did not improve scores on the Sight Word Efficiency test, but the four interventions combined and the three word-level interventions combined improved scores on the Passage Comprehension test in addition to scores on the Word Attack test.

3. Subgroups Defined Jointly by Baseline Phonemic Decoding and Vocabulary Scores

There was some expectation that the impacts of the interventions might be larger for students with low phonemic decoding ability but relatively high vocabulary, as this would create a sample that is more consistent with the way reading disabilities have been defined, and previous studies have found large impacts for students with severe disabilities (Lyon and Shaywitz 2003). We therefore examined impacts within subgroups defined by baseline Word Attack and Peabody Picture Vocabulary test scores. Each subgroup is approximately 25 percent of the full sample. We generally did not find large differences in impacts across subgroups defined by these tests (see Tables IV.6 through IV.8). The following is a summary of the impacts for three groups of students of particular interest defined by these two tests:

- ***Students with Low Word Attack and Low Peabody Picture Vocabulary Test Scores.***[43] Very few impacts are seen among third graders in this group. In fact, the four interventions combined had an impact only on scores on the GRADE test, and the three word-level interventions combined did not have a statistically significant impact on any measure of reading ability. For fifth graders in this group, the four interventions combined had positive impacts on scores on the Phonemic Decoding Efficiency and Sight Word Efficiency tests.

- ***Students with Low Word Attack and High Peabody Picture Vocabulary Test Scores.***[44] Few impacts are seen for students in this group in either grade. Among third graders in this group, the four interventions combined had impacts only on scores on the GRADE test. The three word-level interventions also improved Word Attack scores. Among fifth graders in this group, the four interventions combined and the three word-level interventions combined had impacts only scores on the Word Attack test.

- ***Students with High Word Attack and High Peabody Picture Vocabulary Test Scores.***[45] Among the third graders in this group, the four interventions combined and the three word-level interventions combined improved Word Attack scores. The four interventions combined and the three word-level interventions combined did not have a statistically significant impact on any other test scores, except that the three word-level interventions had a negative impact on Passage Comprehension scores in this group. Among fifth graders in this group, the four interventions combined and the three word-level interventions combined had impacts only on the two measures of phonemic decoding; no impacts were seen on reading fluency and accuracy or comprehension for fifth graders.

[43] Students in this group had low reading ability as measured by the Word Attack test (below the 30th percentile) and low verbal ability, as measured by the Peabody Picture Vocabulary test (below the 30th percentile).

[44] This group of students had low reading ability (below the 30th percentile) but relatively high vocabulary skills (above the 30th percentile) at the beginning of the school year.

[45] These students began the year with relatively high reading ability and vocabulary skills (above the 30th percentile on both tests).

These findings suggest that the large effects found in some previous studies of severely impaired students might not pertain to broader groups of struggling readers that include, for example, students with only moderately impaired phonemic decoding skills.

4. Subgroups Defined by Eligibility Status for Free or Reduced-Price School Lunch

Because of increased attention on schools with a high proportion of low-income students, we examined whether impacts vary with students' socioeconomic status by estimating impacts (in Tables IV.9 and IV.10) within subgroups defined by eligibility for free or reduced-price school lunch (FRPL).[46] Among third graders, larger impacts were seen for the 58 percent of students ineligible for FRPL (with relatively high family income) than for the 42 percent of students eligible for FRPL (with relatively low family income). The four interventions combined and the three word-level interventions combined had an impact only on Word Attack for third-grade students eligible for FRPL, but had impacts on every test for students ineligible for FRPL, with some significant differences between the groups. The large impacts on all tests for third-grade students ineligible for FRPL appear to be primarily attributable to large impacts of Wilson Reading for this group, which may in turn be partially due to the fact that the Wilson Reading control students ineligible for FRPL experienced large declines in almost all test scores.

Few impacts of the four interventions combined are seen for fifth-grade students who are either eligible or ineligible for FRPL (see Tables IV.9 and IV.10). Among the 57 percent of fifth graders who are eligible for FRPL, the four interventions combined and the three word-level interventions combined had a positive impact only on the Sight Word Efficiency test of reading fluency, and the four interventions combined had a negative impact on the GRADE test of comprehension. Among the 43 percent of fifth-grade students ineligible for FRPL, the four interventions combined and the three word-level interventions combined had impacts only on the Word Attack test of phonemic decoding.

G. DO THE INTERVENTIONS CLOSE THE READING GAP?

The impact estimates show that for most outcomes that measured word-level skills and comprehension, third graders in one of the four interventions had better reading scores than the control students who received their ordinary instruction. For fifth graders, impacts of the four interventions combined were found only for Word Attack and Sight Word Efficiency. To assess the extent to which the interventions helped to close the reading gap during the period of the intervention, we assess how much smaller the gap is for students in the interventions than for students in the control group at the end of the school year. Our standard for determining each group's reading gap is the score (of 100) for an average reader in the national population of students. Thus, the gap for the control group, for example, is 100 minus the average standard score for the group. If the average score is 90, the gap is 100 - 90 = 10. The reading gap describes the extent to which the average student in one of the two evaluation groups (intervention or control) is lagging behind the average student in the population.

On most outcomes, the average student in our evaluation was between one-half and one standard deviation—about 7 to 15 standard score points—below the population average before the interventions started (see Figures IV.2-IV.13 and Table IV.31).[47] By the end of the school year when the interventions

[46] Information on students' eligibility for free or reduced-price school lunch was generally obtained from school records. See Appendix C for more details.

[47] In terms of percentiles, the average student in our evaluation was at about the 31st percentile on a measure such as Word Attack and the 18th percentile on the GRADE test.

had ended, third-grade students in the control group were still generally between one-half and one standard deviation below the population average, while fifth-grade students in the control group were about one-third to three-quarters of a standard deviation below.

Reflecting the estimated pattern of impacts, the gaps at the end of the school year for students in the interventions were smaller than those for the students in the control group, although as noted above, only some of the impacts are statistically significant. To quantify the effect of the interventions on closing the gap, we computed a statistic that shows the reduction in the gap due to the interventions relative to the size of the gap for the control group at the end of the school year.[48]

Table IV.31 shows that the gap for third-grade students in the control group in phonemic decoding skills on the Phonemic Decoding Efficiency subtest of the TOWRE, for example, is about 11 standard score points at the end of the third grade (100 - 89). Students in the intervention group had an average standard score that was about 8 points below the population mean (100 - 92). The 3 point difference in the reading gap for those in the intervention and control groups represents the impact of the interventions and shows that being in one of the interventions reduced the gap by about one-quarter (3/11 = 0.27). As another example, the almost 5 point impact on the GRADE, which is a measure of reading comprehension, also results in a gap reduction of about 25 percent. The result for GRADE is particularly interesting because third graders in the control group lost ground relative to the national average between the beginning and the end of the school year, which increased the gap in reading comprehension for these struggling readers. Students in the intervention group did not fall farther behind and, thus, the end of the year reading gap was smaller by about 5 points. However, despite this effect of the interventions, the average student in the interventions was approximately 14 standard score points below the average student in the nation at the end of the year. Results for the other outcomes show that the largest reduction in the reading gap for third graders occurred on the Word Attack test (69 percent reduction). On the tests for other word-level skills and reading comprehension, the interventions reduced the gap by about one-fifth or one-quarter after one year.

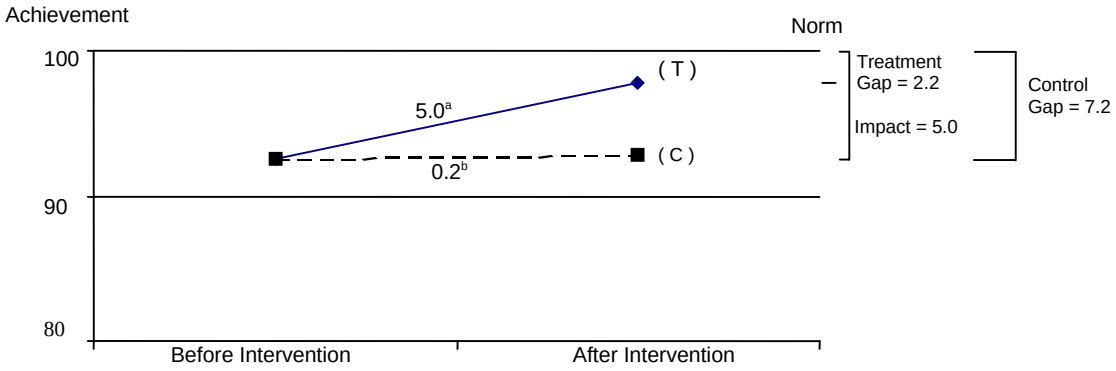
For fifth graders, the interventions reduced the gap by more than 50 percent on Word Attack and by about 12 percent on Sight Word Efficiency. For most of the other outcomes, for which impacts were not statistically significant, negligible reductions were observed. At the end of the school year, the gap for the average intervention student was approximately 2 points for Word Attack, 10 points for Sight Word Efficiency, and 8 points for the GRADE test of reading comprehension.[49]

[48] The relative gap reduction due to the intervention was computed as: RGR=[(100-Mean for Control Group)-(100-Mean for Treatment Group at Follow-up)]/(100-Mean for Control Group at Follow-up)=IMPACT/(100-Mean for Control Group at Follow-up), where 100 is the mean for the normed population.

[49] These analyses examine whether the interventions closed the gap for the average student in the interventions. In future analyses, we plan to explore another approach for estimating the impact of the interventions on closing the reading gap. This approach will contrast the percentage of students in the intervention groups and the control groups who scored within the "normal range" on the standardized tests

Figure IV.2

Third-Grade Gains in Word Attack



[a]Treatment group gain
[b]Control group gain

Figure IV.3

Third-Grade Gains in Phonemic Decoding Efficiency

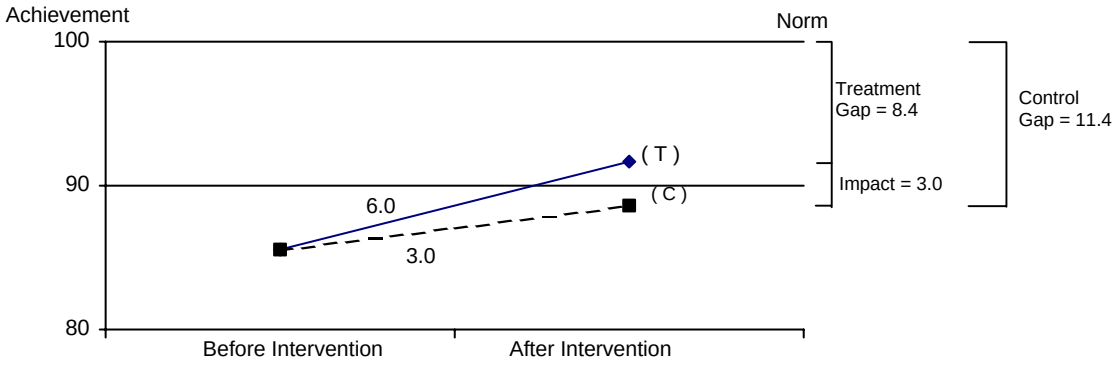


Figure IV.4

Third-Grade Gains in Word Identification

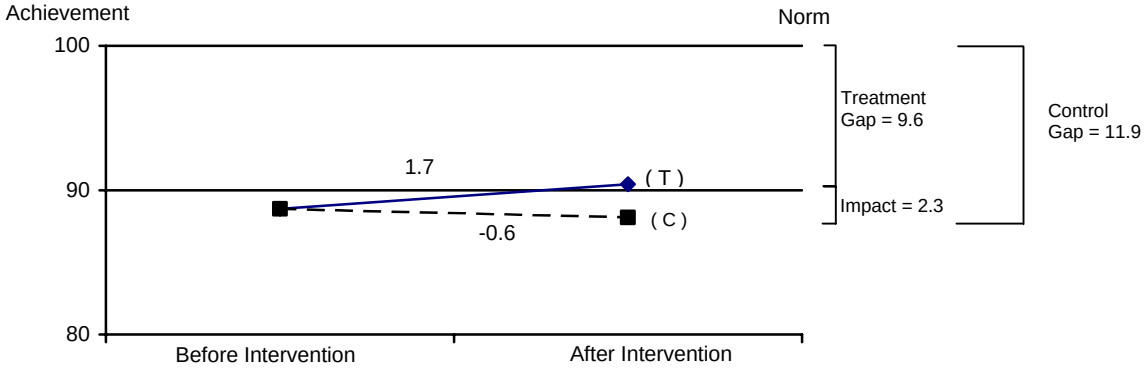


Figure IV.5

Third-Grade Gains in Sight Word Efficiency

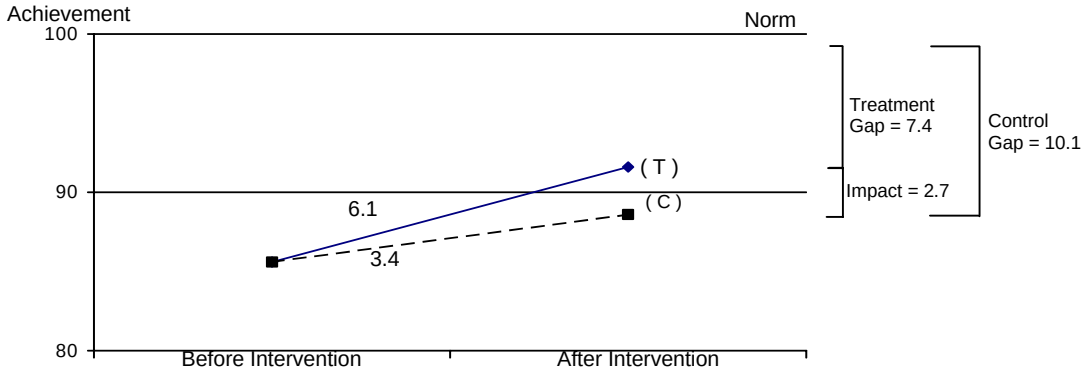


Figure IV.6

Third-Grade Gains in Passage Comprehension

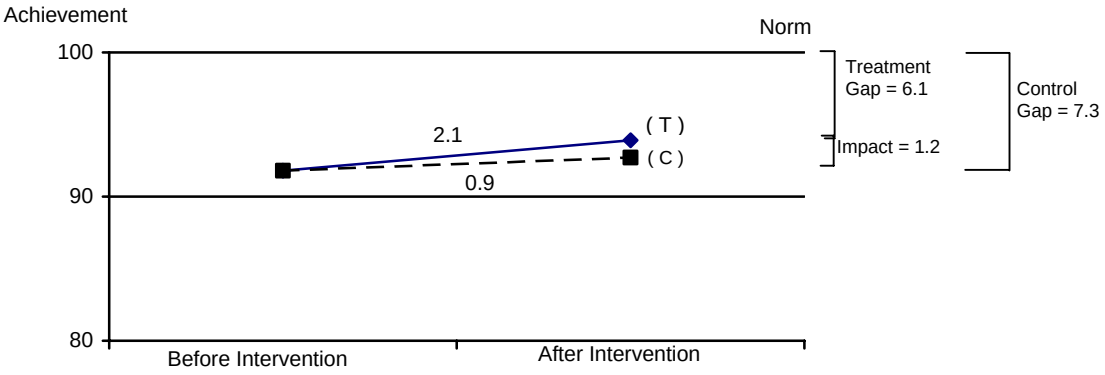


Figure IV.7

Third-Grade Gains in GRADE Test

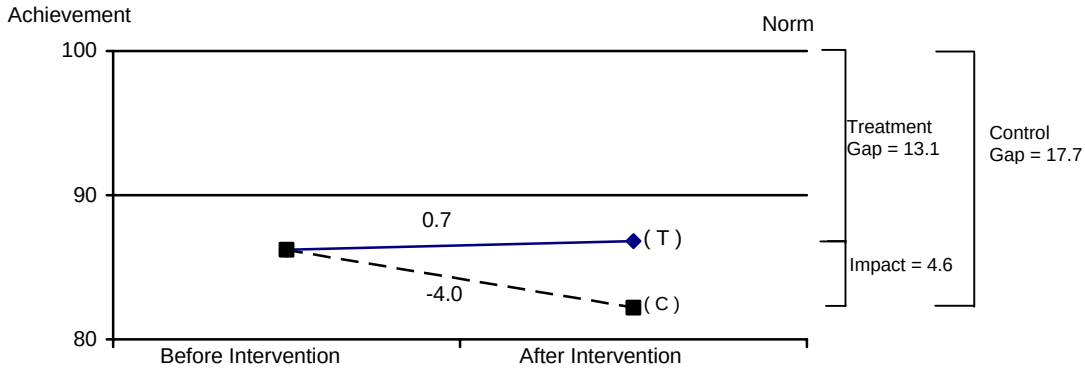


Figure IV.8

Fifth-Grade Gains in Word Attack

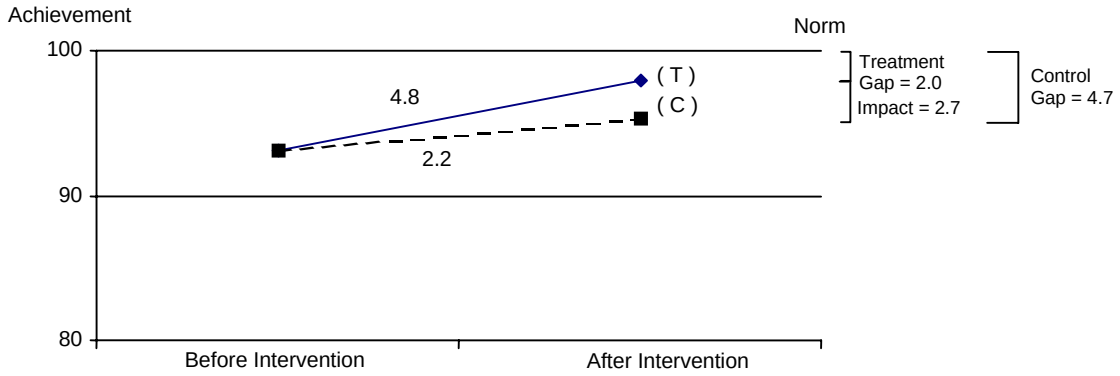


Figure IV.9

Fifth-Grade Gains in Phonemic Decoding Efficiency

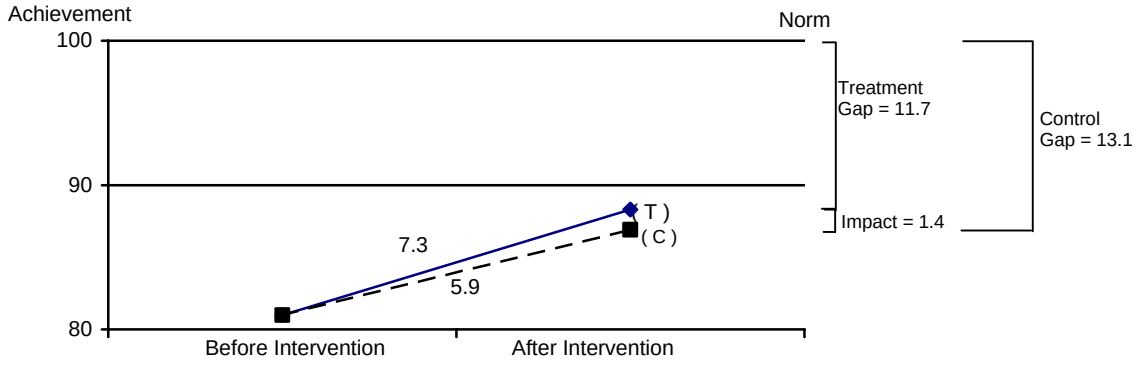


Figure IV.10

Fifth-Grade Gains in Word Identification

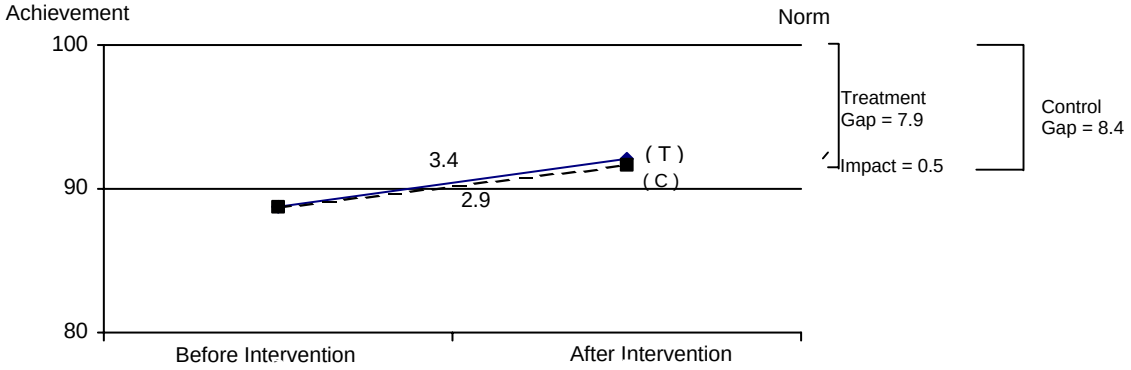


Figure IV.11

Fifth-Grade Gains in Sight Word Efficiency

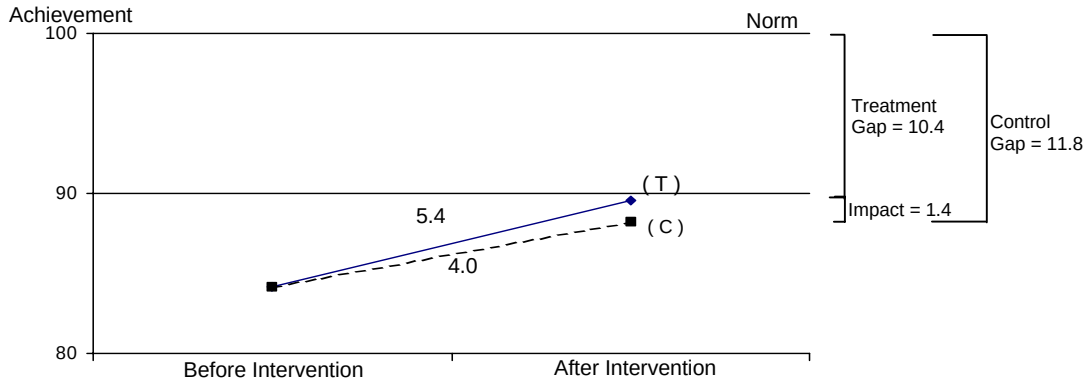


Figure IV.12

Fifth-Grade Gains in Passage Comprehension

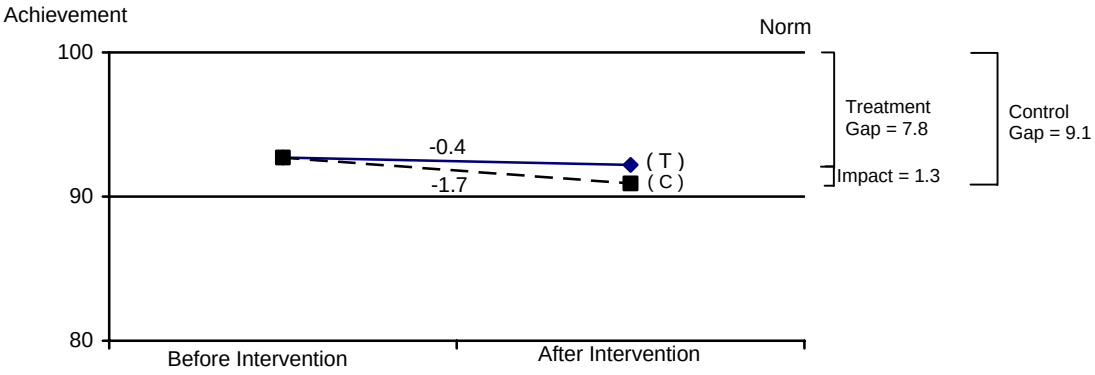


Figure IV.13

Fifth-Grade Gains in GRADE Test

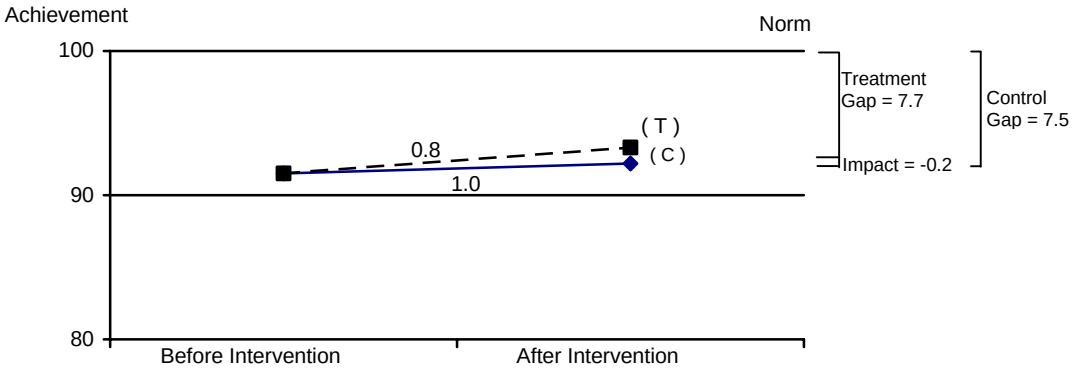


Table IV.1

Impacts for 3rd and 5th Graders

		All Interventions		Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading			Corrective Reading	
		Control	ABCD	Control	BCD	Control	A	Control	B	Control	C		Control	D
Grade 3	Baseline	Gain	Impact	Gain	Impact	Gain	Impact	Gain	Impact	Gain	Impact		Gain	Impact
Word Attack	92.6	0.2	5.0 *	0.0	6.8 *	0.7	-0.5	2.5	6.5 *	-3.0	8.8 *		0.5	5.2 *
TOWRE PDE	85.6	3.0	3.0 *	2.6	4.4 *	4.1	-1.3	4.1	7.1 *	0.2	5.8 *		3.6	0.4
Word Identification	88.7	-0.6	2.3 *	-0.6	2.6 *	-0.5	1.3	0.4	2.0	-2.3	2.5		0.1	3.3 *
TOWRE SWE	86.5	3.4	2.7 *	3.6	2.8 *	2.9	2.6	4.9	0.7	3.5	3.1		2.4	4.6 *
Aimsweb	40.9	20.6	4.9 *	20.3	5.9 *	21.5	1.9	22.6	1.0	17.5	6.0		20.9	10.7 *
Passage Comprehension	91.8	0.9	1.2	1.5	0.7	-0.8	2.7	2.4	0.2	-0.5	1.0		2.6	0.9
GRADE	86.2	-4.0	4.6 *	-3.1	4.4	-6.5	5.3	-4.2	4.9	-4.3	4.2		-0.9	4.2
Sample Size	335													

		All Interventions		Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading			Corrective Reading	
		Control	ABCD	Control	BCD	Control	A	Control	B	Control	C		Control	D
Grade 5	Baseline	Gain	Impact	Gain	Impact	Gain	Impact	Gain	Impact	Gain	Impact		Gain	Impact
Word Attack	93.1	2.2	2.7 *	2.4	3.9 *	1.3	-0.9	3.2	5.3 *	2.0	4.4 *		2.1	1.9
TOWRE PDE	81.0	5.9	1.4	6.3	1.5	4.6	1.1	7.9	4.1 *	6.8	-1.4	#	4.3	1.9
Word Identification	88.7	2.9	0.5	2.8	0.9	3.1	-0.6	2.8	0.1	2.6	2.1		3.1	0.3
TOWRE SWE	84.2	4.0	1.4 *	4.5	1.3	2.4	1.7	5.6	2.1	4.6	-0.5		3.4	2.2
Aimsweb	77.4	19.1	2.0	18.7	2.8	20.5	-0.3	19.6	3.6	19.4	-0.1		17.1	4.9
Passage Comprehension	92.7	-1.7	1.3	-2.1	1.6	-0.6	0.3	-1.2	0.6	-3.7	2.5		-1.4	1.8
GRADE	91.5	1.0	-0.2	0.8	0.3	1.6	-1.6	-0.5	-0.7	-0.7	1.3		3.6	0.3
Sample Size	407													

* Impact is statistically significant at the 0.05 level.

Impact is statistically different from the 3rd grade impact at the 0.05 level.

Table IV.2

Impacts for 3rd and 5th Graders with Low Baseline Word Attack Scores

		All Interventions		Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 3	Baseline	Control Gain	ABCD Impact	Control Gain	BCD Impact	Control Gain	A Impact	Control Gain	B Impact	Control Gain	C Impact	Control Gain	D Impact
Word Attack	84.6	2.8	4.5 *	2.7	6.5 *	3.0	-1.5	5.9	6.0	-0.6	7.8 *	2.8	5.9
TOWRE PDE	82.1	2.3	3.3 *	1.5	5.4 *	4.5	-2.7	4.0	7.8 *	-0.9	5.7	1.4	2.5
Word Identification	85.2	-0.4	1.7 *	-0.5	2.1 *	0.0	0.6	1.3	0.6	-3.8	3.6 *	0.9	2.1
TOWRE SWE	82.8	3.2	2.1	3.0	2.4	3.8	1.1	5.4	-0.8	0.9	4.3	2.8	3.6
Aimsweb	31.9	21.0	1.6	20.9	2.0	21.3	0.4	22.9	-5.3	15.0	7.2	24.7	4.2
Passage Comprehension	86.8	2.2	1.3	2.2	1.3	2.4	1.5	3.8	1.7	-1.7	3.1	4.4	-1.0
GRADE	83.0	-6.5	6.7 *	-6.3	7.1 *	-7.1	5.5	-7.5	7.3	-6.7	4.6	-4.7	9.4
Sample Size	173												

		All Interventions			Word-level interventions		Failure Free Reading			Spell Read			Wilson Reading		Corrective Reading	
Grade 5	Baseline	Control Gain	ABCD Impact		Control Gain	BCD Impact	Control Gain	A Impact		Control Gain	B Impact		Control Gain	C Impact	Control Gain	D Impact
Word Attack	84.5	4.6	3.3 *		4.9	4.7 *	3.7	-0.8		3.3	8.0 *		6.9	2.9	4.5	3.2
TOWRE PDE	75.5	6.4	1.6		7.1	1.8	4.0	1.0		7.5	6.6 *		9.0	-3.0	4.9	1.9
Word Identification	84.1	2.6	1.7 *	#	2.7	1.6 *	2.2	2.0	#	3.5	0.1		2.4	2.1	2.1	2.7
TOWRE SWE	81.3	3.7	1.3		4.3	1.4	1.8	0.7		5.4	2.7		5.8	-2.2	1.8	3.8 *
Aimsweb	67.4	20.3	-1.4	#	20.1	-1.0	20.8	-2.7		26.0	-5.8	#	17.0	-1.7	17.3	4.5
Passage Comprehension	89.2	-0.8	1.0		-1.3	1.6	0.5	-0.8		0.0	1.0		-1.6	0.9	-2.2	2.8
GRADE	88.2	0.9	1.2		1.4	0.9	-0.5	2.1		0.1	0.1		-0.6	4.1	4.5	-1.4
Sample Size	201															

* Impact is statistically significant at the 0.05 level.

Impact is statistically different from the impact for all students in that grade at the 0.05 level.

Table IV.3

Impacts for 3rd and 5th Graders with high baseline Word Attack scores

		All Interventions			Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 3	Baseline	Control Gain	ABCD Impact		Control Gain	BCD Impact	Control Gain	A Impact	Control Gain	B Impact	Control Gain	C Impact	Control Gain	D Impact
Word Attack	101.1	-1.8	4.3 *		-1.9	5.9 *	-1.4	-0.3	-1.2	6.2 *	-3.3	6.8 *	-1.3	4.6
TOWRE PDE	89.3	3.7	2.8 *		3.2	3.8 *	5.0	-0.2	4.6	5.9 *	0.4	5.7 *	4.6	-0.4
Word Identification	92.5	-1.3	3.0 *		-1.5	3.2 *	-0.7	2.3	-0.1	1.6	-2.7	2.9	-1.7	5.0 *
TOWRE SWE	90.4	5.0	2.1 *		5.6	1.4	3.3	4.4 *	5.3	0.9	7.0	0.3	4.4	3.0
Aimsweb	50.4	22.4	3.9		21.2	5.3	26.0	-0.3	22.5	2.4	17.2	6.1	23.9	7.5
Passage Comprehension	97.2	1.8	-1.2		3.5	-2.7	-3.1	3.2	1.0	-1.0	7.9	-8.3 *	1.5	1.1
GRADE	89.8	-0.1	0.7		0.6	0.0	-2.1	2.6	0.8	1.1	0.3	-0.4	0.7	-0.5
Sample Size	162													

		All Interventions			Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 5	Baseline	Control Gain	ABCD Impact		Control Gain	BCD Impact	Control Gain	A Impact	Control Gain	B Impact	Control Gain	C Impact	Control Gain	D Impact
Word Attack	101.7	0.5	1.6		0.9	2.5	-0.8	-1.0	3.1	2.8	0.1	2.7	-0.4	1.9
TOWRE PDE	86.3	5.8	1.3		6.0	1.1	5.2	2.0	8.3	1.9	5.3	0.3	4.5	1.2
Word Identification	93.2	3.3	-0.6	#	3.2	0.0	3.8	-2.4 #	2.5	0.2	3.5	1.1	3.5	-1.3
TOWRE SWE	87.1	4.4	1.7		5.0	1.1	2.4	3.5 *	6.1	2.1	4.4	0.7	4.6	0.6
Aimsweb	87.3	17.1	5.3 *	#	16.7	5.5 *	18.3	4.7	12.7	12.3 * #	21.6	-0.6	15.8	4.9
Passage Comprehension	96.1	-1.7	0.3		-1.9	0.3	-1.1	0.6	-2.2	0.3	-3.6	1.6	0.1	-1.2
GRADE	94.7	1.2	-1.7		0.1	-0.3	4.6	-5.8	-3.1	1.3	-0.3	-2.6	3.7	0.4
Sample Size	206													

* Impact is statistically significant at the 0.05 level.

Impact is statistically different from the impact for all students in that grade at the 0.05 level.

Table IV.4

Impacts for 3rd and 5th Graders with Low Screening Peabody Picture Vocabulary Test Scores

		All Interventions		Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 3	Baseline	Control Gain	ABCD Impact	Control Gain	BCD Impact	Control Gain	A Impact	Control Gain	B Impact	Control Gain	C Impact	Control Gain	D Impact
Word Attack	92.1	-0.4	4.9 *	-0.2	6.1 *	-1.0	1.2	3.0	7.5 *	-1.9	6.6 *	-1.6	4.2
TOWRE PDE	85.3	3.3	2.3	4.0	2.5	1.0	1.5	6.9	4.1	2.4	3.0	2.8	0.5
Word Identification	87.9	-0.1	1.3	0.1	1.3	-0.4	1.1	1.9	0.0	-0.6	0.6	-1.1	3.5 *
TOWRE SWE	86.1	3.2	3.6 *	3.0	4.1 *	3.6	2.2	3.6	2.2	3.3	4.9	2.1	5.1
Aimsweb	38.9	21.8	0.5	22.2	0.0	20.6	2.1	25.0	-2.4	19.1	0.9	22.6	1.3
Passage Comprehension	90.0	0.5	1.4	1.4	0.7	-2.4	3.4	3.9	-0.6	1.5	-1.6	-1.1	4.4
GRADE	83.6	-5.5	5.2 *	-4.9	5.2	-7.5	4.9	-4.0	4.1	-6.6	3.6	-4.0	8.0
Sample Size	148												

		All Interventions		Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 5	Baseline	Control Gain	ABCD Impact	Control Gain	BCD Impact	Control Gain	A Impact	Control Gain	B Impact	Control Gain	C Impact	Control Gain	D Impact
Word Attack	91.9	1.2	4.0 *	1.4	5.6 *	0.6	-0.7	4.8	4.5 *	1.0	6.5 *	-1.5	5.7 *
TOWRE PDE	80.1	5.3	2.6 *	6.1	2.5	2.9	3.0	9.2	3.3 *	4.6	2.0	4.4	2.1
Word Identification	86.8	1.6	1.5	1.3	2.4 *	2.4	-1.2	3.2	-0.8	0.1	5.1 * #	0.7	2.9
TOWRE SWE	83.5	2.5	3.9 * #	2.8	4.0 * #	1.7	3.7	4.7	3.4	3.2	2.8	0.4	5.7 *
Aimsweb	72.5	20.4	1.0	19.7	1.3	22.5	-0.1	18.1	6.0	18.5	1.8	22.4	-3.8
Passage Comprehension	89.6	-1.5	1.9	-1.7	2.3	-0.8	0.8	0.1	1.4	-1.1	1.4	-4.1	4.3
GRADE	87.5	-2.0	1.2	-2.0	1.7	-1.9	-0.1	-1.2	0.4	-1.1	-1.8	-3.7	6.3
Sample Size	200												

* Impact is statistically significant at the 0.05 level.

Impact is statistically different from the impact for all students in that grade at the 0.05 level.

Table IV.5

Impacts for 3rd and 5th Graders with High screening Peabody Picture Vocabulary Test Scores

		All Interventions			Word-level interventions			Failure Free Reading		Spell Read		Wilson Reading			Corrective Reading	
Grade 3	Baseline	Control Gain	ABCD Impact		Control Gain	BCD Impact		Control Gain	A Impact	Control Gain	B Impact	Control Gain	C Impact		Control Gain	D Impact
Word Attack	93.0	0.4	5.5 *		0.2	7.4 *		0.9	-0.2	0.9	7.6 *	-4.4	11.6 *		4.1	3.1
TOWRE PDE	85.8	3.5	2.6		2.8	4.5 *		5.8	-3.1	2.0	9.2 *	2.2	4.2		4.2	0.1
Word Identification	89.4	-0.8	2.5 *		-1.1	3.1 *		0.2	0.7	-0.5	2.6	-3.8	4.5 *		0.9	2.3
TOWRE SWE	86.8	3.2	2.8 *		3.7	2.3		1.5	4.3	4.8	0.9	3.1	2.4		3.3	3.7
Aimsweb	42.5	19.5	6.7 *		18.6	8.5 *		22.4	1.5	20.9	1.7	14.1	9.4		20.8	14.2 *
Passage Comprehension	93.2	0.1	2.0		0.3	1.7		-0.3	3.0	1.1	1.1	-2.7	2.6		2.4	1.3
GRADE	88.3	-3.6	5.6 *		-2.3	4.9		-7.3	7.6	-4.3	5.5	-4.3	7.4		1.6	1.7
Sample Size	187															

		All Interventions			Word-level interventions			Failure Free Reading		Spell Read		Wilson Reading			Corrective Reading	
Grade 5	Baseline	Control Gain	ABCD Impact		Control Gain	BCD Impact		Control Gain	A Impact	Control Gain	B Impact	Control Gain	C Impact		Control Gain	D Impact
Word Attack	94.2	1.5	3.0 *		1.7	4.1 *		0.9	-0.2	2.0	6.6 *	1.0	3.9		2.1	1.8
TOWRE PDE	81.7	5.0	1.5		5.0	1.9		5.2	0.4	4.8	6.3 *	6.5	-3.0		3.5	2.3
Word Identification	90.2	3.0	0.3		2.9	0.3		3.1	0.5	2.1	0.5	3.7	0.1	#	3.0	0.2
TOWRE SWE	84.8	4.4	0.3	#	4.9	0.2	#	2.6	0.6	6.3	1.5	4.2	-1.6		4.3	0.6
Aimsweb	81.5	19.2	1.2		19.2	1.8		19.0	-0.6	21.9	-0.4	19.4	-1.6		16.4	7.5
Passage Comprehension	95.3	-3.1	2.4 *		-3.9	2.9 *		-0.6	0.8	-2.2	0.2	-9.1	6.0 *		-0.4	2.4
GRADE	94.8	1.3	0.3		1.1	0.8		2.0	-1.1	-1.6	0.6	-0.6	3.7		5.3	-2.1
Sample Size	207															

* Impact is statistically significant at the 0.05 level.

Impact is statistically different from the impact for all students in that grade at the 0.05 level.

Table IV.6

Impacts for 3rd and 5th Graders with Low Baseline Word Attack and Low Screening Peabody Picture Vocabulary Test Scores

		All Interventions		Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 3	Baseline	Control Gain	ABCD Impact	Control Gain	BCD Impact	Control Gain	A Impact	Control Gain	B Impact	Control Gain	C Impact	Control Gain	D Impact
Word Attack	85.5	1.1	3.5	1.4	5.0	0.2	-1.1	1.4	11.5 *	3.6	-2.6	-0.6	6.0
TOWRE PDE	82.1	1.4	2.7	1.9	3.6	-0.1	0.2	6.7	4.4	-1.7	3.3	0.6	3.0
Word Identification	85.0	0.0	0.4	0.0	0.3	-0.2	0.7	2.6	-1.6	-1.1	-0.9	-1.4	3.4
TOWRE SWE	82.5	4.0	2.1	4.5	1.8	2.5	3.0	5.8	-0.3	5.6	0.7	2.1	5.1
Aimsweb	31.8	21.4	-0.9	20.6	-0.1	23.8	-3.5	22.4	-4.0	25.1	-5.6	14.3	9.3
Passage Comprehension	85.5	1.5	1.4	2.9	0.1	-2.6	5.5	7.0	-1.9	2.5	-2.5	-1.0	4.6
GRADE	81.7	-7.4	6.4 *	-6.3	5.7	-11.0	8.4	-5.1	4.0	-3.6	0.8	-10.0	12.2 *
Sample Size	81												

		All Interventions			Word-level interventions			Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 5	Baseline	Control Gain	ABCD Impact		Control Gain	BCD Impact		Control Gain	A Impact	Control Gain	B Impact	Control Gain	C Impact	Control Gain	D Impact
Word Attack	83.7	3.3	3.8		2.8	5.9 *		4.6	-2.6	2.8	8.3 *	3.7	3.9	2.0	5.5
TOWRE PDE	74.7	4.6	4.1 *		6.1	4.2		0.2	3.8	7.5	6.1 *	6.5	1.8	4.2	4.6
Word Identification	82.6	2.8	0.9		3.1	0.5		1.9	2.0	3.9	-1.0	2.8	0.9	2.5	1.7
TOWRE SWE	80.0	4.3	3.6 *		5.2	3.5		1.3	4.0	5.3	4.0	6.2	2.7	4.2	3.8
Aimsweb	62.7	25.8	-7.7	#	25.2	-7.4	#	27.8	-8.9	25.3	-4.9	20.9	-4.6	29.4	-12.5 #
Passage Comprehension	86.0	-3.6	5.6		-5.2	7.7		1.5	-0.8	-1.5	4.4	-6.1	7.4	-8.2	11.3
GRADE	84.4	-3.8	3.1		-4.6	4.2		-1.1	-0.1	-4.4	4.1	-2.7	-0.2	-6.8	8.5
Sample Size	111														

* Impact is statistically significant at the 0.05 level.

Impact is statistically different from the impact for all students in that grade at the 0.05 level.

Table IV.7

Impacts for 3rd and 5th Graders With Low Baseline Word Attack and high Peabody Picture Vocabulary Test Scores

		All Interventions		Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading			Corrective Reading	
		Control	ABCD	Control	BCD	Control	A	Control	B	Control	C		Control	D
Grade 3	Baseline	Gain	Impact	Gain	Impact	Gain	Impact	Gain	Impact	Gain	Impact		Gain	Impact
Word Attack	83.9	5.2	4.0	5.8	5.3 *	3.5	0.2	7.3	4.4	-1.6	13.1 *		11.8	-1.7
TOWRE PDE	82.1	3.8	2.3	3.1	4.4	5.8	-4.1	2.7	8.6 *	2.0	4.8		4.6	-0.1
Word Identification	85.4	0.1	1.7	0.0	2.4	0.6	-0.4	1.2	0.5	-4.7	5.7 *		3.4	0.9
TOWRE SWE	83.1	2.1	2.4	1.9	2.6	2.8	1.7	4.4	0.1	-2.2	6.4		3.4	1.4
Aimsweb	32.1	21.3	1.4	21.1	1.9	21.8	-0.4	23.8	-7.6	13.0	10.4		26.4	3.0
Passage Comprehension	87.9	2.7	2.5	1.9	3.0	5.0	1.0	3.8	2.6	-4.4	8.5	#	6.3	-2.2
GRADE	84.1	-4.9	7.4 *	-4.5	7.5 *	-5.9	6.9	-6.2	5.3	-13.0	16.2 *	#	5.7	1.1
Sample Size	92													

		All Interventions		Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading			Corrective Reading	
		Control	ABCD	Control	BCD	Control	A	Control	B	Control	C		Control	D
Grade 5	Baseline	Gain	Impact	Gain	Impact	Gain	Impact	Gain	Impact	Gain	Impact		Gain	Impact
Word Attack	85.4	4.5	3.5 *	4.8	4.7 *	3.7	0.1	4.1	8.2 *	6.2	3.5		4.0	2.4
TOWRE PDE	76.3	6.3	0.6	6.5	1.3	5.9	-1.4	7.2	8.2 *	7.7	-4.8		4.5	0.3
Word Identification	85.6	2.7	1.4	2.9	1.2	2.2	2.1	3.6	0.1	2.9	0.9		2.3	2.6
TOWRE SWE	82.6	3.3	-0.1	3.8	0.2	1.7	-0.9	4.6	3.4	4.7	-5.0		2.0	2.2
Aimsweb	72.2	18.3	0.6	17.8	2.0	20.0	-3.9	25.3	-3.1	14.1	-0.3		13.8	9.4
Passage Comprehension	92.4	-1.6	1.0	-2.1	1.3	-0.2	0.0	-0.1	-0.2	-6.6	3.5		0.3	0.6
GRADE	92.1	1.3	2.5	2.1	1.6	-1.2	5.2	1.6	-0.7	-2.5	9.1		7.0	-3.6
Sample Size	90													

* Impact is statistically significant at the 0.05 level.

Impact is statistically different from the impact for all students in that grade at the 0.05 level.

Table IV.8

Impacts for 3rd and 5th Graders with High Baseline Word Attack and High Screening Peabody Picture Vocabulary Test Scores

		All Interventions			Word-level interventions			Failure Free Reading		Spell Read			Wilson Reading		Corrective Reading	
Grade 3	Baseline	Control Gain	ABCD Impact		Control Gain	BCD Impact		Control Gain	A Impact	Control Gain	B Impact		Control Gain	C Impact	Control Gain	D Impact
Word Attack	101.7	-4.1	6.4 *		-4.9	8.9 *		-1.8	-1.2	-7.4	12.6 *	#	-5.6	8.0 *	-1.8	6.2 *
TOWRE PDE	89.4	4.0	2.5		3.3	4.0		6.3	-2.0	3.0	7.3 *		1.2	6.0	5.6	-1.3
Word Identification	93.2	-1.2	2.8		-1.8	3.3		0.4	1.4	0.2	1.3		-3.9	3.4	-1.7	5.0
TOWRE SWE	90.3	6.7	0.5		8.1	-1.0		2.4	4.9	7.4	-2.1		11.5	-3.5	# 5.4	2.7
Aimsweb	52.5	20.1	6.9		18.0	8.9		26.7	1.0	20.3	2.4		10.9	12.0	22.6	12.2
Passage Comprehension	98.3	3.0	-3.5		5.1	-5.8 *	#	-3.5	3.4	0.2	-1.0		15.5	-18.7 *	# -0.2	2.2
GRADE	92.3	1.8	-1.8	#	3.8	-3.5	#	-4.4	3.1	4.1	-4.2	#	7.3	-4.1	0.0	-2.0
Sample Size	95															

		All Interventions			Word-level interventions			Failure Free Reading		Spell Read			Wilson Reading		Corrective Reading	
Grade 5	Baseline	Control Gain	ABCD Impact		Control Gain	BCD Impact		Control Gain	A Impact	Control Gain	B Impact		Control Gain	C Impact	Control Gain	D Impact
Word Attack	101.6	-0.6	3.1 *		-0.5	4.1 *		-0.8	0.1	0.0	6.8 *		-1.9	3.2	0.4	2.2
TOWRE PDE	86.3	4.0	3.5 *		4.0	3.8 *		4.0	2.4	3.2	5.8		5.4	2.6	3.5	3.1
Word Identification	94.1	3.6	-0.8		3.2	-0.8		4.6	-0.7	1.4	0.7		4.0	-0.7	4.2	-2.2 #
TOWRE SWE	86.7	4.6	1.9		5.7	1.4		1.3	3.6	4.9	2.7		6.0	2.3	6.1	-1.0 #
Aimsweb	89.4	18.1	4.1		18.7	3.8		16.1	5.0	13.8	7.1		25.6	-1.0	16.7	5.4
Passage Comprehension	97.7	-3.6	2.4		-4.2	2.6		-1.8	2.1	-5.9	3.0		-5.8	2.6	-0.9	2.0
GRADE	97.1	1.6	-1.1		0.4	0.9		5.1	-7.3	-5.0	4.7		1.0	-1.4	5.1	-0.6
Sample Size	117															

* Impact is statistically significant at the 0.05 level.

Impact is statistically different from the impact for all students in that grade at the 0.05 level.

Table IV.9

Impacts for 3rd and 5th Graders Eligible for Free or Reduced Price School Lunch

		All Interventions			Word-level interventions			Failure Free Reading		Spell Read			Wilson Reading			Corrective Reading		
Grade 3	Baseline	Control Gain	ABCD Impact		Control Gain	BCD Impact		Control Gain	A Impact	Control Gain	B Impact		Control Gain	C Impact		Control Gain	D Impact	
Word Attack	92.2	1.3	4.7 *		1.6	5.9 *		0.7	1.3	1.7	8.4 *		0.2	6.0 *	#	2.8	3.3	
TOWRE PDE	85.3	4.6	1.8		4.5	2.6	#	4.9	-0.7	5.1	6.2 *		1.9	3.6	#	6.5	-2.0	
Word Identification	88.0	0.2	1.1		0.3	1.1		-0.2	1.0	2.3	-0.6		-1.4	1.2		0.0	2.8	
TOWRE SWE	85.5	3.5	1.3		4.0	0.7		2.2	3.0	4.1	-0.8		3.9	2.5		3.9	0.4	#
Aimsweb	38.6	20.3	2.0		19.6	3.1		22.5	-1.1	22.0	-1.9		16.1	6.4		20.7	4.7	
Passage Comprehension	90.4	3.3	-0.8	#	4.2	-1.2	#	0.7	0.4	3.5	0.5		4.5	-2.6	#	4.5	-1.5	
GRADE	84.4	-2.0	0.1	#	-0.7	-0.8	#	-6.0	2.5	-2.6	1.6		-1.4	-2.1	#	1.8	-1.7	
Sample Size	193																	

		All Interventions			Word-level interventions			Failure Free Reading		Spell Read			Wilson Reading			Corrective Reading		
Grade 5	Baseline	Control Gain	ABCD Impact		Control Gain	BCD Impact		Control Gain	A Impact	Control Gain	B Impact		Control Gain	C Impact		Control Gain	D Impact	
Word Attack	92.5	3.5	0.6		4.1	1.5		1.7	-2.3	5.7	0.8	#	3.7	3.0		2.8	0.8	
TOWRE PDE	80.1	6.5	0.6		6.6	1.0		6.2	-0.5	8.9	2.9		7.2	-1.2		3.8	1.3	
Word Identification	87.8	2.4	0.4		2.5	0.7		2.2	-0.4	2.5	-1.2		2.1	3.0 *		3.0	0.2	
TOWRE SWE	83.2	2.6	3.7 *	#	2.9	3.8 *	#	1.6	3.2	4.5	3.9 *		4.1	1.0		0.3	6.5 *	#
Aimsweb	73.4	14.7	3.1		14.0	4.5		16.6	-1.1	16.0	8.6 *		13.7	0.7		12.4	4.4	
Passage Comprehension	90.6	-0.1	-0.3		-0.3	-0.1		0.5	-0.9	0.6	-0.8		-0.8	-0.8		-0.8	1.3	
GRADE	88.6	3.2	-4.1 *	#	3.1	-3.7		3.3	-5.4	4.9	-6.1 *		1.0	-4.2		3.3	-0.8	
Sample Size	230																	

* Impact is statistically significant at the 0.05 level.

Impact is statistically different from the impact for all students in that grade at the 0.05 level.

Table IV.10

Impacts for 3rd and 5th Graders not Eligible for Free or Reduced Price School Lunch

		All Interventions		Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 3	Baseline	Control Gain	ABCD Impact	Control Gain	BCD Impact	Control Gain	A Impact	Control Gain	B Impact	Control Gain	C Impact	Control Gain	D Impact
Word Attack	93.3	-2.7	7.8 *	-3.8	10.9 *	0.7	-1.7	0.8	8.3 *	-13.2	19.5 * #	0.9	5.0
TOWRE PDE	86.1	0.1	5.3 *	-1.2	8.0 * #	4.1	-3.1	4.8	6.2 *	-12.1	17.6 * #	3.7	0.3
Word Identification	89.9	-2.4	3.6 *	-3.1	4.6 *	-0.2	0.5	-1.1	2.4	-7.8	7.8	-0.3	3.6
TOWRE SWE	87.9	3.0	3.0 *	2.6	3.9 *	4.1	0.2	6.8	-0.5	-0.1	5.2	1.1	6.9 * #
Aimsweb	44.1	19.0	7.6 *	19.0	8.4 *	19.1	5.1	23.1	1.1	13.0	9.6	20.9	14.5 *
Passage Comprehension	93.8	-5.0	6.1 * #	-5.9	6.7 * #	-2.1	4.2	2.7	-2.8	-20.9	19.5 * #	0.5	3.6
GRADE	88.9	-8.6	9.5 * #	-8.9	10.6 * #	-7.5	6.4	-5.5	6.0	-17.9	19.2 * #	-3.4	6.6
Sample Size	142												

		All Interventions		Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 5	Baseline	Control Gain	ABCD Impact	Control Gain	BCD Impact	Control Gain	A Impact	Control Gain	B Impact	Control Gain	C Impact	Control Gain	D Impact
Word Attack	94.0	1.4	3.7 *	1.5	5.1 *	0.9	-0.5	1.3	8.9 * #	1.4	4.1	1.9	2.2
TOWRE PDE	82.0	5.3	1.2	6.1	1.0	3.0	1.6	6.3	4.8 *	6.9	-2.1	5.0	0.5
Word Identification	89.7	3.6	0.0	3.1	0.5	4.8	-1.6	2.5	0.9	3.8	0.5	3.1	0.0
TOWRE SWE	85.4	4.8	0.0 #	5.7	-0.7 #	1.9	2.0	5.3	1.1	5.0	-0.4	6.8	-2.8 #
Aimsweb	82.2	22.1	0.3	21.7	0.2	23.5	0.5	21.0	-0.7	22.0	0.0	22.0	1.4
Passage Comprehension	95.1	-2.9	2.1	-3.2	2.4	-1.9	1.4	-2.4	1.3	-6.9	5.3 *	-0.3	0.5
GRADE	94.9	0.3	1.2 #	-0.2	1.9	1.9	-0.7	-4.5	1.8	0.1	2.8	3.8	1.0
Sample Size	177												

* Impact is statistically significant at the 0.05 level.

Impact is statistically different from the impact for all students in that grade at the 0.05 level.

Table IV.11

Effect Sizes for 3rd and 5th Graders

Grade 3	All interventions ABCD Effect Size	Word-level interventions BCD Effect Size	Failure Free Reading A Effect Size	Spell Read B Effect Size	Wilson Reading C Effect Size	Corrective Reading D Effect Size
Word Attack	0.33 *	0.45 *	-0.04	0.43 *	0.59 *	0.35 *
TOWRE PDE	0.20 *	0.29 *	-0.09	0.47 *	0.39 *	0.03
Word Identification	0.15 *	0.17 *	0.09	0.13	0.16	0.22 *
TOWRE SWE	0.18 *	0.18 *	0.17	0.04	0.20	0.30 *
Aimsweb	0.12 *	0.15 *	0.05	0.03	0.15	0.27 *
Passage Comprehension	0.08	0.05	0.18	0.02	0.07	0.06
GRADE	0.31 *	0.29	0.35	0.33	0.28	0.28

Grade 5	All interventions ABCD Effect Size	Word-level interventions BCD Effect Size	Failure Free Reading A Effect Size	Spell Read B Effect Size	Wilson Reading C Effect Size	Corrective Reading D Effect Size
Word Attack	0.18 *	0.26 *	-0.06	0.35 *	0.29 *	0.12
TOWRE PDE	0.09	0.10	0.07	0.28 *	-0.10 #	0.13
Word Identification	0.03	0.06	-0.04	0.01	0.14	0.02
TOWRE SWE	0.09 *	0.08	0.11	0.14	-0.03	0.15
Aimsweb	0.04	0.06	-0.01	0.08	0.00	0.10
Passage Comprehension	0.09	0.11	0.02	0.04	0.17	0.12
GRADE	-0.01	0.02	-0.11	-0.05	0.09	0.02

Note: Population standard deviation = 15 for all tests except AimsWeb
AimsWeb SD (Fall) 3rd grade = 39.2; AimsWeb SD (fall) 5th grade = 47

* Impact is statistically significant at the 0.05 level.

Impact is statistically different from the 3rd grade impact at the 0.05 level

Table IV.12

Effect Sizes for 3rd and 5th Graders With Low Baseline Word Attack Scores

Grade 3	All interventions ABCD Effect Size	Word-level interventions BCD Effect Size	Failure Free Reading A Effect Size	Spell Read B Effect Size	Wilson Reading C Effect Size	Corrective Reading D Effect Size
Word Attack	0.30 *	0.44 *	-0.10	0.40	0.52 *	0.39
TOWRE PDE	0.22 *	0.36 *	-0.18	0.52 *	0.38	0.17
Word Identification	0.11 *	0.14 *	0.04	0.04	0.24 *	0.14
TOWRE SWE	0.14	0.16	0.07	-0.05	0.29	0.24
Aimsweb	0.04	0.05	0.01	-0.13	0.18	0.11
Passage Comprehension	0.09	0.09	0.10	0.11	0.21	-0.06
GRADE	0.45 *	0.47 *	0.37	0.48	0.31	0.63

Grade 5	All interventions ABCD Effect Size	Word-level interventions BCD Effect Size	Failure Free Reading A Effect Size	Spell Read B Effect Size	Wilson Reading C Effect Size	Corrective Reading D Effect Size
Word Attack	0.22 *	0.31 *	-0.05	0.54 *	0.20	0.21
TOWRE PDE	0.11	0.12	0.07	0.44 *	-0.20	0.13
Word Identification	0.12 * #	0.11 *	0.13 #	0.01	0.14	0.18
TOWRE SWE	0.08	0.10	0.05	0.18	-0.15	0.25 *
Aimsweb	-0.03 #	-0.02	-0.06	-0.12 #	-0.04	0.09
Passage Comprehension	0.07	0.11	-0.06	0.07	0.06	0.19
GRADE	0.08	0.06	0.14	0.01	0.27	-0.09

Note: Population standard deviation = 15 for all tests except AimsWeb
AimsWeb SD (Fall) 3rd grade = 39.2; AimsWeb SD (fall) 5th grade = 47

* Impact statistically significant at the 0.05 level.

Impact is statistically different from the overall impact for that grade at the 0.05 level

Table IV.13

Effect Sizes for 3rd and 5th Graders with High Baseline Word Attack Scores

	All interventions		Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
	ABCD		BCD		A		B		C		D	
Grade 3	Effect Size		Effect Size		Effect Size		Effect Size		Effect Size		Effect Size	
Word Attack	0.29		0.39		-0.02		0.41		0.45		0.31	
TOWRE PDE	0.19		0.25		-0.01		0.40		0.38		-0.03	
Word Identification	0.20		0.21		0.15		0.11		0.19		0.34	
TOWRE SWE	0.14		0.09		0.29		0.06		0.02		0.20	
Aimsweb	0.10		0.14		-0.01		0.06		0.15		0.19	
Passage Comprehension	-0.08		-0.18		0.21		-0.06		-0.56		0.07	
GRADE	0.05		0.00		0.18		0.07		-0.03		-0.04	
	All interventions		Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
	ABCD		BCD		A		B		C		D	
Grade 5	Effect Size		Effect Size		Effect Size		Effect Size		Effect Size		Effect Size	
Word Attack	0.11		0.17		-0.07		0.19		0.18		0.13	
TOWRE PDE	0.09		0.08		0.13		0.13		0.02		0.08	
Word Identification	-0.04	#	0.00		-0.16	#	0.02		0.07		-0.08	#
TOWRE SWE	0.12		0.08		0.24		0.14		0.04		0.04	
Aimsweb	0.11	#	0.12		0.10		0.26	#	-0.01		0.10	
Passage Comprehension	0.02		0.02		0.04		0.02		0.11		-0.08	
GRADE	-0.11		-0.02		-0.39		0.09		-0.17		0.02	

Note: Population standard deviation = 15 for all tests except AimsWeb
AimsWeb SD (Fall) 3rd grade = 39.2; AimsWeb SD (fall) 5th grade = 47

* Impact statistically significant at the 0.05 level.

Impact is statistically different from the overall impact for that grade at the 0.05 level

Table IV.14

Effect Sizes for 3rd and 5th Graders with Low Peabody Picture Vocabular Test Scores

Grade 3	All interventions ABCD Effect Size	Word-level interventions BCD Effect Size	Failure Free Reading A Effect Size	Spell Read B Effect Size	Wilson Reading C Effect Size	Corrective Reading D Effect Size
Word Attack	0.33 *	0.41 *	0.08	0.50 *	0.44 *	0.28
TOWRE PDE	0.15	0.17	0.10	0.27	0.20	0.03
Word Identification	0.09	0.09	0.08	0.00	0.04	0.23 *
TOWRE SWE	0.24 *	0.27 *	0.15	0.15	0.33	0.34
Aimsweb	0.01	0.00	0.05	-0.06	0.02	0.03
Passage Comprehension	0.09	0.05	0.23	-0.04	-0.10	0.29
GRADE	0.34 *	0.35	0.32	0.27	0.24	0.53

Grade 5	All interventions ABCD Effect Size	Word-level interventions BCD Effect Size	Failure Free Reading A Effect Size	Spell Read B Effect Size	Wilson Reading C Effect Size	Corrective Reading D Effect Size
Word Attack	0.27 *	0.37 *	-0.05	0.30 *	0.43 *	0.38 *
TOWRE PDE	0.17 *	0.16	0.20	0.22 *	0.13	0.14
Word Identification	0.10	0.16 *	-0.08	-0.05	0.34 * #	0.19
TOWRE SWE	0.26 * #	0.26 * #	0.25	0.23	0.19	0.38 *
Aimsweb	0.02	0.03	0.00	0.13	0.04	-0.08
Passage Comprehension	0.13	0.16	0.05	0.09	0.09	0.28
GRADE	0.08	0.11	-0.01	0.03	-0.12	0.42

Note: Population standard deviation = 15 for all tests except AimsWeb
AimsWeb SD (Fall) 3rd grade = 39.2; AimsWeb SD (fall) 5th grade = 47

* Impact statistically significant at the 0.05 level.

Impact is statistically different from the overall impact for that grade at the 0.05 level

Table IV.15

Effect Sizes for 3rd and 5th Graders with High Screening Peabody Picture Vocabulary Test Scores

Grade 3	All interventions ABCD Effect Size	Word-level interventions BCD Effect Size	Failure Free Reading A Effect Size	Spell Read B Effect Size	Wilson Reading C Effect Size	Corrective Reading D Effect Size
Word Attack	0.37 *	0.50 *	-0.02	0.51 *	0.78 *	0.20
TOWRE PDE	0.17	0.30 *	-0.21	0.61 *	0.28	0.00
Word Identification	0.17 *	0.21 *	0.05	0.17	0.30 *	0.15
TOWRE SWE	0.19 *	0.16	0.29	0.06	0.16	0.25
Aimsweb	0.17 *	0.22 *	0.04	0.04	0.24	0.36 *
Passage Comprehension	0.13	0.11	0.20	0.07	0.17	0.09
GRADE	0.37 *	0.33	0.51	0.37	0.50	0.11

Grade 5	All interventions ABCD Effect Size	Word-level interventions BCD Effect Size	Failure Free Reading A Effect Size	Spell Read B Effect Size	Wilson Reading C Effect Size	Corrective Reading D Effect Size
Word Attack	0.20 *	0.27 *	-0.01	0.44 *	0.26	0.12
TOWRE PDE	0.10	0.12	0.03	0.42 *	-0.20	0.15
Word Identification	0.02	0.02	0.04	0.04	0.01 #	0.02
TOWRE SWE	0.02 #	0.01 #	0.04	0.10	-0.11	0.04
Aimsweb	0.03	0.04	-0.01	-0.01	-0.03	0.16
Passage Comprehension	0.16 *	0.19 *	0.05	0.01	0.40 *	0.16
GRADE	0.02	0.05	-0.07	0.04	0.25	-0.14

Note: Population standard deviation = 15 for all tests except AimsWeb
AimsWeb SD (Fall) 3rd grade = 39.2; AimsWeb SD (fall) 5th grade = 47

* Impact statistically significant at the 0.05 level.

Impact is statistically different from the overall impact for that grade at the 0.05 level

Table IV.16

Effect Sizes for 3rd and 5th Graders with Low Baseline Word Attack and Low Screening PPVT Test Scores

Grade 3	All interventions ABCD Effect Size	Word-level interventions BCD Effect Size	Failure Free Reading A Effect Size	Spell Read B Effect Size	Wilson Reading C Effect Size	Corrective Reading D Effect Size
Word Attack	0.23	0.33	-0.07	0.77 *	-0.18	0.40
TOWRE PDE	0.18	0.24	0.01	0.29	0.22	0.20
Word Identification	0.03	0.02	0.04	-0.11	-0.06	0.23
TOWRE SWE	0.14	0.12	0.20	-0.02	0.04	0.34
Aimsweb	-0.02	0.00	-0.09	-0.10	-0.14	0.24
Passage Comprehension	0.09	0.00	0.37	-0.13	-0.17	0.31
GRADE	0.42 *	0.38	0.56	0.27	0.05	0.82 *

Grade 5	All interventions ABCD Effect Size	Word-level interventions BCD Effect Size	Failure Free Reading A Effect Size	Spell Read B Effect Size	Wilson Reading C Effect Size	Corrective Reading D Effect Size
Word Attack	0.25	0.39 *	-0.17	0.55 *	0.26	0.36
TOWRE PDE	0.27 *	0.28	0.25	0.41 *	0.12	0.31
Word Identification	0.06	0.04	0.13	-0.07	0.06	0.12
TOWRE SWE	0.24 *	0.23	0.27	0.26	0.18	0.26
Aimsweb	-0.16 #	-0.16 #	-0.19	-0.11	-0.10	-0.27 #
Passage Comprehension	0.37	0.51	-0.05	0.29	0.49	0.75
GRADE	0.21	0.28	0.00	0.28	-0.01	0.57

Note: Population standard deviation = 15 for all tests except AimsWeb
AimsWeb SD (Fall) 3rd grade = 39.2; AimsWeb SD (fall) 5th grade = 47

* Impact statistically significant at the 0.05 level.

Impact is statistically different from the overall impact for that grade at the 0.05 level

Table IV.17

Effect Sizes for 3rd and 5th Graders with Low Baseline Word Attack and High Screening PVVT Test Scores

	All interventions	Word-level interventions	Failure Free Reading	Spell Read	Wilson Reading	Corrective Reading
Grade 3	ABCD Effect Size	BCD Effect Size	A Effect Size	B Effect Size	C Effect Size	D Effect Size
Word Attack	0.27	0.35 *	0.01	0.30	0.88 *	-0.12
TOWRE PDE	0.15	0.29	-0.27	0.57 *	0.32	-0.01
Word Identification	0.11	0.16	-0.03	0.04	0.38 *	0.06
TOWRE SWE	0.16	0.17	0.11	0.00	0.42	0.09
Aimsweb	0.03	0.05	-0.01	-0.19	0.27	0.08
Passage Comprehension	0.17	0.20	0.07	0.17	0.57 #	-0.15
GRADE	0.49 *	0.50 *	0.46	0.36	1.08 * #	0.07

	All interventions	Word-level interventions	Failure Free Reading	Spell Read	Wilson Reading	Corrective Reading
Grade 5	ABCD Effect Size	BCD Effect Size	A Effect Size	B Effect Size	C Effect Size	D Effect Size
Word Attack	0.24 *	0.31 *	0.00	0.54 *	0.23	0.16
TOWRE PDE	0.04	0.08	-0.09	0.55 *	-0.32	0.02
Word Identification	0.09	0.08	0.14	0.01	0.06	0.17
TOWRE SWE	-0.01	0.01	-0.06	0.22	-0.33	0.14
Aimsweb	0.01	0.04	-0.08	-0.07	-0.01	0.20
Passage Comprehension	0.06	0.09	0.00	-0.01	0.24	0.04
GRADE	0.17	0.11	0.35	-0.04	0.61	-0.24

Note: Population standard deviation = 15 for all tests except AimsWeb
AimsWeb SD (Fall) 3rd grade = 39.2; AimsWeb SD (fall) 5th grade = 47

* Impact statistically significant at the 0.05 level.

Impact is statistically different from the overall impact for that grade at the 0.05 level

Table IV.18

Effect Sizes for 3rd and 5th Graders with High Baseline Word Attack and High Screening PVVT Test Scores

Grade 3	All interventions ABCD Effect Size	Word-level interventions BCD Effect Size	Failure Free Reading A Effect Size	Spell Read B Effect Size	Wilson Reading C Effect Size	Corrective Reading D Effect Size
Word Attack	0.43 *	0.60 *	-0.08	0.84 * #	0.53 *	0.41 *
TOWRE PDE	0.17	0.27	-0.13	0.49 *	0.40	-0.08
Word Identification	0.19	0.22	0.09	0.09	0.23	0.33
TOWRE SWE	0.03	-0.07	0.33	-0.14	-0.23 #	0.18
Aimsweb	0.18	0.23	0.03	0.06	0.31	0.31
Passage Comprehension	-0.23	-0.39 * #	0.23	-0.07	-1.25 * #	0.15
GRADE	-0.12 #	-0.23 #	0.21	-0.28 #	-0.27	-0.14

Grade 5	All interventions ABCD Effect Size	Word-level interventions BCD Effect Size	Failure Free Reading A Effect Size	Spell Read B Effect Size	Wilson Reading C Effect Size	Corrective Reading D Effect Size
Word Attack	0.20 *	0.27 *	0.00	0.45 *	0.21	0.15
TOWRE PDE	0.23 *	0.25 *	0.16	0.38	0.17	0.21
Word Identification	-0.05	-0.05	-0.05	0.04	-0.05	-0.15 #
TOWRE SWE	0.13	0.09	0.24	0.18	0.15	-0.06 #
Aimsweb	0.09	0.08	0.11	0.15	-0.02	0.12
Passage Comprehension	0.16	0.17	0.14	0.20	0.17	0.14
GRADE	-0.08	0.06	-0.49	0.31	-0.09	-0.04

Note: Population standard deviation = 15 for all tests except AimsWeb
AimsWeb SD (Fall) 3rd grade = 39.2; AimsWeb SD (fall) 5th grade = 47

* Impact statistically significant at the 0.05 level.

Impact is statistically different from the overall impact for that grade at the 0.05 level

Table IV.19

Effect Sizes for 3rd and 5th Graders Eligible for Free or Reduced Price School Lunch

	All interventions		Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
	ABCD		BCD		A		B		C		D	
Grade 3	Effect Size		Effect Size		Effect Size		Effect Size		Effect Size		Effect Size	
Word Attack	0.32	*	0.39	*	0.09		0.56	*	0.40	* #	0.22	
TOWRE PDE	0.12		0.17	#	-0.05		0.41	*	0.24	#	-0.13	
Word Identification	0.07		0.08		0.07		-0.04		0.08		0.19	
TOWRE SWE	0.08		0.05		0.20		-0.05		0.17		0.02	#
Aimsweb	0.05		0.08		-0.03		-0.05		0.16		0.12	
Passage Comprehension	-0.05	#	-0.08	#	0.03		0.04		-0.17	#	-0.10	
GRADE	0.00	#	-0.05	#	0.17		0.11		-0.14	#	-0.11	

	All interventions		Word-level interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
	ABCD		BCD		A		B		C		D	
Grade 5	Effect Size		Effect Size		Effect Size		Effect Size		Effect Size		Effect Size	
Word Attack	0.04		0.10		-0.15		0.05	#	0.20		0.05	
TOWRE PDE	0.04		0.07		-0.03		0.19		-0.08		0.09	
Word Identification	0.03		0.04		-0.03		-0.08		0.20	*	0.01	
TOWRE SWE	0.24	* #	0.25	* #	0.22		0.26	*	0.07		0.44	* #
Aimsweb	0.07		0.10		-0.02		0.18	*	0.01		0.09	
Passage Comprehension	-0.02		-0.01		-0.06		-0.05		-0.06		0.09	
GRADE	-0.28	* #	-0.25		-0.36		-0.41	*	-0.28		-0.06	

Note: Population standard deviation = 15 for all tests except AimsWeb
AimsWeb SD (Fall) 3rd grade = 39.2; AimsWeb SD (fall) 5th grade = 47

* Impact statistically significant at the 0.05 level.

Impact is statistically different from the overall impact for that grade at the 0.05 level

Table IV.20

Effect Sizes for 3rd and 5th Graders not Eligible for Free or Reduced Price School Lunch

	All interventions	Word-level interventions	Failure Free Reading	Spell Read	Wilson Reading	Corrective Reading
	ABCD	BCD	A	B	C	D
Grade 3	Effect Size	Effect Size	Effect Size	Effect Size	Effect Size	Effect Size
Word Attack	0.52 *	0.73 *	-0.12	0.55 *	1.30 * #	0.33
TOWRE PDE	0.35 *	0.54 * #	-0.20	0.42 *	1.17 * #	0.02
Word Identification	0.24 *	0.31 *	0.03	0.16	0.52	0.24
TOWRE SWE	0.20 *	0.26 *	0.02	-0.03	0.34	0.46 * #
Aimsweb	0.19 *	0.21 *	0.13	0.03	0.25	0.37 *
Passage Comprehension	0.41 * #	0.45 * #	0.28	-0.19	1.30 * #	0.24
GRADE	0.64 * #	0.70 * #	0.43	0.40	1.28 * #	0.44

	All interventions	Word-level interventions	Failure Free Reading	Spell Read	Wilson Reading	Corrective Reading
	ABCD	BCD	A	B	C	D
Grade 5	Effect Size	Effect Size	Effect Size	Effect Size	Effect Size	Effect Size
Word Attack	0.25 *	0.34 *	-0.03	0.59 * #	0.27	0.15
TOWRE PDE	0.08	0.07	0.10	0.32 *	-0.14	0.03
Word Identification	0.00	0.03	-0.11	0.06	0.03	0.00
TOWRE SWE	0.00 #	-0.04 #	0.13	0.07	-0.02	-0.18 #
Aimsweb	0.01	0.01	0.01	-0.02	0.00	0.03
Passage Comprehension	0.14	0.16	0.09	0.09	0.36 *	0.04
GRADE	0.08 #	0.12	-0.05	0.12	0.19	0.07

Note: Population standard deviation = 15 for all tests except AimsWeb
AimsWeb SD (Fall) 3rd grade = 39.2; AimsWeb SD (fall) 5th grade = 47

* Impact statistically significant at the 0.05 level.

Impact is statistically different from the overall impact for that grade at the 0.05 level

Table IV.21

Effect of the Treatment on the Treated Impacts for 3rd and 5th Graders

	All interventions			Word-level interventions			Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading		
	ABCD			BCD			A			B			C			D		
Grade 3	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80
Word Attack	5.0	5.0	5.4	6.8	6.9	7.4	-0.5	-0.5	-0.6	6.5	6.5	7.0	8.8	8.9	9.5	5.2	5.2	5.6
TOWRE PDE	3.0	3.0	3.2	4.4	4.5	4.8	-1.3	-1.3	-1.4	7.1	7.2	7.7	5.8	5.8	6.2	0.4	0.4	0.4
Word Identification	2.3	2.3	2.5	2.6	2.6	2.8	1.3	1.4	1.5	2.0	2.0	2.2	2.5	2.5	2.7	3.3	3.3	3.6
TOWRE SWE	2.7	2.8	2.9	2.8	2.8	3.0	2.6	2.6	2.8	0.7	0.7	0.7	3.1	3.1	3.3	4.6	4.6	4.9
Aimsweb	4.9	4.9	5.3	5.9	5.9	6.4	1.9	1.9	2.0	1.0	1.0	1.1	6.0	6.0	6.4	10.7	10.8	11.6
Passage Comprehension	1.2	1.2	1.3	0.7	0.7	0.8	2.7	2.8	2.9	0.2	0.3	0.3	1.0	1.0	1.1	0.9	0.9	0.9
GRADE	4.6	4.7	5.0	4.4	4.5	4.8	5.3	5.3	5.7	4.9	4.9	5.3	4.2	4.2	4.5	4.2	4.2	4.5

	All interventions			Word-level interventions			Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading		
	ABCD			BCD			A			B			C			D		
Grade 5	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80
Word Attack	2.7	2.7	2.9	3.9	3.9	4.2	-0.9	-0.9	-1.0	5.3	5.3	5.7	4.4	4.4	4.8	1.9	1.9	2.0
TOWRE PDE	1.4	1.4	1.5	1.5	1.5	1.7	1.1	1.1	1.2	4.1	4.2	4.5	-1.4	-1.4	-1.5	1.9	1.9	2.1
Word Identification	0.5	0.5	0.5	0.9	0.9	0.9	-0.6	-0.6	-0.7	0.1	0.1	0.1	2.1	2.1	2.3	0.3	0.3	0.4
TOWRE SWE	1.4	1.4	1.5	1.3	1.3	1.4	1.7	1.7	1.8	2.1	2.1	2.2	-0.5	-0.5	-0.5	2.2	2.2	2.4
Aimsweb	2.0	2.0	2.2	2.8	2.8	3.0	-0.3	-0.3	-0.4	3.6	3.6	3.9	-0.1	-0.1	-0.1	4.9	4.9	5.3
Passage Comprehension	1.3	1.3	1.4	1.6	1.6	1.7	0.3	0.3	0.3	0.6	0.6	0.7	2.5	2.5	2.7	1.8	1.8	1.9
GRADE	-0.2	-0.2	-0.2	0.3	0.3	0.3	-1.6	-1.6	-1.7	-0.7	-0.7	-0.8	1.3	1.3	1.4	0.3	0.3	0.3

Note: Instruction amounts are Any (99.2%), Over 80 hours (92.5%)

Table IV.22

Effect of the Treatment on the Treated Impacts for 3rd and 5th Graders with Low Baseline Word Attack Scores

	All interventions			Word-level interventions			Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading		
	ABCD			BCD			A			B			C			D		
Grade 3	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80
Word Attack	4.5	4.6	4.9	6.5	6.6	7.1	-1.5	-1.5	-1.6	6.0	6.0	6.5	7.8	7.8	8.4	5.9	5.9	6.4
TOWRE PDE	3.3	3.4	3.6	5.4	5.4	5.8	-2.7	-2.7	-2.9	7.8	7.9	8.5	5.7	5.8	6.2	2.5	2.5	2.7
Word Identification	1.7	1.7	1.8	2.1	2.1	2.3	0.6	0.6	0.6	0.6	0.6	0.6	3.6	3.6	3.9	2.1	2.1	2.2
TOWRE SWE	2.1	2.1	2.2	2.4	2.4	2.6	1.1	1.1	1.2	-0.8	-0.8	-0.8	4.3	4.3	4.7	3.6	3.7	3.9
Aimsweb	1.6	1.7	1.8	2.0	2.1	2.2	0.4	0.4	0.5	-5.3	-5.3	-5.7	7.2	7.3	7.8	4.2	4.2	4.5
Passage Comprehension	1.3	1.4	1.5	1.3	1.3	1.4	1.5	1.5	1.6	1.7	1.7	1.8	3.1	3.2	3.4	-1.0	-1.0	-1.0
GRADE	6.7	6.8	7.2	7.1	7.1	7.7	5.5	5.6	6.0	7.3	7.3	7.9	4.6	4.6	5.0	9.4	9.5	10.2

	All interventions			Word-level interventions			Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading		
	ABCD			BCD			A			B			C			D		
Grade 5	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80
Word Attack	3.3	3.4	3.6	4.7	4.7	5.1	-0.8	-0.8	-0.8	8.0	8.1	8.7	2.9	3.0	3.2	3.2	3.2	3.4
TOWRE PDE	1.6	1.6	1.7	1.8	1.8	2.0	1.0	1.0	1.1	6.6	6.7	7.1	-3.0	-3.1	-3.3	1.9	1.9	2.1
Word Identification	1.7	1.7	1.9	1.6	1.6	1.8	2.0	2.0	2.2	0.1	0.1	0.1	2.1	2.1	2.3	2.7	2.7	2.9
TOWRE SWE	1.3	1.3	1.4	1.4	1.5	1.6	0.7	0.7	0.8	2.7	2.8	3.0	-2.2	-2.2	-2.4	3.8	3.8	4.1
Aimsweb	-1.4	-1.4	-1.5	-1.0	-1.0	-1.1	-2.7	-2.7	-2.9	-5.8	-5.8	-6.3	-1.7	-1.7	-1.8	4.5	4.5	4.8
Passage Comprehension	1.0	1.0	1.1	1.6	1.6	1.7	-0.8	-0.8	-0.9	1.0	1.0	1.1	0.9	0.9	1.0	2.8	2.8	3.0
GRADE	1.2	1.2	1.3	0.9	0.9	1.0	2.1	2.1	2.2	0.1	0.1	0.1	4.1	4.1	4.4	-1.4	-1.4	-1.5

Note: Instruction amounts are Any (99.2%), Over 80 hours (92.5%)

Table IV.23

Effect of the Treatment on the Treated Impacts for 3rd and 5th Graders with High Baseline Word Attack Scores

	All interventions			Word-level interventions			Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading		
	ABCD			BCD			A			B			C			D		
Grade 3	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80
Word Attack	4.3	4.4	4.7	5.9	5.9	6.3	-0.3	-0.3	-0.3	6.2	6.2	6.7	6.8	6.8	7.3	4.6	4.7	5.0
TOWRE PDE	2.8	2.8	3.0	3.8	3.8	4.1	-0.2	-0.2	-0.2	5.9	6.0	6.4	5.7	5.8	6.2	-0.4	-0.4	-0.4
Word Identification	3.0	3.0	3.2	3.2	3.2	3.4	2.3	2.3	2.5	1.6	1.6	1.8	2.9	2.9	3.1	5.0	5.1	5.4
TOWRE SWE	2.1	2.2	2.3	1.4	1.4	1.5	4.4	4.4	4.7	0.9	0.9	1.0	0.3	0.3	0.4	3.0	3.0	3.2
Aimsweb	3.9	3.9	4.2	5.3	5.3	5.7	-0.3	-0.3	-0.3	2.4	2.4	2.6	6.1	6.1	6.5	7.5	7.6	8.1
Passage Comprehension	-1.2	-1.3	-1.3	-2.7	-2.7	-2.9	3.2	3.2	3.5	-1.0	-1.0	-1.0	-8.3	-8.4	-9.0	1.1	1.1	1.2
GRADE	0.7	0.7	0.7	0.0	0.0	0.0	2.6	2.7	2.8	1.1	1.1	1.2	-0.4	-0.4	-0.5	-0.5	-0.6	-0.6

	All interventions			Word-level interventions			Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading		
	ABCD			BCD			A			B			C			D		
Grade 5	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80
Word Attack	1.6	1.6	1.7	2.5	2.5	2.7	-1.0	-1.0	-1.1	2.8	2.9	3.1	2.7	2.8	3.0	1.9	1.9	2.1
TOWRE PDE	1.3	1.4	1.5	1.1	1.2	1.2	2.0	2.0	2.1	1.9	1.9	2.1	0.3	0.3	0.3	1.2	1.2	1.3
Word Identification	-0.6	-0.6	-0.6	0.0	0.0	0.0	-2.4	-2.4	-2.6	0.2	0.2	0.3	1.1	1.1	1.1	-1.3	-1.3	-1.4
TOWRE SWE	1.7	1.7	1.9	1.1	1.1	1.2	3.5	3.6	3.8	2.1	2.2	2.3	0.7	0.7	0.7	0.6	0.6	0.6
Aimsweb	5.3	5.3	5.7	5.5	5.6	6.0	4.7	4.7	5.0	12.3	12.4	13.3	-0.6	-0.6	-0.7	4.9	4.9	5.3
Passage Comprehension	0.3	0.3	0.4	0.3	0.3	0.3	0.6	0.6	0.6	0.3	0.3	0.3	1.6	1.7	1.8	-1.2	-1.2	-1.3
GRADE	-1.7	-1.7	-1.8	-0.3	-0.3	-0.3	-5.8	-5.8	-6.3	1.3	1.3	1.4	-2.6	-2.6	-2.8	0.4	0.4	0.4

Note: Instruction amounts are Any (99.2%), Over 80 hours (92.5%)

Table IV.24

Effect of the Treatment on the Treated Impacts for 3rd and 5th Graders with Low Screening Peabody Picture Vocabulary Test Scores

	All interventions			Word-level interventions			Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading		
	ABCD			BCD			A			B			C			D		
Grade 3	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80
Word Attack	4.9	4.9	5.3	6.1	6.2	6.6	1.2	1.3	1.3	7.5	7.5	8.1	6.6	6.7	7.2	4.2	4.3	4.6
TOWRE PDE	2.3	2.3	2.4	2.5	2.5	2.7	1.5	1.5	1.6	4.1	4.1	4.4	3.0	3.0	3.2	0.5	0.5	0.5
Word Identification	1.3	1.3	1.4	1.3	1.4	1.5	1.1	1.2	1.2	0.0	0.0	0.0	0.6	0.6	0.6	3.5	3.5	3.7
TOWRE SWE	3.6	3.6	3.9	4.1	4.1	4.4	2.2	2.2	2.4	2.2	2.2	2.4	4.9	4.9	5.3	5.1	5.1	5.5
Aimsweb	0.5	0.5	0.5	0.0	0.0	0.0	2.1	2.1	2.3	-2.4	-2.4	-2.5	0.9	0.9	1.0	1.3	1.3	1.4
Passage Comprehension	1.4	1.4	1.5	0.7	0.7	0.8	3.4	3.5	3.7	-0.6	-0.6	-0.7	-1.6	-1.6	-1.7	4.4	4.4	4.7
GRADE	5.2	5.2	5.6	5.2	5.3	5.7	4.9	4.9	5.3	4.1	4.1	4.4	3.6	3.7	3.9	8.0	8.1	8.7

	All interventions			Word-level interventions			Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading		
	ABCD			BCD			A			B			C			D		
Grade 5	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80
Word Attack	4.0	4.0	4.3	5.6	5.6	6.0	-0.7	-0.7	-0.7	4.5	4.6	4.9	6.5	6.6	7.0	5.7	5.7	6.2
TOWRE PDE	2.6	2.6	2.8	2.5	2.5	2.7	3.0	3.0	3.2	3.3	3.3	3.6	2.0	2.0	2.1	2.1	2.1	2.3
Word Identification	1.5	1.5	1.6	2.4	2.4	2.6	-1.2	-1.3	-1.3	-0.8	-0.8	-0.8	5.1	5.2	5.6	2.9	2.9	3.1
TOWRE SWE	3.9	3.9	4.2	4.0	4.0	4.3	3.7	3.8	4.0	3.4	3.4	3.7	2.8	2.8	3.1	5.7	5.7	6.1
Aimsweb	1.0	1.0	1.1	1.3	1.4	1.5	-0.1	-0.1	-0.1	6.0	6.1	6.5	1.8	1.9	2.0	-3.8	-3.8	-4.1
Passage Comprehension	1.9	2.0	2.1	2.3	2.3	2.5	0.8	0.8	0.9	1.4	1.4	1.5	1.4	1.4	1.5	4.3	4.3	4.6
GRADE	1.2	1.2	1.3	1.7	1.7	1.8	-0.1	-0.1	-0.1	0.4	0.4	0.5	-1.8	-1.8	-1.9	6.3	6.4	6.8

Note: Instruction amounts are Any (99.2%), Over 80 hours (92.5%)

Table IV.25

Effect of the Treatment on the Treated Impacts for 3rd and 5th Graders with High Screening Peabody Picture Vocabulary Test Scores

	All interventions			Word-level interventions			Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading		
	ABCD			BCD			A			B			C			D		
Grade 3	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80
Word Attack	5.5	5.6	6.0	7.4	7.5	8.0	-0.2	-0.2	-0.2	7.6	7.7	8.2	11.6	11.7	12.6	3.1	3.1	3.3
TOWRE PDE	2.6	2.6	2.8	4.5	4.5	4.8	-3.1	-3.1	-3.3	9.2	9.2	9.9	4.2	4.2	4.5	0.1	0.1	0.1
Word Identification	2.5	2.5	2.7	3.1	3.1	3.4	0.7	0.7	0.8	2.6	2.6	2.8	4.5	4.5	4.9	2.3	2.3	2.4
TOWRE SWE	2.8	2.9	3.1	2.3	2.3	2.5	4.3	4.4	4.7	0.9	0.9	0.9	2.4	2.4	2.6	3.7	3.8	4.0
Aimsweb	6.7	6.8	7.3	8.5	8.5	9.1	1.5	1.5	1.6	1.7	1.7	1.8	9.4	9.5	10.2	14.2	14.3	15.4
Passage Comprehension	2.0	2.0	2.2	1.7	1.7	1.8	3.0	3.1	3.3	1.1	1.1	1.2	2.6	2.6	2.8	1.3	1.3	1.4
GRADE	5.6	5.6	6.0	4.9	4.9	5.3	7.6	7.7	8.2	5.5	5.6	6.0	7.4	7.5	8.0	1.7	1.7	1.8

	All interventions			Word-level interventions			Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading		
	ABCD			BCD			A			B			C			D		
Grade 5	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80
Word Attack	3.0	3.0	3.3	4.1	4.1	4.4	-0.2	-0.2	-0.2	6.6	6.6	7.1	3.9	4.0	4.2	1.8	1.8	2.0
TOWRE PDE	1.5	1.5	1.6	1.9	1.9	2.0	0.4	0.4	0.4	6.3	6.4	6.8	-3.0	-3.0	-3.2	2.3	2.3	2.5
Word Identification	0.3	0.4	0.4	0.3	0.3	0.3	0.5	0.5	0.6	0.5	0.5	0.6	0.1	0.1	0.1	0.2	0.2	0.2
TOWRE SWE	0.3	0.3	0.3	0.2	0.2	0.2	0.6	0.6	0.7	1.5	1.5	1.6	-1.6	-1.6	-1.8	0.6	0.6	0.7
Aimsweb	1.2	1.2	1.3	1.8	1.8	2.0	-0.6	-0.6	-0.6	-0.4	-0.4	-0.5	-1.6	-1.6	-1.7	7.5	7.6	8.1
Passage Comprehension	2.4	2.4	2.5	2.9	2.9	3.1	0.8	0.8	0.8	0.2	0.2	0.2	6.0	6.1	6.5	2.4	2.4	2.6
GRADE	0.3	0.3	0.3	0.8	0.8	0.8	-1.1	-1.1	-1.2	0.6	0.6	0.7	3.7	3.8	4.0	-2.1	-2.1	-2.3

Note: Instruction amounts are Any (99.2%), Over 80 hours (92.5%)

Table IV.26

Effect of the Treatment on the Treated Impacts for 3rd and 5th Graders with Low Baseline Word Attack Scores and Low Screening Peabody Picture Vocabulary Test Scores

	All interventions			Word-level interventions			Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading		
	ABCD			BCD			A			B			C			D		
Grade 3	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80
Word Attack	3.5	3.5	3.7	5.0	5.0	5.4	-1.1	-1.1	-1.2	11.5	11.6	12.5	-2.6	-2.6	-2.8	6.0	6.1	6.5
TOWRE PDE	2.7	2.7	2.9	3.6	3.6	3.8	0.2	0.2	0.2	4.4	4.4	4.7	3.3	3.4	3.6	3.0	3.0	3.2
Word Identification	0.4	0.4	0.4	0.3	0.3	0.3	0.7	0.7	0.7	-1.6	-1.6	-1.7	-0.9	-0.9	-1.0	3.4	3.5	3.7
TOWRE SWE	2.1	2.1	2.3	1.8	1.8	2.0	3.0	3.0	3.2	-0.3	-0.3	-0.3	0.7	0.7	0.7	5.1	5.1	5.5
Aimsweb	-0.9	-0.9	-1.0	-0.1	-0.1	-0.1	-3.5	-3.5	-3.7	-4.0	-4.0	-4.3	-5.6	-5.6	-6.0	9.3	9.4	10.1
Passage Comprehension	1.4	1.4	1.5	0.1	0.1	0.1	5.5	5.5	5.9	-1.9	-1.9	-2.1	-2.5	-2.5	-2.7	4.6	4.6	5.0
GRADE	6.4	6.4	6.9	5.7	5.7	6.1	8.4	8.5	9.1	4.0	4.1	4.4	0.8	0.8	0.8	12.2	12.3	13.2

	All interventions			Word-level interventions			Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading		
	ABCD			BCD			A			B			C			D		
Grade 5	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80
Word Attack	3.8	3.8	4.1	5.9	5.9	6.4	-2.6	-2.6	-2.8	8.3	8.4	9.0	3.9	3.9	4.2	5.5	5.5	5.9
TOWRE PDE	4.1	4.1	4.4	4.2	4.2	4.5	3.8	3.8	4.1	6.1	6.2	6.6	1.8	1.9	2.0	4.6	4.7	5.0
Word Identification	0.9	0.9	1.0	0.5	0.5	0.6	2.0	2.0	2.2	-1.0	-1.0	-1.1	0.9	0.9	1.0	1.7	1.7	1.9
TOWRE SWE	3.6	3.6	3.9	3.5	3.5	3.8	4.0	4.0	4.3	4.0	4.0	4.3	2.7	2.7	2.9	3.8	3.9	4.2
Aimsweb	-7.7	-7.8	-8.4	-7.4	-7.4	-7.9	-8.9	-9.0	-9.6	-4.9	-5.0	-5.3	-4.6	-4.6	-4.9	-12.5	-12.6	-13.5
Passage Comprehension	5.6	5.6	6.0	7.7	7.8	8.3	-0.8	-0.8	-0.9	4.4	4.4	4.8	7.4	7.5	8.0	11.3	11.4	12.2
GRADE	3.1	3.1	3.4	4.2	4.2	4.5	-0.1	-0.1	-0.1	4.1	4.2	4.5	-0.2	-0.2	-0.2	8.5	8.6	9.2

Note: Instruction amounts are Any (99.2%), Over 80 hours (92.5%)

Table IV.27

Effect of the Treatment on the Treated Impacts for 3rd and 5th Graders with Low Baseline Word Attack and High Screening Peabody Picture Vocabulary Test Scores

	All interventions			Word-level interventions			Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading		
	ABCD			BCD			A			B			C			D		
Grade 3	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80
Word Attack	3.5	3.5	3.7	5.0	5.0	5.4	-1.1	-1.1	-1.2	11.5	11.6	12.5	-2.6	-2.6	-2.8	6.0	6.1	6.5
TOWRE PDE	2.7	2.7	2.9	3.6	3.6	3.8	0.2	0.2	0.2	4.4	4.4	4.7	3.3	3.4	3.6	3.0	3.0	3.2
Word Identification	0.4	0.4	0.4	0.3	0.3	0.3	0.7	0.7	0.7	-1.6	-1.6	-1.7	-0.9	-0.9	-1.0	3.4	3.5	3.7
TOWRE SWE	2.1	2.1	2.3	1.8	1.8	2.0	3.0	3.0	3.2	-0.3	-0.3	-0.3	0.7	0.7	0.7	5.1	5.1	5.5
Aimsweb	-0.9	-0.9	-1.0	-0.1	-0.1	-0.1	-3.5	-3.5	-3.7	-4.0	-4.0	-4.3	-5.6	-5.6	-6.0	9.3	9.4	10.1
Passage Comprehension	1.4	1.4	1.5	0.1	0.1	0.1	5.5	5.5	5.9	-1.9	-1.9	-2.1	-2.5	-2.5	-2.7	4.6	4.6	5.0
GRADE	6.4	6.4	6.9	5.7	5.7	6.1	8.4	8.5	9.1	4.0	4.1	4.4	0.8	0.8	0.8	12.2	12.3	13.2

	All interventions			Word-level interventions			Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading		
	ABCD			BCD			A			B			C			D		
Grade 5	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80
Word Attack	3.8	3.8	4.1	5.9	5.9	6.4	-2.6	-2.6	-2.8	8.3	8.4	9.0	3.9	3.9	4.2	5.5	5.5	5.9
TOWRE PDE	4.1	4.1	4.4	4.2	4.2	4.5	3.8	3.8	4.1	6.1	6.2	6.6	1.8	1.9	2.0	4.6	4.7	5.0
Word Identification	0.9	0.9	1.0	0.5	0.5	0.6	2.0	2.0	2.2	-1.0	-1.0	-1.1	0.9	0.9	1.0	1.7	1.7	1.9
TOWRE SWE	3.6	3.6	3.9	3.5	3.5	3.8	4.0	4.0	4.3	4.0	4.0	4.3	2.7	2.7	2.9	3.8	3.9	4.2
Aimsweb	-7.7	-7.8	-8.4	-7.4	-7.4	-7.9	-8.9	-9.0	-9.6	-4.9	-5.0	-5.3	-4.6	-4.6	-4.9	-12.5	-12.6	-13.5
Passage Comprehension	5.6	5.6	6.0	7.7	7.8	8.3	-0.8	-0.8	-0.9	4.4	4.4	4.8	7.4	7.5	8.0	11.3	11.4	12.2
GRADE	3.1	3.1	3.4	4.2	4.2	4.5	-0.1	-0.1	-0.1	4.1	4.2	4.5	-0.2	-0.2	-0.2	8.5	8.6	9.2

Note: Instruction amounts are Any (99.2%), Over 80 hours (92.5%)

Table IV.28

Effect of the Treatment on the Treated Impacts for 3rd and 5th Graders with High Baseline Word Attack and High Screening Peabody Picture Vocabulary Test Scores

	All interventions			Word-level interventions			Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading		
	ABCD			BCD			A			B			C			D		
Grade 3	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80
Word Attack	6.4	6.5	6.9	8.9	9.0	9.7	-1.2	-1.2	-1.3	12.6	12.7	13.7	8.0	8.1	8.7	6.2	6.3	6.7
TOWRE PDE	2.5	2.6	2.7	4.0	4.1	4.4	-2.0	-2.0	-2.1	7.3	7.4	7.9	6.0	6.1	6.5	-1.3	-1.3	-1.4
Word Identification	2.8	2.8	3.0	3.3	3.3	3.5	1.4	1.4	1.5	1.3	1.3	1.4	3.4	3.5	3.7	5.0	5.1	5.4
TOWRE SWE	0.5	0.5	0.5	-1.0	-1.0	-1.1	4.9	4.9	5.3	-2.1	-2.1	-2.3	-3.5	-3.6	-3.8	2.7	2.7	2.9
Aimsweb	6.9	7.0	7.5	8.9	9.0	9.6	1.0	1.0	1.1	2.4	2.5	2.6	12.0	12.1	13.0	12.2	12.3	13.2
Passage Comprehension	-3.5	-3.5	-3.8	-5.8	-5.9	-6.3	3.4	3.5	3.7	-1.0	-1.0	-1.1	-18.7	-18.9	-20.3	2.2	2.2	2.4
GRADE	-1.8	-1.8	-2.0	-3.5	-3.5	-3.7	3.1	3.1	3.4	-4.2	-4.3	-4.6	-4.1	-4.1	-4.4	-2.0	-2.1	-2.2

	All interventions			Word-level interventions			Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading		
	ABCD			BCD			A			B			C			D		
Grade 5	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80
Word Attack	3.1	3.1	3.3	4.1	4.1	4.4	0.1	0.1	0.1	6.8	6.9	7.4	3.2	3.2	3.4	2.2	2.2	2.4
TOWRE PDE	3.5	3.5	3.7	3.8	3.9	4.1	2.4	2.4	2.6	5.8	5.8	6.2	2.6	2.6	2.8	3.1	3.1	3.4
Word Identification	-0.8	-0.8	-0.8	-0.8	-0.8	-0.8	-0.7	-0.8	-0.8	0.7	0.7	0.7	-0.7	-0.7	-0.8	-2.2	-2.2	-2.4
TOWRE SWE	1.9	1.9	2.1	1.4	1.4	1.5	3.6	3.6	3.9	2.7	2.8	3.0	2.3	2.3	2.5	-1.0	-1.0	-1.0
Aimsweb	4.1	4.2	4.5	3.8	3.9	4.2	5.0	5.0	5.4	7.1	7.2	7.7	-1.0	-1.0	-1.1	5.4	5.5	5.9
Passage Comprehension	2.4	2.5	2.6	2.6	2.6	2.8	2.1	2.1	2.2	3.0	3.1	3.3	2.6	2.6	2.8	2.0	2.1	2.2
GRADE	-1.1	-1.2	-1.2	0.9	0.9	1.0	-7.3	-7.3	-7.9	4.7	4.7	5.1	-1.4	-1.4	-1.5	-0.6	-0.6	-0.6

Note: Instruction amounts are Any (99.2%), Over 80 hours (92.5%)

Table IV.29

Effect of the Treatment on the Treated Impacts for 3rd and 5th Graders Eligible for Free or Reduced Price School Lunch

	All interventions			Word-level interventions			Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading		
	ABCD			BCD			A			B			C			D		
Grade 3	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80
Word Attack	4.7	4.8	5.1	5.9	5.9	6.4	1.3	1.3	1.4	8.4	8.5	9.1	6.0	6.1	6.5	3.3	3.3	3.5
TOWRE PDE	1.8	1.8	1.9	2.6	2.6	2.8	-0.7	-0.8	-0.8	6.2	6.3	6.7	3.6	3.6	3.8	-2.0	-2.0	-2.2
Word Identification	1.1	1.1	1.2	1.1	1.1	1.2	1.0	1.0	1.1	-0.6	-0.6	-0.7	1.2	1.2	1.3	2.8	2.9	3.1
TOWRE SWE	1.3	1.3	1.4	0.7	0.7	0.8	3.0	3.0	3.2	-0.8	-0.8	-0.9	2.5	2.6	2.7	0.4	0.4	0.4
Aimsweb	2.0	2.0	2.2	3.1	3.1	3.3	-1.1	-1.1	-1.1	-1.9	-1.9	-2.0	6.4	6.4	6.9	4.7	4.7	5.1
Passage Comprehension	-0.8	-0.8	-0.9	-1.2	-1.2	-1.3	0.4	0.4	0.4	0.5	0.5	0.6	-2.6	-2.6	-2.8	-1.5	-1.6	-1.7
GRADE	0.1	0.1	0.1	-0.8	-0.8	-0.8	2.5	2.6	2.7	1.6	1.6	1.7	-2.1	-2.2	-2.3	-1.7	-1.7	-1.9

	All interventions			Word-level interventions			Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading		
	ABCD			BCD			A			B			C			D		
Grade 5	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80
Word Attack	0.6	0.6	0.6	1.5	1.6	1.7	-2.3	-2.3	-2.5	0.8	0.8	0.9	3.0	3.1	3.3	0.8	0.8	0.9
TOWRE PDE	0.6	0.6	0.7	1.0	1.0	1.1	-0.5	-0.5	-0.6	2.9	2.9	3.2	-1.2	-1.2	-1.3	1.3	1.3	1.4
Word Identification	0.4	0.4	0.4	0.7	0.7	0.7	-0.4	-0.4	-0.4	-1.2	-1.2	-1.3	3.0	3.0	3.2	0.2	0.2	0.2
TOWRE SWE	3.7	3.7	4.0	3.8	3.8	4.1	3.2	3.3	3.5	3.9	3.9	4.2	1.0	1.0	1.1	6.5	6.6	7.1
Aimsweb	3.1	3.2	3.4	4.5	4.6	4.9	-1.1	-1.1	-1.2	8.6	8.7	9.3	0.7	0.7	0.7	4.4	4.4	4.7
Passage Comprehension	-0.3	-0.3	-0.3	-0.1	-0.1	-0.1	-0.9	-0.9	-0.9	-0.8	-0.8	-0.9	-0.8	-0.8	-0.9	1.3	1.3	1.4
GRADE	-4.1	-4.2	-4.5	-3.7	-3.7	-4.0	-5.4	-5.5	-5.9	-6.1	-6.2	-6.6	-4.2	-4.2	-4.5	-0.8	-0.8	-0.9

Note: Instruction amounts are Any (99.2%), Over 80 hours (92.5%)

Table IV.30

Effect of the Treatment on the Treated Impacts for 3rd and 5th Graders not Eligible for Free or Reduced Price School Lunch

	All interventions			Word-level interventions			Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading		
	ABCD			BCD			A			B			C			D		
Grade 3	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80
Word Attack	7.8	7.8	8.4	10.9	11.0	11.8	-1.7	-1.7	-1.9	8.3	8.3	8.9	19.5	19.7	21.1	5.0	5.1	5.4
TOWRE PDE	5.3	5.3	5.7	8.0	8.1	8.7	-3.1	-3.1	-3.3	6.2	6.3	6.8	17.6	17.7	19.0	0.3	0.3	0.3
Word Identification	3.6	3.6	3.9	4.6	4.6	5.0	0.5	0.5	0.6	2.4	2.5	2.6	7.8	7.8	8.4	3.6	3.6	3.9
TOWRE SWE	3.0	3.0	3.2	3.9	3.9	4.2	0.2	0.2	0.3	-0.5	-0.5	-0.5	5.2	5.2	5.6	6.9	7.0	7.5
Aimsweb	7.6	7.6	8.2	8.4	8.5	9.1	5.1	5.1	5.5	1.1	1.1	1.2	9.6	9.7	10.4	14.5	14.6	15.7
Passage Comprehension	6.1	6.2	6.6	6.7	6.8	7.3	4.2	4.2	4.6	-2.8	-2.8	-3.0	19.5	19.6	21.1	3.6	3.6	3.9
GRADE	9.5	9.6	10.3	10.6	10.7	11.4	6.4	6.4	6.9	6.0	6.0	6.4	19.2	19.3	20.7	6.6	6.7	7.1

	All interventions			Word-level interventions			Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading		
	ABCD			BCD			A			B			C			D		
Grade 5	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80	ITT Impact	TOT Any	TOT Over 80
Word Attack	3.7	3.7	4.0	5.1	5.1	5.5	-0.5	-0.5	-0.5	8.9	9.0	9.6	4.1	4.1	4.4	2.2	2.2	2.4
TOWRE PDE	1.2	1.2	1.3	1.0	1.1	1.1	1.6	1.6	1.7	4.8	4.8	5.2	-2.1	-2.2	-2.3	0.5	0.5	0.5
Word Identification	0.0	0.0	0.0	0.5	0.5	0.5	-1.6	-1.6	-1.7	0.9	0.9	1.0	0.5	0.5	0.5	0.0	0.0	0.1
TOWRE SWE	0.0	0.0	0.0	-0.7	-0.7	-0.7	2.0	2.0	2.1	1.1	1.1	1.2	-0.4	-0.4	-0.4	-2.8	-2.8	-3.0
Aimsweb	0.3	0.3	0.3	0.2	0.2	0.3	0.5	0.5	0.6	-0.7	-0.7	-0.8	0.0	0.0	0.0	1.4	1.4	1.5
Passage Comprehension	2.1	2.2	2.3	2.4	2.4	2.6	1.4	1.4	1.5	1.3	1.3	1.4	5.3	5.4	5.8	0.5	0.5	0.6
GRADE	1.2	1.2	1.3	1.9	1.9	2.0	-0.7	-0.7	-0.8	1.8	1.8	1.9	2.8	2.8	3.0	1.0	1.0	1.1

Note: Instruction amounts are Any (99.2%), Over 80 hours (92.5%)

Table IV.31

Relative Gap Reduction: All Interventions Combined

3rd Grade	Average at baseline	Gap at baseline (Std. Units)	Average at follow-up: Intervention Group	Average at follow-up: Control Group	Gap at follow-up (Std. Units): Intervention Group	Gap at follow-up (Std. Units): Control Group	Impact	RGR
Word Attack	92.6	0.49	97.8	92.8	0.15	0.48	5.0 *	0.69
TOWRE PDE	85.6	0.96	91.6	88.6	0.56	0.76	3.0 *	0.26
Word Identification	88.7	0.75	90.4	88.1	0.64	0.79	2.3 *	0.19
TOWRE SWE	86.5	0.90	92.6	89.9	0.49	0.67	2.7 *	0.27
Aimsweb	NA	NA	NA	NA	NA	NA	NA	NA
Passage Comprehension	91.8	0.55	93.9	92.7	0.40	0.48	1.2	0.17
GRADE	86.2	0.92	86.9	82.3	0.87	1.18	4.6 *	0.26

5th Grade	Average at baseline	Gap at baseline (Std. Units)	Average at follow-up: Intervention Group	Average at follow-up: Control Group	Gap at follow-up (Std. Units): Intervention Group	Gap at follow-up (Std. Units): Control Group	Impact	RGR
Word Attack	93.1	0.46	98.0	95.3	0.14	0.31	2.7 *	0.56
TOWRE PDE	81.0	1.27	88.3	86.9	0.78	0.87	1.4	0.11
Word Identification	88.7	0.76	92.1	91.6	0.53	0.56	0.5	0.06
TOWRE SWE	84.2	1.05	89.6	88.2	0.69	0.78	1.4 *	0.12
Aimsweb	NA	NA	NA	NA	NA	NA	NA	NA
Passage Comprehension	92.7	0.49	92.2	90.9	0.52	0.60	1.3	0.14
GRADE	91.5	0.57	92.3	92.5	0.51	0.50	-0.2	-0.02

* Impact is statistically significant at the 0.05 level.

Note: RGR defined as RGR = (Impact/(100-Average for Control Group at follow-up).

Note: Gap defined as (100-Average Score)/15, where 100 is the population average and 15 is the population standard deviation.

Note: Values for Aimsweb not available because normed standard scores unavailable.

REFERENCES

Agronin, M.E., J.M. Holahan, B.A. Shaywitz, and S.E. Shaywitz. "The Multi-Grade Inventory for Teachers." In S.E. Shaywitz and B.A. Shaywitz, eds*., Attention Deficit Disorder Comes of Age*. Austin, TX: PRO-ED, 1992, p. 29-67.

Alexander, A., H. Anderson, P.C.Heilman, K.S. Voeller, and J.K Torgesen. "Phonological Awareness Training and Remediation of Analytic Decoding Deficits in a Group of Severe Dyslexics." 1991 41: 193-206.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 1996, 91(434).

Benjamini, Yoav, and Yosef Hochberg. "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society*, Series B, 1995, 57(1): 289-300.

Bloom, Howard. "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review* 1984, 8.

Brown, A.L., A.S. Palincsar, and L. Purcell. "Poor Readers: Teach, Don't Label." In U. Neisser, ed., *The School Achievement of Minority Children: New Perspectives*. Mahwah, NJ: Lawrence Erlbaum Assoc., l986, 105-143.

Bruck, M. "Word Recognition Skills of Adults with Childhood Diagnoses of Dyslexia." *Developmental Psychology* 1990, 26: 439-454.

Dunn, L.M., and Dunn, L.M. *Peabody Picture Vocabulary Test - Third Edition*. Circle Pines, MN: AGS Publishing, 1997.

Elbaum, B., S. Vaughn, M.T. Hughes, and S.W. Moody. "How Effective Are One-to-One Tutoring Programs in Reading for Elementary Students at Risk for Reading Failure? A Meta-Analysis of the Intervention Research." *Journal of Educational Psychology* 2000, 92: 605-619.

Engelmann, S., L. Carnine, G. Johnson. *Corrective Reading, Word Attack Basics, Decoding A*. Columbus, OH: SRA/McGraw-Hill, 1999.

Engelmann, S., L. Meyer, L. Carnine, W. Becker, J. Eisele, and G, Johnson. *Corrective Reading, Decoding Strategies, Decoding B1 and B2*. Columbus, OH: SRA/McGraw-Hill, 1999.

Engelmann, S., L. Meyer, G. Johnson, and L. Carnine. *Corrective Reading, Skill Applications, Decoding C*. Columbus, OH: SRA/McGraw-Hill, 1999.

Foorman, B., and J.K. Torgesen. "Critical Elements of Classroom and Small-Group Instruction to Promote Reading Success in All Children." *Learning Disabilities Research and Practice* 2001, 16: 203-212.

Hanushek, E.A., J.F. Kain, and S.G. Rivkin. "Does Special Education Raise Academic Achievement for Students with Disabilities?" Working Paper No. 6690. Cambridge, MA: National Bureau of Economic Research, 1998.

Hart, B., and T.R. Risley. *Meaningful Differences in the Everyday Experience of Young American Children.* Baltimore: Paul H. Brookes Publishing Co., 1995.

Howe, K.B., and M.M. Shinn. "Standard Reading Assessment Passages for Use in General Outcome Assessment: A Manual Describing Development and Technical Features. Eden Prairie: MN: Edformation, Inc., 2002

Jenkins, J.R., L.S. Fuchs, P. van den Broek, C. Espin, and S.L Deno. "Sources of Individual Differences in Reading Comprehension and Reading Fluency." *Journal of Educational Psychology* 2003, 95: 719-729.

Juel, C. "Learning to Read and Write: A Longitudinal Study of 54 Children from First Through Fourth Grades." *Journal of Educational Psychology* 1988, 80: 437-447.

Little, Roderick J., and Donald B Rubin. "Causal Effects in Clinical and Epidemiological Studies via Potential Outcomes: Concepts and Analytical Approaches." *Annual Review of Public Health* 2000, 21: 121-145.

Little, Roderick J., and Donald B Rubin. *Statistical Analysis with Missing Data, Second Edition.* Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley Interscience, 2002.

Lockavitch, J. "Failure Free Reading." Concord, NC: Failure Free Reading, 1996.

Loney, J., and R. Milich. "Hyperactivity, Inattention, and Aggression in Clinical Practice." In M. Wolraich and D.D. Routh, eds., *Advances in Developmental and Behavioral Pediatrics,* 1982, 3: 1213-147.

Lovett, M.W., L. Lacerenza, S.L. Borden, J.C. Frijters, K.A. Steinbach, and M. DePalma. "Components of Effective Remediation for Developmental Reading Disabilities: Combining Phonological and Strategy-Based Instruction to Improve Outcomes." *Journal of Educational Psychology* 2000, 92: 263-283.

Lyon, G.R., and S.E. Shaywitz. "A Definition of Dyslexia." *Annals of Dyslexia* 2003, 53: 1-14.

MacPhee, K. "Spell Read Phonological Auditory Training (P.A.T.)." Rockville, MD: P.A.T. Learning Systems Inc., 1990.

Manis, F.R., R. Custodio, and P.A. Szeszulski. "Development of Phonological and Orthographic Skill: A Two-year Longitudinal Study of Dyslexic Children." *Journal of Experimental Child Psychology, 56*, 1993, pp. 64-86.

Mastropieri, M.A., and T. Scruggs. "Best Practices in Promoting Reading Comprehension in Students with Learning Disabilities: 1976-1996." *Remedial and Special Education* 1997, 18: 197-213.

Mathematica Policy Research, Inc. "A Proposal for the Evaluation of Reading Interventions Sponsored by The Power4Kids Initiative." Submitted to the Haan Foundation, October 31, 2002 (with American Institutes for Research).

McKinney, J.D. "Longitudinal Research on the Behavioral Characteristics of Children with Learning Disabilities." In J.K. Torgesen, ed., *Cognitive and Behavioral Characteristics of Children with Learning Disabilities.* Austin, TX.: PRO-ED, 1990.

National Reading Panel 2000. "Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and its Implications for Reading Instruction." Washington, DC: National Institute of Child Health and Human Development, 2000.

Rashotte, C.A., K. MacPhee, and J. Torgesen. "The Effectiveness of a Group Reading Instruction Program with Poor Readers in Multiple Grades." *Learning Disability Quarterly* 2001, 24: 119-134.

Raudenbush, Stephen W., and Anthony Bryk. "Hierarchical Linear Models: Applications and Data Analysis Methods, Second Edition." In Jan DeLeeuw and Richard A. Berk, eds*., Advanced Quantitative Techniques in the Social Sciences Series*, Volume 1. Sage Publications: Thousand Oaks, CA, 2002.

Schatschneider, C., J. Buck, J.K. Torgesen, R.K. Wagner, L. Hassler, S. Hecht, and K. Powell-Smith. "A Multivariate Study of Factors That Contribute to Individual Differences in Performance on the Florida Comprehensive Reading Assessment Test." Technical Report 5, Tallahassee, FL: Florida Center for Reading Research, 2004.

Semel, E., E.H. Wiig, and W. Secord. "Clinical Evaluation of Language Fundamentals." Fourth Edition. San Antonio, TX: The Psychological Corporation, 2003.

Share, D. L., and K. Stanovich. "Cognitive Processes in Early Reading Development: A Model of Acquisition and Individual Differences." *Issues in Education: Contributions from Educational Psychology* 1995, 1: -57.

Siegel, L.S. "IQ Is Irrelevant to the Definition of Learning Disabilities." *Journal of Learning Disabilities* 1989, 22: 469-479.

Snow, C.E., M.S. Burns, and P. Griffin. *Preventing Reading Difficulties in Young Children*. Washington, DC: National Academy Press, 1998.

Snowling, M. J. *Dyslexia*, 2nd edition. Oxford: Blackwell Publishers, 2000.

Speece, D.L., and B.K. Keogh. *Research on Classroom Ecologies*. Mahwah, NJ: Lawrence Erlbaum Assoc., 1996.

Stahl, S.A. Four Questions About Vocabulary Knowledge and Reading and Some Answers. In C. Hynd et al., eds. *Learning from Text Across Conceptual Domains*. Mahwah, NJ: Lawrence Erlbaum Associates, 1998.

Stanovich, K. E. "Matthew Effects in Reading: Some Consequences of Individual Differences in Acquisition of Literacy." *Reading Research Quarterly* 1986, 21: 360-407.

Stanovich, K.E. and L.S. Siegel. "The Phenotypic Performance Profile of Reading-Disabled Children: A Regression-Based Test of the Phonological-Core Variable-Difference Model." *Journal of Educational Psychology* 1994: 24-53.

Torgesen, J.K. "Avoiding the Devastating Downward Spiral: The Evidence that Early Intervention Prevents Reading Failure." *American Educator* 2004, 28: 6-19.

Torgesen, J.K. "Recent Discoveries from Research on Remedial Interventions for Children with Dyslexia." In M. Snowling and C. Hulme, eds., *The Science of Reading*. Oxford: Blackwell Publishers, 2005.

Torgesen, J.K., A. W. Alexander, R.K. Wagner, C.A. Rashotte, K. Voeller, T. Conway, and E. Rose. "Intensive Remedial Instruction for Children with Severe Reading Disabilities: Immediate and Long-Term Outcomes from Two Instructional Approaches." *Journal of Learning Disabilities* 2001, 34: 33-58.

Torgesen, J.K., and S.R. Burgess. "Consistency of Reading-Related Phonological Processes throughout Early Childhood: Evidence from Longitudinal-Correlational and Instructional Studies." In J. Metsala and L. Ehri, eds., *Word Recognition in Beginning Reading*. Hillsdale, NJ: Lawrence Erlbaum Assoc., 1998: 161-188.

Torgesen, J.K., and R. Hudson. "Reading Fluency: Critical Issues for Struggling Readers." In S.J. Samuels and A. Farstrup, eds., *Reading Fluency: The Forgotten Dimension of Reading Success*. Newark, DE: International Reading Association (in press).

Torgesen, J.K., C.A. Rashotte, and A. Alexander. "Principles of Fluency Instruction in Reading: Relationships with Established Empirical Outcomes." In M. Wolf, ed., *Dyslexia, Fluency, and the Brain*. Parkton, MD: York Press, 2001: 333-355.

Torgesen, J.K., R. K. Wagner, and C.A. Rashotte. *Test of Word Reading Efficiency*. Austin, TX: PRO-ED Publishing, Inc., 1999.

Truch, S. "Stimulating Basic Reading Processes Using Auditory Discrimination in Depth." *Annals of Dyslexia* 1994, 44: pp. 60-80.

Truch, S. "Comparing Remedial Outcomes Using LIPS and Phono-Graphix: An In-Depth Look from a Clinical Perspective." Calgary, Alberta, Canada: The Reading Foundation, Unpublished manuscript, 2003.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. "The Nation's Report Card: Reading Highlights 2003." NCES 2004-452. Washington, DC: NCES. Available online at [www.nces.ed.gov/nationsreportcard/pdf/main2003/2004452.pdf].

Vaughn, S., S. Moody, and J.S. Schumm. "Broken Promises: Reading Instruction in the Resource Room." *Exceptional Children* 1998, 64: 211-226.

Wagner, R. K., J.K. Torgesen, and C.A. Rashotte. *Comprehensive Test of Phonological Processes*. Austin, TX: PRO-ED Publishing, Inc., 1999.

Williams, K.T. *Group Reading and Diagnostic Evaluation*. Circle Pines, MN: American Guidance Service, 2001.

Wilson, B. *The Wilson Reading System*, Third Edition. Millbury, MA: Wilson Language Training Corp., 2002

Wise, B.W., J. Ring, and R.K. Olson. "Training Phonological Awareness With and Without Explicit Attention to Articulation." *Journal of Experimental Child Psychology* 1999, 72: 271-304.

Wolf, M., and M. Denkla. *Rapid Automatized Naming and Rapid Alternating Stimulus Tests*. Austin, TX: PROED, Inc., 2005.

Woodcock, R.W. *Woodcock Reading Mastery Tests-Revised* NU (WRMT-R/NU). Circle Pines, MN: American Guidance Service, l998.

Woodcock, R.W., K.S. McGrew, and N. Mather. *Woodcock-Johnson III Tests of Achievement.* Itasca, IL: Riverside Publishing, 2001.

Zigmond, N. "Organization and Management of General Education Classrooms." In D.L. Speece and B.K. Keogh, eds., *Research on Classroom Ecologies.* Mahwah, NJ: Lawrence Erlbaum Publishers, 1996: 163-190.

Zigmond, N. "Organization and Management of General Education Classrooms." In D.L. Speece and B.K. Keogh, eds., *Research on Classroom Ecologies.* Mahwah, NJ: Lawrence Erlbaum Publishers, 1996: 163-190.


DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA