National Assessment of Title I
Final Report

Volume II: Closing the Reading Gap: 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

Under IES Grant R305U030002A

OCTOBER 2007

Joseph Torgesen
Florida Center for Reading Research

Allen Schirm
Laura Castner
Sonya Vartivarian
Wendy Mansfield
Mathematica Policy Research

David Myers
Fran Stancavage
American Institutes for Research

Donna Durno
Rosanne Javorsky
Allegheny Intermediate Unit

Cinthia Haan
Haan Foundation

NCEE 2008-4013
U.S. DEPARTMENT OF EDUCATION

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

October 2007

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

To order copies of this report,

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

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

Alternate Formats

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

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 Ambrose Monell Foundation, the Barksdale Reading institute, the Grable Foundation, the Haan Foundation for Children, the Heinz Endowment, the W.K. Kellogg Foundation, the Raymond Foundation, the Rockefeller Foundation, and the U.S. Department of Education's Institute of Education Sciences. 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, and the William and Flora Hewlett Foundation and the Richard King Mellon Foundation for their support of the functional magnetic resonance imaging study, which collaborated with our evaluation of the four reading programs.

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. Alfreda Holmes, Donna Dorsey, and Daryl Hall were instrumental in producing and editing the document.

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 the American Institutes for Research, 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.

DISCLOSURE OF POTENTIAL CONFLICTS OF INTEREST

The research team for this evaluation consists of two grantees: Florida State University with Joe Torgesen as Principal Investigator, and the Corporation for the Advancement of Policy Evaluation with David Myers and Allen Schirm as Principal Investigators. None of these organizations or their key staff has financial interests that could be affected by findings from the evaluation of the four reading interventions reported here. No one from the Scientific Review Board, convened to provide advice and guidance, has financial interests that could be affected by findings from the evaluation.

CONTENTS

IV *(continued)*

EXECUTIVE SUMMARY

KEY EVALUATION QUESTIONS AND STUDY DESCRIPTION

Evaluation Context

According to the National Assessment of Educational Progress (U.S. Department of Education 2006), 36 percent of fourth graders read below the basic level. Such literacy problems can worsen as students advance through school and are exposed to progressively more complex concepts and courses. While schools often are 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 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? For whom do they work best? 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 would 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. Many struggling readers in late elementary school continue to have word-level reading problems affecting the accuracy and fluency with which they read text. For some of these students (those with broad verbal abilities in the average range), these word-level reading difficulties are the primary bottleneck that prevents them from being able to understand and learn from text. Many other struggling readers have not only word-level reading difficulties but also limitations in vocabulary and comprehension processes that directly impact their ability to read grade-level text with understanding. In the original design of this study, Corrective Reading and Wilson Reading were to focus primarily on improving students' word-level reading skills, while Spell Read P.A.T. and Failure Free Reading were to focus on building reading comprehension and vocabulary in addition to word-level skills. This design addressed the question of whether it would be more effective to focus the limited amount of time available for remedial instruction on improving word-level reading skills as much as possible, or whether it would be more effective to spread instructional time between word-level skills and vocabulary and comprehension processes. 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

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

be substantially improved by providing, for a sustained period of time, the kinds of skillful, systematic, and explicit instruction offered by the programs in this study (Torgesen 2005).

Evaluation Purpose and Design

Conducted just outside Pittsburgh, Pennsylvania, in the Allegheny Intermediate Unit (AIU), the evaluation has explored 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 provides educators with rigorous evidence of what could happen in terms of reading 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 designed to address the following questions:

- 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 the interventions 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, we based the impact study 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 three and five 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 third- and fifth-grade students identified as struggling readers, 1,502 were screened, and 1,042 met the test-score criteria. Of these eligible students, 779 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 of the regular instruction received by the treatment group beyond the interventions.

We have collected test data and other information on students, parents, teachers, classrooms, and schools several times over a two-year period. Key data collection points include the period just before the interventions began, when baseline information was collected, and the periods immediately after and one year after the interventions ended, when follow-up data were collected.

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: Corrective Reading, Failure Free Reading, Spell Read P.A.T., and Wilson 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. Developing word-level skills helps children overcome two of the three problems that struggling readers in late elementary school generally face, namely, accuracy and fluency. Struggling readers rely heavily on guessing based on the context of the passage, and they encounter more words that they cannot read "by sight" than do average readers. The interventions designed to focus on both word-level skills and reading comprehension will directly address the third type of reading problem faced by struggling readers, comprehending the text.

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. However, in a time-by-activity analysis of the instruction that was actually delivered (Torgesen et al. 2006), 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. Although reclassifying Spell Read P.A.T. for our analyses does not invalidate the conclusions of this study, the impacts of that intervention and the other interventions that focused primarily on word-level skills in this study could be different if they were implemented differently in another context, with, for example, more emphasis on the development of comprehension skills.

- ***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 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 classified as one of the two "word-level plus comprehension" interventions, but after the time-by-activity analysis, we

determined that it was more appropriately classified as a "word-level" intervention. Because the word-level instructional content in Spell Read is more structured than the instruction designed to build reading comprehension, the relatively short instructional sessions in this study led to a different balance of word-level and comprehension instruction than was anticipated. That is, to accomplish the highly specified word-level instruction contained in the program, the teachers reduced the amount of time they spent on the comprehension components. In clinical settings, Spell Read is typically provided in 70-minute sessions, whereas the sessions in this study averaged closer to 55 minutes in length.

- ***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 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 (phonics and phonemic analysis), 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, this intervention was 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 noncontrolled 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 Wilson Reading was originally designated as a "word-level" intervention, the developer was 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 several times for the evaluation to assess student progress in learning to read. These measures 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. 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. (See the note on Table 1 for more information about the means and standard deviations for scores on the AIMSweb test.)

In addition to administering these seven reading tests for the evaluation, we collected reading and mathematics scores from the tests administered by the AIU schools as part of the Pennsylvania System of School Assessment (PSSA). Designed by advisory committees of Pennsylvania educators, the PSSA reading tests measure students' skills in comprehending text, while the PSSA mathematics tests measure skills ranging from recalling specific facts to solving problems. Scaled scores for the PSSA tests are derived using item response theory (Rasch) models. Students in the evaluation sample took these standards-based PSSA tests from late March to early April of the 2003-04 school year, the year during which the interventions took place. For the seven tests that we administered, the scores that we analyze in this report are from the tests that the students took one year later—near the end of the 2004-05 school year, one year after the interventions ended. Our results from analyzing the scores on these tests from one year earlier—the end of the intervention year—are presented in Torgesen et al. (2006).

Implementing the Interventions

The interventions provided instruction to students in the treatment group from November 2003 through 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 55 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. For third graders, the interventions' small-group instruction substituted

for the large-group instruction provided in their classrooms, with a negligible effect on the total amount of reading instruction received. For fifth graders, the interventions increased both the amount of small-group reading instruction and the total amount of reading instruction during the intervention year.

Teachers were recruited from participating schools on the basis of experience and characteristics and skills relevant to teaching struggling readers. They received, on average, nearly 70 hours of training and professional development support during the intervention year.

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 28 percent were African American, and 72 percent were white. Fewer than two percent were Hispanic. Roughly 32 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 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 of 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 (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 GRADE they scored, on average, at the 23rd percentile.

This sample, as a whole, was substantially less impaired in basic reading skills than most other samples assessed in previous research with older reading disabled students (Torgesen 2005). These earlier studies typically examined samples in which the average students' phonemic decoding and word reading accuracy skills 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 evaluated 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.

KEY FINDINGS

This evaluation assesses the impacts of the four interventions on the treatment groups in comparison with the control groups. In the first report from the evaluation (Torgesen et al. 2006), we presented impacts on reading test scores at the end of the intervention year, when the students in the evaluation

were third and fifth graders. In this second (and last) report from the evaluation, we present estimates of impacts on scores from the same tests as of the end of the following year, when most of the students were fourth and sixth graders. We also present estimates of impacts on PSSA scores. However, the scores that we were able to obtain were from the tests taken the previous year, that is, during the intervention year when the students were third and fifth graders. In other words, this report presents impacts at two points in time: late March to early April of the intervention year (for PSSA scores) and the end of the following school year (for the tests that we administered). Because most of the third and fifth graders in the first year are, respectively, fourth and sixth graders in the following year, we refer to and analyze student outcomes according to their "grade cohorts," which refer to the students' grade level when they entered the evaluation. Specifically, we present estimates for the "third-grade cohort" and the "fifth-grade cohort."

We provide detailed estimates of 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. An impact of the four interventions combined is not the impact of implementing the four interventions simultaneously. Rather, it can be interpreted as the impact of providing a struggling reader in third or fifth grade with the opportunity to receive a substantial amount of instruction in small groups with reasonably well-trained teachers, although as noted elsewhere, the content and instructional focus across the four interventions varied considerably. Such an impact is of greatest relevance to federal and state policymakers who can support broad programmatic approaches to instruction but cannot generally endorse specific products. In contrast, school district and school administrators must select specific products. For that purpose, the impact of being randomly assigned to an individual intervention—as modified or partially implemented for this study—is most relevant.

In addition to the impacts for the entire third- and fifth-grade cohorts, the full report also estimates impacts for various subgroups. The subgroups include 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.

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 follow up, or as the change in test scores between baseline and follow-up (the "gain"). In the tables of impacts for tests that we administered (see Table 2, for example), we show three types of numbers. The baseline score shows the average standard score for students at the beginning of the intervention 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 control group. The gain in the intervention group students' average test scores between the baseline and the follow-up can be calculated by adding the control group gain and the impact. Given the timing of the seven tests that we administered, our impact estimates based on those tests compare the effects of one year with the interventions followed by one year without for the intervention group students to the effects of two years without the interventions for the control group students. In contrast, the impact estimates based on PSSA test scores show the effects of, roughly, the first two-thirds to three-quarters of the interventions.

In practice, impacts were estimated using hierarchical linear models, with separate models for the third- and fifth-grade cohorts. The models include a student-level model and a school-level model. In the student-level model, we include an indicator for treatment status and 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 includes 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.

Our key findings are as follows:

- ***The interventions improved some reading skills.*** For students in the third-grade cohort, the four interventions combined had impacts on phonemic decoding, word reading accuracy and fluency, and reading comprehension, although impacts were not detected for all measures of accuracy and fluency or comprehension (see Table 2). For students in the fifth-grade cohort, the four interventions combined improved phonemic decoding on one measure, but led to a small reduction in oral reading fluency. The three word-level interventions combined had similar impacts to those for all four interventions combined, although they did not have an impact on either measure of comprehension for students in the third-grade cohort, and they did have impacts on both measures of phonemic decoding for students in the fifth-grade cohort. For students in the third-grade cohort, Failure Free Reading (the only word level plus comprehension program) had impacts on one measure of phonemic decoding, two of the three measures of word reading accuracy and fluency, and one measure of comprehension. However, this intervention did not have any impacts for students in the fifth-grade cohort.

- ***The interventions did not improve PSSA scores.*** For students in the third-grade cohort, we did not detect significant impacts of the four interventions combined on reading and mathematics test scores from the Pennsylvania System of School Assessment (see Table 3). For students in the fifth-grade cohort, the four interventions combined lowered the reading and mathematics scores.

- ***Younger students benefited more.*** The interventions generally helped students in the third-grade cohort more than students in the fifth-grade cohort (see Tables 2 and 3). However, the interventions did not consistently benefit any one subgroup more than another.

- ***The interventions narrowed some reading gaps.*** The four interventions combined generally narrowed the reading gap for students in the intervention groups compared with students in the control group for the third-grade cohort. The reading gap describes the extent to which the average student in one of the evaluation groups (intervention or control) is lagging behind the average student in the population (see Figures 1-12 and Table 4). The reduction in the reading gap attributable to the interventions 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 in Word Attack skills by about two-thirds for students in the third-grade cohort. On other word-level tests and a measure of reading comprehension, the interventions reduced the gap for students in the third-grade cohort by about one-sixth to one-third. For students in the fifth-grade cohort, the interventions reduced the gap in Word Attack skills by one-half.

The key findings presented in this report for the seven tests administered for this study one year after the interventions ended are similar to the findings from the end of the intervention year. In our earlier report (Torgesen et al. 2006) we found that the four interventions combined and the three word-level interventions had impacts for students in the third-grade cohort on phonemic decoding, word reading accuracy and fluency, and reading comprehension. We found fewer significant impacts for students in the fifth-grade cohort than for students in third-grade cohort. Also, for the four interventions combined, the reading gaps for students in the intervention group were generally smaller than the gaps for students in the control group.

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 (%)	72	70	74
Race--African American (%)	28	30	26
Race--Other (%)	a	a	a
Family income less than $30,000 (%)	49	48	49
Family income between $30,000 and $60,000 (%)	35	34	36
Family income over $60,000 (%)	16	18	15
Eligible for Free or Reduced Price Lunch (%)	45	45	46
Has any learning or other disability (%)	32	34	30
Mother has bachelor's degree or higher (%)	13	13	13

	Standard		Standard		Standard	
Reading Tests	Score	Percentile	Score	Percentile	Score	Percentile
Screening Tests						
TOWRE Sight Word Efficiency	84.3	15	84.4	15	84.1	15
TOWRE Phonemic Decoding Efficiency	83.0	13	85.6	17	80.6	10
Peabody Picture Vocabulary Test--Revised	94.7	36	94.4	35	95.0	37
Baseline Tests						
WRM Word Identification	88.7	23	88.7	22	88.7	23
TOWRE Phonemic Decoding Efficiency	83.2	13	85.5	17	81.0	10
WRM Word Attack	92.8	32	92.4	31	93.2	33
TOWRE Sight Word Efficiency	85.3	16	86.6	19	84.2	15
AIMSweb (Raw score)	NA	NA	40.9	NA	77.0	NA
WRM Passage Comprehension	92.1	30	91.6	29	92.6	31
GRADE	88.8	23	86.1	18	91.2	28
Woodcock Johnson Spelling	89.8	25	88.5	22	90.9	27
Woodcock Johnson Calculation	94.8	36	95.4	38	94.2	35
Sample Size	729		329		400	

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

Note: All standard scores have mean 100 and standard deviation 15. The mean raw scores for AIMSweb tests, administered to students across the country in the fall in the school years 2000-2001, 2001-2002, and 2002-2003, were 75 and 112 for third and fifth graders, respectively. The respective standard deviations were 39 and 47.

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

a Values suppressed to protect student confidentiality.

Table 2

Impacts on Reading Test Scores for 3rd and 5th Grade Cohorts
One Year After the Intervention Year

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 3 Cohort	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	92.4	-0.3	5.3 *	0.3	5.5 *	-2.2	4.9 *	-0.3	5.4 *	0.5	5.8 *	0.7	5.2 *
TOWRE PDE	85.5	1.1	4.0 *	1.0	4.9 *	1.3	1.3	3.4	4.9 *	1.8	4.1 *	-2.3	5.5 *
Word Identification	88.7	-0.1	2.3 *	0.0	2.5 *	-0.4	1.8	1.8	0.7	-2.2	4.1 *	0.6	2.6 *
TOWRE SWE	86.6	3.5	1.7 *	3.7	1.6 *	2.7	2.0 *	4.0	0.9	2.8	2.6 *	4.3	1.4
AIMSweb	40.9	33.7	5.3	33.4	4.5	34.4	7.9 *	30.1	6.0 *	31.9	3.6	38.3	3.9
Passage Comprehension	91.6	-0.3	2.1 *	0.3	1.3	-2.1	4.4 *	1.2	0.1	-2.5	3.5	2.3	0.3
GRADE	86.1	-7.5	1.0	-6.9	0.4	-9.3	2.8	-10.0	2.1	-10.4	0.1	-0.1	-1.1
Sample Size	329		329		240		89		91		70		79

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 5 Cohort	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	93.2	1.5	2.7 * #	1.8	3.8 *	0.4	-0.8 #	0.0	3.5	2.3	7.8 *	3.1	0.2 #
TOWRE PDE	81.0	5.3	1.7	5.2	2.4 *	5.4	-0.3	4.9	3.2	4.2	2.6	6.6	1.4
Word Identification	88.7	3.0	-0.6 #	3.4	-0.6 #	1.7	-0.6 #	1.5	0.1	4.3	0.0 #	4.3	-1.9 #
TOWRE SWE	84.2	3.0	1.4	3.1	1.4	3.0	1.5	1.5	3.4 *	2.7	1.1	5.0	-0.4
AIMSweb	77.0	30.9	-3.9 * #	30.7	-3.9 * #	31.6	-4.1 #	26.5	-3.3 #	29.7	-3.0	35.9	-5.3
Passage Comprehension	92.6	-0.4	-1.1 #	-0.8	-0.7	0.8	-2.5 #	-2.8	-0.9	-1.6	0.9	1.9	-2.1
GRADE	91.2	-3.5	0.7	-4.7	1.2	0.2	-0.9	-4.8	-1.1	-8.9	4.7	-0.4	0.0
Sample Size	400		400		272		128		100		88		84

Note: The Failure Free, Spell Read, Wilson Reading, and Corrective Reading interventions are labeled A, B, C, and D, respectively. These labels are arbitrary and not related to performance. ABCD is the label for the four interventions combined and BCD is the label for the three word-level intervention combined.

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

Note: According to the first row of estimates, students in the third-grade cohort achieved an average standardized Word Attack score of 92.4 at "baseline," that is, shortly after the beginning of third grade—the intervention year. For the Failure Free Reading intervention, the average standardized Word Attack score one year after the intervention year fell by 2.2 points from the baseline score for the students in the control group (the "control gain"). Also one year after the intervention year, the average score for the students in the treatment group for the Failure Free Reading intervention was 4.9 points higher than for the students in the control group (the "impact"), a difference that is statistically significant, as indicated by the asterisk. According to the columns for "All Interventions," the average score for the control group was 0.3 points lower than the baseline score, and the average score for the treatment group was 5.3 points higher than the average for the control group, a statistically significant difference.

* Impact is statistically significant at the 0.05 level.
# Impact is statistically different from the 3rd grade cohort impact at the 0.05 level.

Table 3

Impacts on PSSA Reading and Math Scores for 3rd and 5th Grade Cohorts
Late March/Early April of the Intervention Year

Grade 3 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
PSSA Reading	-15.6	-3.8	-51.1 *	-39.9	52.5	-23.8
PSSA Math	20.2	14.2	38.4	-15.5	56.6 *	1.4
Sample Size	329	240	89	92	71	77

Grade 5 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
PSSA Reading	-27.3 *	-25.3	-33.4 *	-30.0	-23.8	-22.1
PSSA Math	-28.8 * #	-34.0 * #	-13.4	-20.1	-56.4 * #	-25.4 *
Sample Size	408	280	128	102	92	86

Note: The Failure Free, Spell Read, Wilson Reading, and Corrective Reading interventions are labeled A, B, C, and D, respectively. These labels are arbitrary and not related to performance. ABCD is the label for the four interventions combined and BCD is the label for the three word-level intervention combined.

Note: According to the first row of estimates, students in the third-grade cohort assigned to the Failure Free Reading intervention achieved a standardized score on the PSSA Reading test that was 51.1 points lower than the average score achieved by the students in the control group, a statistically significant difference, as indicated by the asterisk. The average standardized score for students participating in any intervention was 15.6 points lower than the average score for students assigned to a control group, a difference that is not statistically significant.

* Impact is statistically significant at the 0.05 level.
# Impact is statistically different form the 3rd grade impact at the 0.05 level.

Table 4

Relative Gap Reduction (RGR): All Interventions Combined
One Year After the Intervention Year

Grade 3 Cohort	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	92.4	0.50	97.4	92.1	0.17	0.53	5.3 *	0.68
TOWRE PDE	85.5	0.97	90.5	86.6	0.63	0.90	4.0 *	0.29
Word Identification	88.7	0.76	90.9	88.6	0.61	0.76	2.3 *	0.20
TOWRE SWE	86.6	0.90	91.7	90.0	0.55	0.67	1.7 *	0.17
AIMSweb	NA	NA	NA	NA	NA	NA	NA	NA
Passage Comprehension	91.6	0.56	93.4	91.3	0.44	0.58	2.1 *	0.24
GRADE	86.1	0.93	79.6	78.6	1.36	1.42	1.0	0.05

Grade 5 Cohort	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	93.2	0.45	97.3	94.7	0.18	0.36	2.7 *	0.50
TOWRE PDE	81.0	1.27	88.0	86.3	0.80	0.91	1.7	0.13
Word Identification	88.7	0.75	91.1	91.7	0.60	0.56	-0.6	-0.07
TOWRE SWE	84.2	1.05	88.6	87.2	0.76	0.85	1.4	0.11
AIMSweb	NA	NA	NA	NA	NA	NA	NA	NA
Passage Comprehension	92.6	0.49	91.0	92.2	0.60	0.52	-1.1	-0.15
GRADE	91.2	0.59	88.4	87.7	0.77	0.82	0.7	0.06

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

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

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

Note: According to the first row of estimates, students in the third-grade cohort achieved an average standardized score of 92.4 on the Word Attack test at "baseline," that is, shortly after the beginning of third grade—the intervention year. One year after the intervention year, that is, at "follow-up," the students participating in any intervention achieved an average standardized score of 97.4, and the students in the control group achieved an average standardized score of 92.1, implying a statistically significant impact of 5.3 points. The "gap at baseline," measured as the difference between the population average (100) and the study sample average (92.4) divided by the population standard deviation (15), was 0.5. One year after the intervention year, the gap was reduced 68 percent (see the "RGR"), when the reduction is measured as the impact (5.3) divided by the difference between the population average (100) and the control group average (92.1). The calculations described in this note might produce results that are slightly different from the estimates in the table due to rounding.

* Impact is statistically significant at the 0.05 level.

Figure 1

Gap Reduction for Third-Grade Cohort: Word Attack

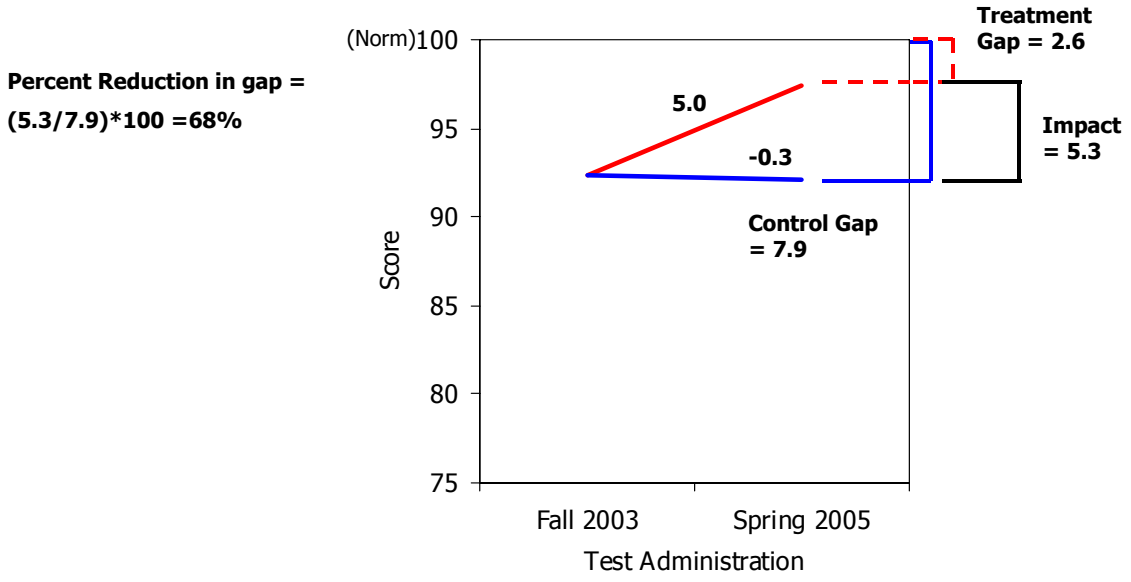


Figure 2

Gap Reduction for Third-Grade Cohort: TOWRE PDE

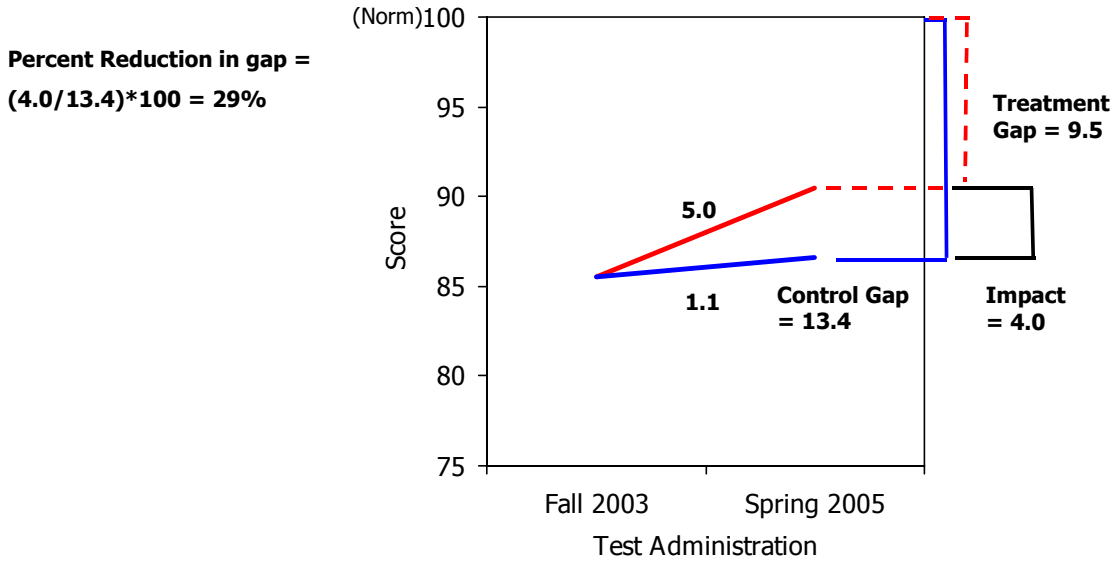


Figure 3

Gap Reduction for Third-Grade Cohort: Word Identification

Percent Reduction in gap =
(2.3/11.4)*100 = 20%

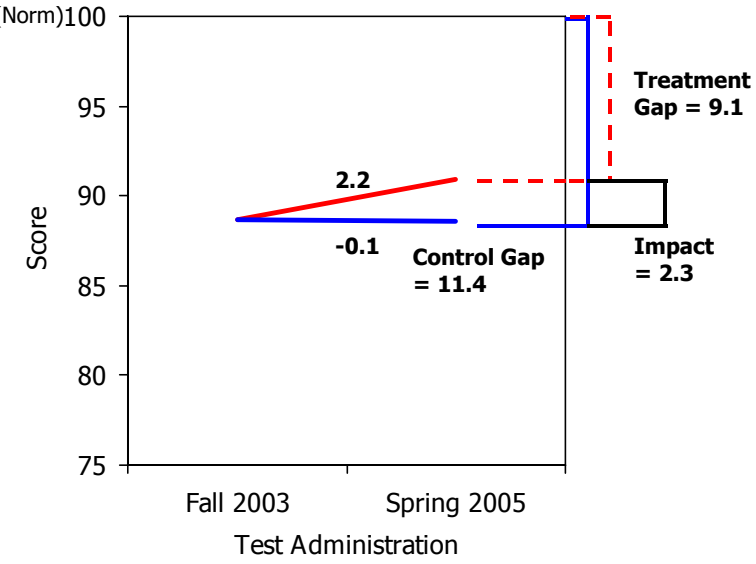


Figure 4

Gap Reduction for Third-Grade Cohort: TOWRE SWE

Percent Reduction in gap =
(1.7/10.0)*100 = 17%

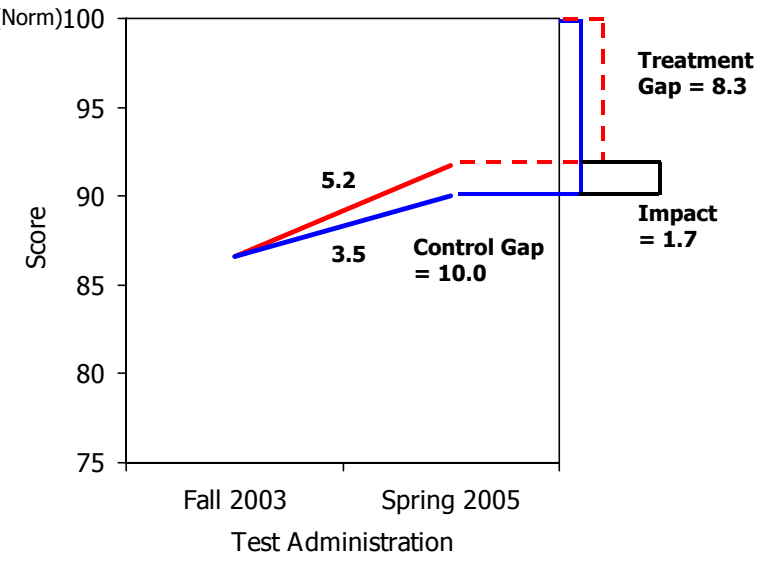


Figure 5

Gap Reduction for Third-Grade Cohort: Passage Comprehension

Percent Reduction in gap =
(2.1/8.7)*100 =24%

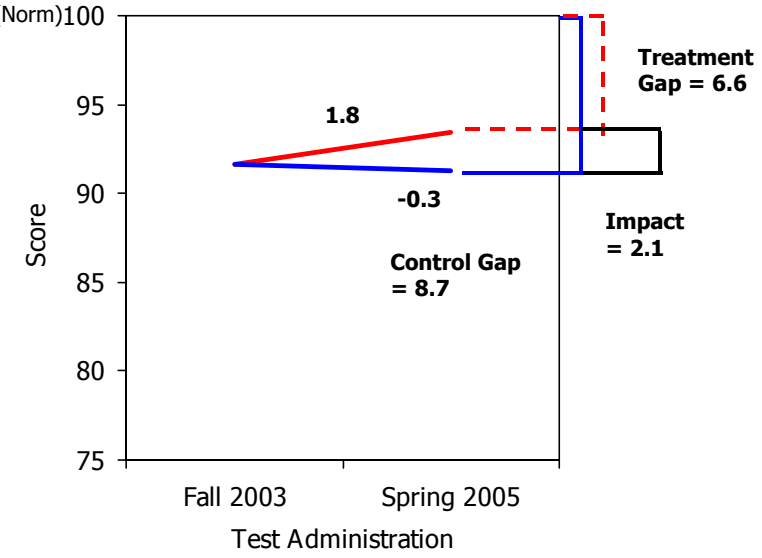


Figure 6

Gap Reduction for Third-Grade Cohort: GRADE

Percent Reduction in gap =
(1.0/21.4)*100 = 5%

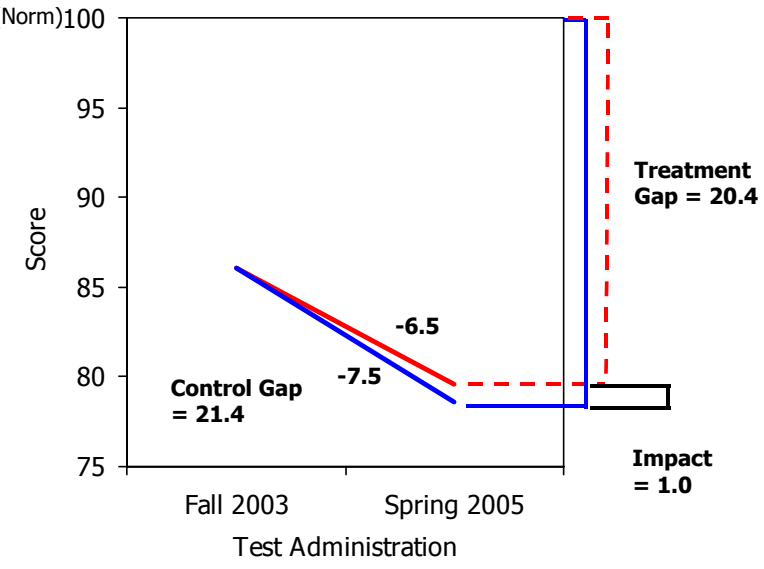


Figure 7

Gap Reduction for Fifth-Grade Cohort: Word Attack

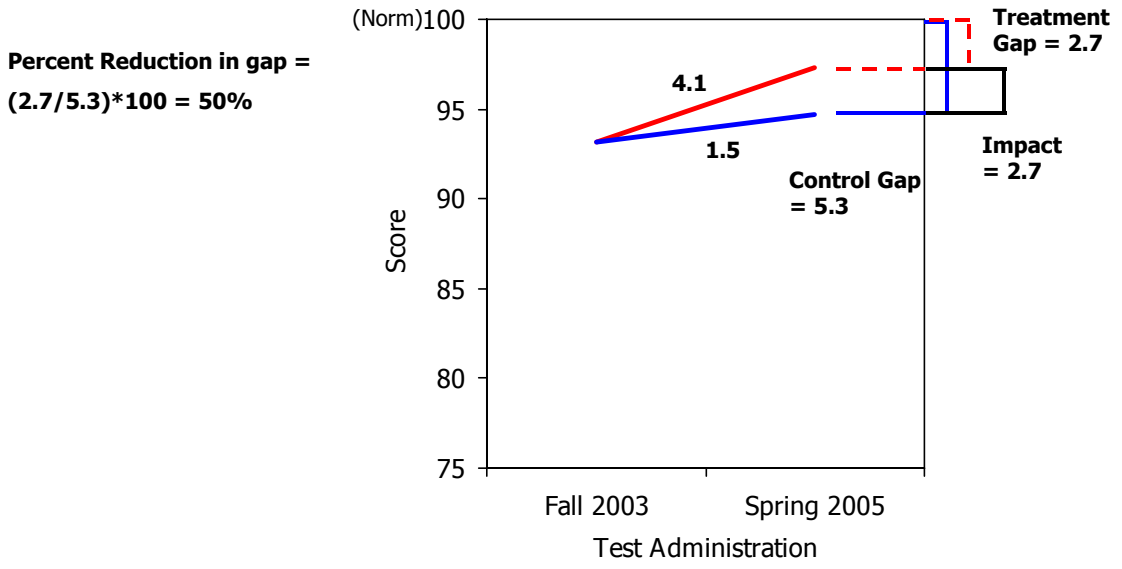


Figure 8

Gap Reduction for Fifth-Grade Cohort: TOWRE PDE

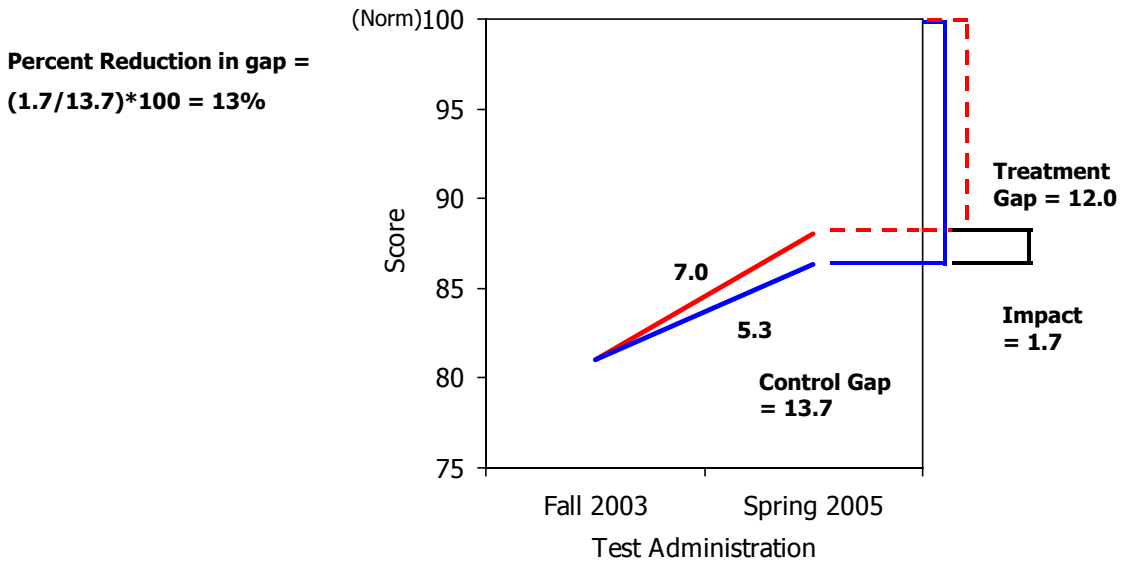


Figure 9

Gap Reduction for Fifth-Grade Cohort: Word Identification

Percent Reduction in gap =
(-0.6/8.3)*100 = -7%

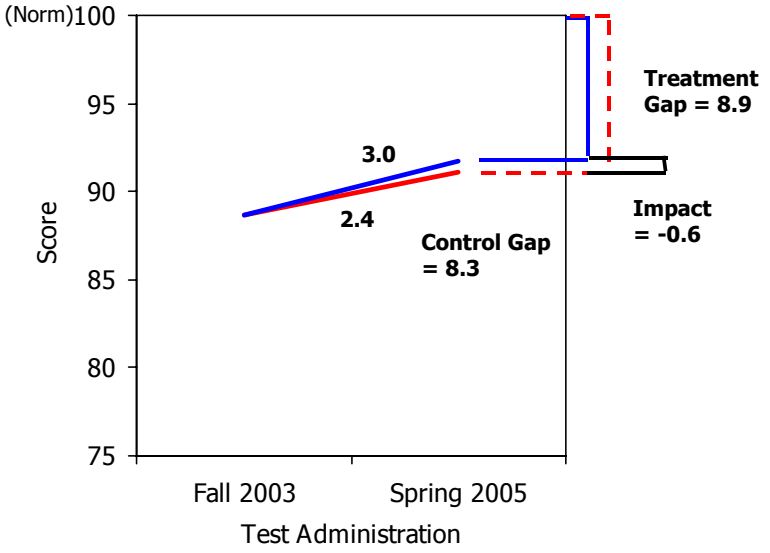


Figure 10

Gap Reduction for Fifth-Grade Cohort: TOWRE SWE

Percent Reduction in gap =
(1.4/12.8)*100 = 11%

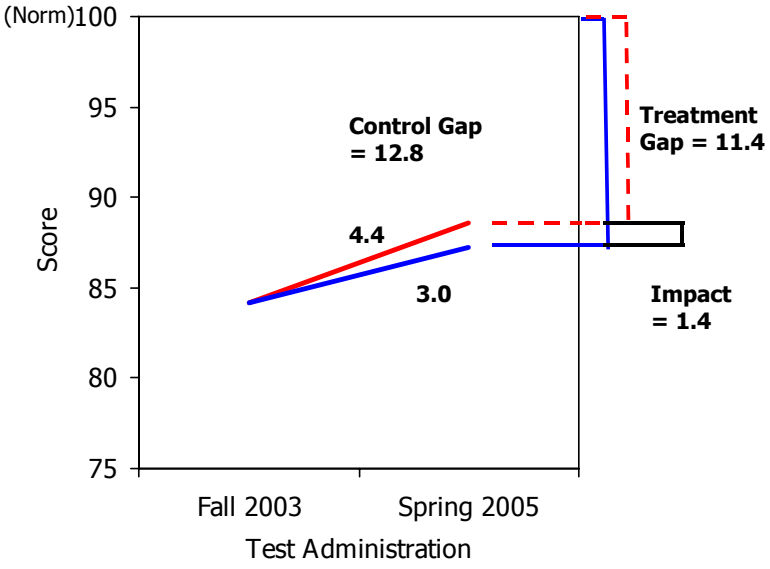


Figure 11

Gap Reduction for Fifth-Grade Cohort: Passage Comprehension

Percent Reduction in gap =
(-1.1/7.8)*100 = -15%

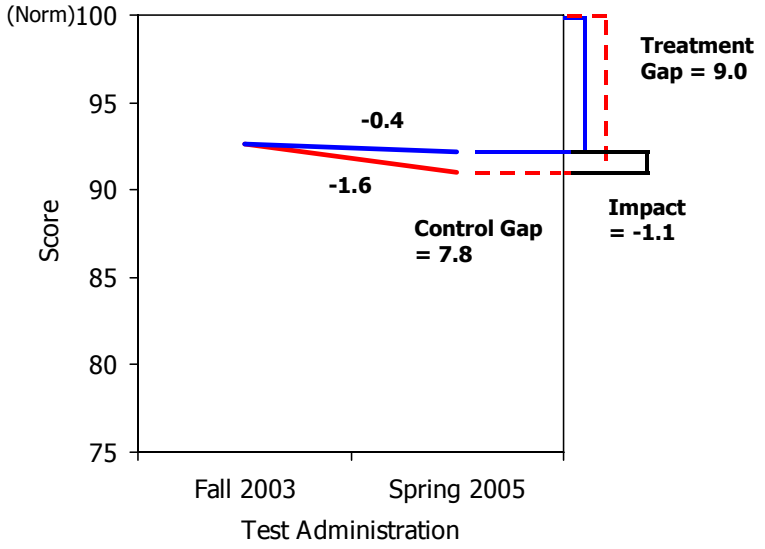
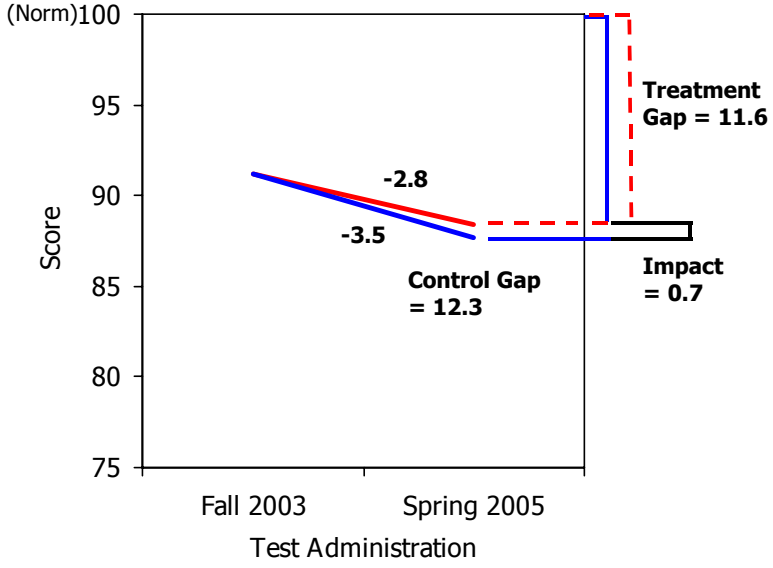


Figure 12

Gap Reduction for Fifth-Grade Cohort: GRADE

Percent Reduction in gap =
(0.7/12.3)*100 = 6%



I. INTRODUCTION

A. OVERVIEW

According to the National Assessment of Educational Progress (U.S. Department of Education 2006), 36 percent of 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" that separates them from average readers. Recent studies have found that children placed in special education after third grade typically achieve no more than a year's gain 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 (Hanushek, Kain, and Rivkin 1998; Vaughn, Moody, and Schumm 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, MacFee, and Torgesen 2001; Torgesen et al. 2001; Truch, 2003, 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

A large fraction of students in the late elementary school grades are unable to read at a basic level (U.S. Department of Education 2006). 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 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, 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*. 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 et al. 2003; Torgesen and Hudson 2006; 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 2006). 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, that is, the ability to construct meaning from the text. For some 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 not only to accuracy and fluency problems but also to general verbal skills—particularly vocabulary skills—that are significantly below average (Snow, Burns, and Griffin 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 (Pressley 2000; 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 programs on the basis of previous research showing their potential to improve reading skills in children of the same age as those in this study. To be included in the study, the program developers also had to have the capacity to provide initial professional development and ongoing support to the intervention teachers. The selection of interventions 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.[2] 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 the study 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. Two programs were included within each of the classifications (word level and word level plus comprehension) to increase the generalizability of the contrast. That is, if only one instructional program had been used to represent each class of instruction, any differences emerging between the two treatment classes might be as easily tied to specific features of each program as to the more general difference between instruction focusing primarily on word-level skills versus instruction spread across a broader array of reading skills.

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 designed to be held constant across students within each program, although 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.

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

[2] A complete list of members of the advisory board is provided in Appendix J.

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.[3]

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 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 either to the increased intensity of instruction or to the particular focus of the intervention.

D. EVALUATION DESIGN AND IMPLEMENTATION

We designed the evaluation to answer 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 the interventions 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 three and five 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.

[3] 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 did occur. In grade 3, students in both the treatment and control groups received, on average, the same number of hours of reading instruction per week during the intervention year, although more of the treatment group hours were delivered in small group settings. In grade 5, there was an overall increase in hours of reading instruction per week for the treatment group, but the increase was substantially less than the 4.5 hours per week contributed by the intervention. (See Figures III.1 and III.2.) The impact analysis was not designed to estimate the impact of each hour of reading instruction. It was not possible to control experimentally the reading instruction received outside of the interventions.

Students in the control groups received the same instruction in reading that they would have received ordinarily.

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 fifth percentile on a vocabulary test. From an original pool of 1,576 third- and fifth-grade students identified as struggling readers, 1,502 were screened and 1,042 met the test-score criteria. Of these eligible students, 779 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 of 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 55 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. Although some previous studies of struggling readers of this age level have employed more intensive instruction than was used here (e.g. Torgesen, et al. (2001) provided approximately 70 hours of one-one-one instruction), other studies (Torgesen 2005) have obtained substantial improvements in students' reading levels through instruction provided in small groups for amounts of time similar to that provided in this study. These previous studies provide support for the idea that the instructional conditions in this study would allow a reasonable estimate of potential differences in the impact of word-level or word level plus vocabulary and comprehension approaches on the reading ability of struggling readers in third and fifth grade.

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 have collected test data and other information on students, parents, teachers, classrooms, and schools several times over a two-year period. Key data collection points include the period just before the interventions began, when baseline information was collected, and the periods immediately after the interventions ended and one year after, when follow-up data were collected. In this report, we focus on estimates of the impacts of the interventions one year after the interventions ended. We also present estimates of impacts on state assessments administered near the end of the intervention year.

II. DESIGN AND IMPLEMENTATION OF STUDY

This evaluation has two main elements: (1) an impact study, and (2) an implementation study. The implementation study has examined 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. We describe the design and main findings of that study in Torgesen et al. (2006). We summarize the findings and present some additional findings in the next chapter.

This chapter focuses 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, subsequent differences in outcomes can be attributed to the interventions and not to pre-existing differences between the groups.[4] 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,

[4] 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. Subsequently, 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 correlations (see Appendix I). For the power calculations, the two-tailed significance level is 0.05 with a power of 0.80. 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 and 0.35 (in standard deviation units), respectively; the MDI for testing whether an individual intervention had an impact is approximately 0.6. When testing impacts for a subgroup that is half of the sample within a grade cohort, the MDI's for all interventions combined and for each intervention individually are approximately 0.35 and 0.7, respectively.

within four strata defined by the percentage of students eligible for free or reduced-price school lunch.[5] 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.[6,7]

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. Torgesen et al. (2006) also compared 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. 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, 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 a few differences. With just 32 school units randomized, however, it is not surprising to observe such differences among the four groups.[8] While small differences could 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 generally very similar to each other (see Tables II.2 through II.5), as discussed in detail later.

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]

[5] We did not restrict our pool of schools to those that had no other reading interventions in third or fifth grade. One school was a Reading First school, and several provided reading support using Title I funds. Several others implemented specialized reading programs such as Reading Recovery and Read to Succeed. Additional information pertaining to these programs was not obtained.

[6] 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).

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

[8] When adjustments for multiple comparisons are made, using methods discussed in more detail in Chapter IV and Appendix D, we find that only the difference in the percentage male within the third-grade cohort remains significant.

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

- ***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]

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					
Fraction 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	
Fraction white	0.85	0.70	0.76	0.82	
Fraction African American	0.14	0.29	0.23	0.16	
School-wide Title I	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 is given equal weight.

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

- ***Determine Eligibility***. For the 1,502 students screened, eligibility was determined using the following 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); 1,042 were eligible based on the two screening tests
  - Having written parental consent to participate in the study;[11] 779 of the test-score eligible students received this consent.[12]

[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.

[11] Both the parental consent form and the student assent form requested approval for the student to participate in the reading program, if selected, as well as approval for participation in all data collection activities.

Table II.2

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

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	8.7		8.7	8.7	
Male (%)	53	58		73	59		39	21		56	48	#
Hispanic (%)	a	a		a	a		a	a		a	a	
Race--White (%)	77	81		65	68		55	57		74	82	#
Race--African American (%)	23	19		35	32		45	43		26	18	
Race--Other (%)	a	a		a	a		a	a		a	a	
Family income less than $30,000 (%)	39	41		57	44		48	62		41	56	
Family income between $30,000 and $60,000 (%)	52	42		20	31		32	38		41	14	* #
Family income over $60,000 (%)	9	17		23	25		a	a	*	18	30	#
Eligible for free or reduced price lunch (%)	45	49		47	36		37	56	*	42	48	
Has any learning or other disability (%)	38	46		36	21		32	24		30	43	
Mother has bachelor's degree or higher (%)	14	9		13	24		a	a		19	11	
Screening Tests												
TOWRE Sight Word Efficiency	84.6	82.0		85.5	85.1		87.0	83.6	*	85.3	82.2	
TOWRE Phonemic Decoding Efficiency	84.2	85.1		85.7	85.4		86.2	85.5		85.7	87.1	
Peabody Picture Vocabulary Test--Revised	93.1	94.6		95.3	98.0		91.0	89.1		97.6	96.5	#
Baseline Tests												
WRM Word Identification	89.1	87.2		89.5	86.9		90.5	88.9		89.7	87.7	
TOWRE Phonemic Decoding Efficiency	84.4	84.3		86.1	84.7		86.7	85.3		87.1	85.9	
WRM Word Attack	90.5	89.2		93.7	91.3		94.5	92.7		93.8	94.7	
TOWRE Sight Word Efficiency	87.4	84.5		89.4	86.5	*	88.9	84.5	*	86.9	84.0	
AIMSweb (Raw score)	38.2	33.6		46.8	41.8		49.1	40.3	*	43.4	34.4	
WRM Passage Comprehension	91.2	88.5		95.1	89.1	*	93.5	91.7		94.2	89.7	
GRADE	86.3	84.9		87.8	82.9		88.3	85.2	*	89.8	84.1	*
Woodcock Johnson Spelling	90.2	86.5		89.5	88.9		89.3	87.1		90.5	85.9	
Woodcock Johnson Calculation	92.7	96.8		99.2	95.6	*	96.8	91.5		96.9	92.8	*
Sample Size	51	38		57	34		51	19		44	35	

Note: Weights were 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. The mean raw scores for AIMSweb tests, administered to students across the country in the fall in the school years 2000-2001, 2001-2002, and 2002-2003, were 75 and 112 for third and fifth graders, respectively. The respective standard deviations were 39 and 47.

* 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.

(continued)

[12] Parents did not provide consent for 250 of the 1,042 test-score eligible students. Another 13 students were excluded for "other" reasons, such as home schooling, grade retention, expulsion, or a determination that the interventions would not be appropriate in light of a student's specific disability.

Table II.3

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

Baseline Means	Failure Free Reading			Spell Read			Wilson Reading			Corrective Reading			
	Treat.	Cont.		Treat.	Cont.		Treat.	Cont.		Treat.	Cont.		
Student Characteristics													
Age	10.6	10.6		10.9	10.7		10.8	10.5	*	10.7	10.6		
Male (%)	53	51		55	58		54	66		49	59		
Hispanic (%)	a	a		a	a		a	a		a	a		
Race--White (%)	78	83		76	67		55	59		83	86		
Race--African American (%)	22	17		24	33		45	41		18	14		
Race--Other (%)	a	a		a	a		a	a		a	a		
Family income less than $30,000 (%)	41	50		51	54		73	47	*	32	46		
Family income between $30,000 and $60,000 (%)	43	33		39	34		23	36		43	36		
Family income over $60,000 (%)	16	17		10	12		a	a	*	25	18		
Eligible for free or reduced price lunch (%)	42	46		53	41	*	58	41	*	41	47		
Has any learning or other disability (%)	27	36		29	30		32	28		29	29		
Mother has bachelor's degree or higher (%)	12	17		5	9		a	a	*	15	29	*	
Screening Tests													
TOWRE Sight Word Efficiency	84.1	85.3		84.1	85.3		83.1	85.0		82.7	83.5		
TOWRE Phonemic Decoding Efficiency	81.7	79.8	*	78.1	79.6		82.3	82.3		80.1	81.2		#
Peabody Picture Vocabulary Test--Revised	94.8	95.1		92.3	92.6		91.8	99.7	*	95.2	98.3	*	#
Baseline Tests													
WRM Word Identification	90.4	89.0		87.1	87.7		87.8	90.2		87.6	89.7		
TOWRE Phonemic Decoding Efficiency	82.1	81.8		77.5	80.0		82.6	81.5		80.9	82.0		#
WRM Word Attack	93.5	92.7		90.5	92.3		93.1	94.6		93.7	95.3		
TOWRE Sight Word Efficiency	84.2	85.7		83.1	85.7	*	83.8	84.9		83.6	82.5		
AIMSweb (Raw score)	79.0	75.6		79.4	79.3		74.2	81.2	*	76.1	71.5		
WRM Passage Comprehension	92.4	92.2		91.3	92.8		90.3	96.5	*	91.9	93.7		
GRADE	91.4	92.2		89.8	89.2		91.9	95.1		88.3	92.0		
Woodcock Johnson Spelling	93.8	92.2		89.8	91.4		91.0	92.4		88.6	88.4		
Woodcock Johnson Calculation	94.0	93.5		94.9	94.5		93.3	95.0		94.1	94.5		
Sample Size	62	66		56	44		52	36		54	30		

Note: Weights were 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. The mean raw scores for AIMSweb tests, administered to students across the country in the fall in the school years 2000-2001, 2001-2002, and 2002-2003, were 75 and 112 for third and fifth graders, respectively. The respective standard deviations were 39 and 47.

* 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 Sample for Three Word-level Interventions, by Treatment Status: 3rd Grade Analysis Sample

Baseline Means	All Interventions			Word-level Interventions		
	Treatment	Control		Treatment	Control	
Student Characteristics						
Age	8.7	8.7		8.7	8.7	
Male (%)	56	47		57	43	
Hispanic (%)	2	3		a	a	
Race--White (%)	68	72		65	69	
Race--African American (%)	32	28		35	31	
Race--Other (%)	a	a		a	a	
Family income less than $30,000 (%)	47	50		50	53	
Family income between $30,000 and $60,000 (%)	36	33		30	29	
Family income over $60,000 (%)	17	18		20	18	
Eligible for free or reduced price lunch (%)	43	47		42	46	
Has any learning or other disability (%)	34	33		33	29	
Mother has bachelor's degree or higher (%)	13	13		12	14	
Screening Tests						
TOWRE Sight Word Efficiency	85.6	83.3	*	85.9	83.7	*
TOWRE Phonemic Decoding Efficiency	85.4	85.7		85.9	85.9	
Peabody Picture Vocabulary Test--Revised	94.2	94.6		94.6	94.6	
Baseline Tests						
WRM Word Identification	89.7	87.7	*	89.9	87.8	*
TOWRE Phonemic Decoding Efficiency	86.0	85.0		86.6	85.3	*
WRM Word Attack	93.0	91.8		94.0	92.8	
TOWRE Sight Word Efficiency	88.2	84.9	*	88.5	85.1	*
AIMSweb (Raw score)	44.3	37.6	*	46.5	39.0	*
WRM Passage Comprehension	93.5	89.7	*	94.3	90.2	*
GRADE	88.0	84.3	*	88.6	84.0	*
Woodcock Johnson Spelling	89.9	87.2		89.7	87.4	
Woodcock Johnson Calculation	96.4	94.3	*	97.7	93.4	*
Sample Size	203	126		152	88	

Note: Weights were 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. The mean raw scores for AIMSweb tests, administered to students across the country in the fall in the school years 2000-2001, 2001-2002, and 2002-2003, were 75 and 112 for third and fifth graders, respectively. The respective standard deviations were 39 and 47.

* 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 Sample for Three Word-level Interventions, by Treatment Status: 5th Grade Analysis Sample

Baseline Means	All Interventions			Word-level Interventions		
	Treatment	Control		Treatment	Control	
Student Characteristics						
Age	10.7	10.6	*	10.8	10.6	*
Male (%)	53	58		53	61	
Hispanic (%)	a	a		a	a	
Race--White (%)	73	74		72	71	
Race--African American (%)	27	26		28	29	
Race--Other (%)	a	a		a	a	
Family income less than $30,000 (%)	48	49		51	49	
Family income between $30,000 and $60,000 (%)	38	35		36	35	
Family income over $60,000 (%)	14	16	*	14	16	*
Eligible for free or reduced price lunch (%)	48	44		51	43	
Has any learning or other disability (%)	29	31		30	29	
Mother has bachelor's degree or higher (%)	8	18	*	7	18	*
Screening Tests						
TOWRE Sight Word Efficiency	83.5	84.8		83.3	84.6	
TOWRE Phonemic Decoding Efficiency	80.5	80.7		80.1	81.0	
Peabody Picture Vocabulary Test--Revised	93.6	96.4	*	93.2	96.8	*
Baseline Tests						
WRM Word Identification	88.2	89.1		87.5	89.2	
TOWRE Phonemic Decoding Efficiency	80.7	81.3		80.2	81.2	
WRM Word Attack	92.7	93.7		92.4	94.0	
TOWRE Sight Word Efficiency	83.7	84.7		83.5	84.3	
AIMSweb (Raw score)	77.2	76.8		76.6	77.2	
WRM Passage Comprehension	91.5	93.7		91.2	94.3	*
GRADE	90.3	92.1		89.9	92.0	
Woodcock Johnson Spelling	90.8	91.1		89.8	90.7	
Woodcock Johnson Calculation	94.1	94.3		94.1	94.6	
Sample Size	224	176		162	110	

Note: Weights were 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. The mean raw scores for AIMSweb tests, administered to students across the country in the fall in the school years 2000-2001, 2001-2002, and 2002-2003, were 75 and 112 for third and fifth graders, respectively. The respective standard deviations were 39 and 47.

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

a Values suppressed to protect student confidentiality.

- ***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.[13] Within each school unit and grade, 3, 6, or 12 eligible students were randomly chosen to receive the intervention.[14] 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.

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).[15] 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. [16] 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 and very little at 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.[17]

For this report, we estimate impacts on tests that we administered for this evaluation (the Word Attack, Word Identification, and Passage Comprehension subtests of the WRMT-R; the Phonemic Decoding

[13] 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).

[14] 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).

[15] Among third graders, the difference in Peabody Picture Vocabulary test scores between eligible and ineligible students was not statistically significant at the 0.05 level. The scores were significantly different between eligible and ineligible fifth-grade students.

[16] The group of 263 students includes 13 students who, as described above, were excluded for "other" reasons, such as home schooling or determination that the interventions would be inappropriate in light of a student's specific disability.

[17] 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 dropouts 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.

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.

Efficiency and Sight Word Efficiency subtests of the TOWRE; and the GRADE) and on the reading and mathematics tests of the PSSA. The tests that we administered were taken a full year after the "intervention year," that is, a full year after the interventions took place, when most students were in the 4th and 6th grades. The PSSA tests were taken at the end of the intervention year, while the students were still third and fifth graders. Because much of our analysis refers to the time period when the students were no longer in the third and fifth grade, we will refer to the students as the third- and fifth grade- "cohorts," referencing their grade when the interventions took place.

We constructed one analysis sample for the tests that we administered and one for the PSSA tests. We derived weights for the samples as described in detail in Appendix C. The analysis samples include 729 and 737 students, respectively, fewer than the 772 students subject to random assignment. The first sample excludes 34 students whom we were unable to test, and the second excludes 26 students for whom we were unable to obtain PSSA scores. Both samples exclude an additional 9 treatment students in one school unit that was assigned to Corrective Reading but did not have enough eligible students to form a control group. 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.[18] Tables II.2-II.5 present baseline characteristics for students in the first analysis sample. Baseline characteristics for students in the second analysis sample are similar and are reported in Appendix G.

Table II.7

Comparison of Consenting and Nonconsenting Students, Among All Eligible

	Consenting	Not consenting	
Screening test scores	Mean	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.

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 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 33rd 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

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

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 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).[19]

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 and grade cohorts.

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. 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 these unequal treatment probabilities, ensuring that when weighted, the treatment and control students 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. Full details of the weighting procedure, including the adjustment of weights to compensate for nonresponse, are given in Appendix C.

[19] 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, most of the significant differences that are scattered across characteristics and interventions are no longer significant. Likewise, nearly all differences for the four interventions combined and the three word-level interventions combined become insignificant when we apply the Bonferroni correlation. Most of the differences for the third-grade cohort and some of the differences for the fifth-grade cohort remain insignificant when we use the Benjamini-Hochberg method. 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.

B. DATA

Test data and other information on students, parents, teachers, classrooms, and schools were collected several times over a two-year period. The second follow-up data collection at the end of the two-year period focused on two main types of information: measures of student performance and measures of student characteristics and the instruction they received.[20] These data are described next. The data collected before the second follow-up are described in Torgesen et al. (2006).

1. Measures of Student Performance

In this report, the tests used to assess student performance fall into three categories. The first category includes seven measures of reading skill that we administered for this evaluation to assess student progress in learning to read. The second category includes two other measures that we administered. 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. Descriptions of each of the seven reading tests, the spelling test, and the calculation test can be found in Exhibit 1 at the end of this chapter. The third category of student performance measures includes two tests—one in reading and one in mathematics—administered by the AIU schools as part of the Pennsylvania System of School Assessment (PSSA). These PSSA tests were administered near the end of the school year (2003-04) during which the reading interventions took place, in contrast to the nine tests in the first two categories, which we administered approximately one year later—one year after the interventions ended.

a. Measures of Reading Administered for the Evaluation

The seven measures of reading skills administered for the evaluation 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

[20] See Appendix B for a detailed description of the second follow-up data collection activities.

- 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)

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.[21,22] 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.[23] For the AIMSweb passages, the score used in this analysis is the median correct words per minute from three grade-level passages. (See the note on Table II.2 for more information about the means and standard deviations for the scores on the AIMSweb tests.) Table II.8 shows estimates of test reliability, and Tables II.9 and II.10 present correlations between tests for the third-grade students and fifth-grade students, respectively. The shaded boxes in Tables II.9 and II.10 indicate tests that measure similar constructs: tests measuring phonemic decoding skills, tests measuring reading fluency and accuracy, and tests measuring reading comprehension.

Even though the tests can be grouped by the skills they measure, the correlations—even between tests measuring similar constructs—were not always large. For example, for the third-grade and fifth-grade students, respectively, the correlations between the Word Attack and Phonemic Decoding Efficiency tests were, 0.65 and 0.69, the average correlations among the three tests measuring word reading accuracy and fluency were 0.70 and 0.69, and the correlations between the Passage Comprehension and GRADE tests were 0.47 and 0.44. 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 0.91 between the Word Attack and Phonemic Decoding Efficiency tests for a sample of at-risk third-grade students. A correlation of 0.87 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 0.92 and 0.86, respectively. The manual

[21] When possible, we standardized scores to the grade and month of administration.

[22] 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.

[23] 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.

for the WRMT-R (Woodcock 1998) reports a correlation between the Word Identification measure and Passage Comprehension measure of 0.67 for third graders and 0.59 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.

b. Measures of Spelling and Mathematics Calculation Ability Administered for the Evaluation

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. Table II.8 includes estimated reliabilities for these tests as well as the seven reading tests.

c. Measures of Reading and Mathematics from the PSSA

The PSSA is a standards-based, criterion-referenced assessment used to measure the attainment in certain areas of each Pennsylvania student in particular grades (http://www.pde.state.pa.us/a_and_t/site/default.asp). In the 2003-04 school year, students in grades three and five completed the PSSA in reading and math. (Fifth-grade students were also assessed in writing.) The test was administered to all students during late March and early April of the year in which they received the interventions.

Guided by the Pennsylvania Academic Standards, advisory committees of Pennsylvania educators developed the content of the PSSA tests, which include both multiple-choice and open-ended items. The PSSA reading tests are designed to measure students' skills in comprehending text. The test for third graders covers three standards: learning to read independently; reading critically in all content areas; and reading, analyzing, and interpreting literature. The test for fifth graders covers those three standards plus two others: English language characteristics and research. The PSSA mathematics tests cover 11 standards and measure skills ranging from recalling specific facts to solving problems. Scaled scores for the PSSA tests are derived using item response theory (Rasch) models. Estimates of test reliability vary by subject and grade, but generally exceed 0.8 and are often at least 0.9 (internal consistency reliability).

Table II.8

Tests Administered for the Second Follow-Up (End of the 2004-05 School Year)

Test	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]
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

Table II.9

Correlations Among Reading Tests, 3rd Grade Analysis Sample

	Word Attack	TOWRE PDE	Word Identification	TOWRE SWE	AIMSweb	Passage Comprehension	GRADE
Word Attack	1.00	0.65	0.56	0.46	0.38	0.58	0.35
TOWRE PDE		1.00	0.53	0.54	0.53	0.47	0.29
Word Identification			1.00	0.71	0.58	0.67	0.43
TOWRE SWE				1.00	0.80	0.67	0.46
AIMSweb					1.00	0.62	0.47
Passage Comprehension						1.00	0.47
GRADE							1.00

Table II.10

Correlations Among Reading Tests , 5th Grade Analysis Sample

	Word Attack	TOWRE PDE	Word Identification	TOWRE SWE	AIMSweb	Passage Comprehension	GRADE
Word Attack	1.00	0.69	0.70	0.49	0.47	0.50	0.34
TOWRE PDE		1.00	0.65	0.66	0.60	0.44	0.35
Word Identification			1.00	0.64	0.64	0.53	0.39
TOWRE SWE				1.00	0.78	0.51	0.35
AIMSweb					1.00	0.54	0.43
Passage Comprehension						1.00	0.44
GRADE							1.00

2. Measures of Student Characteristics and Instruction Received

During the second follow-up, we obtained measures of student characteristics and instruction received with a classroom teacher survey and a school records form.

a. Classroom Teacher Survey

Each child's regular classroom teacher completed a survey near the end of spring 2005. The survey asked the teacher to characterize the reading instruction the 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. The teacher also provided classroom behavior ratings for the child. The behavior rating scales were adapted from the Multigrade Behavior Inventory (Agronin et al. 1992) and Iowa-Connors Teacher Rating Scale (Loney and Milich l982).

b. School Records Form

At the end of the 2004-05 academic year, we obtained data on each student using a school records form. We collected information on enrollment, attendance, and suspensions; characteristics such as limited English proficiency status, eligibility for the free or reduced-price lunch program, and disabilities; reading services; Individual Education Plan (IEP) or Service Agreement specifications; grade promotion and retention; course grades; and reading and math standardized test scores.

EXHIBIT 1. SECOND FOLLOW-UP 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.

III. IMPLEMENTATION ANALYSIS

The purpose of this evaluation is to estimate the impact of four reading interventions when they are delivered with as much fidelity and skill as can be attained in a standard school setting. Our procedures to ensure high quality implementation of the interventions included careful selection of teachers to deliver the interventions, training and supervision of intervention teachers 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 minimize disruptions so that each student could receive at least 100 hours of instruction. We then collected information to evaluate the quality and fidelity of the intervention implementations, as well as to understand how the interventions fit into the overall reading instruction for each child. A detailed discussion of our findings from this assessment, and a description of the procedures for selecting, training, and supporting the intervention teachers can be found in Torgesen et al. (2006). In this report, we summarize the key implementation findings from the prior report and present some new findings pertaining to students' hours of reading instruction in the year after the interventions.

As described in Torgesen et al. (2006), we used a variety of instruments to evaluate implementation. These included documentation of the amount of training received by each of the intervention teachers, daily attendance logs for all intervention sessions, video-tapes of a sample of intervention sessions, ratings of the intervention teachers by program trainers and study coordinators, and questionnaires completed for each participating student by his or her classroom teacher. The video tapes were used in three separate analyses: an analysis of program fidelity and general teacher quality, a verification of session length (which had implications for total hours of instruction), and an analysis of intervention program content. The classroom teacher surveys were used to describe each student's total reading instruction; these surveys were completed for both the intervention year and the year following the intervention.

Later in this section, we present an integrated discussion of our findings regarding hours of total reading instruction during both years of the study. We begin, however, with a brief summary of our other key findings from the first year of the study and a description of the instructional elements and procedures for each of the four interventions.

A. SUMMARY OF KEY IMPLEMENTATION FINDINGS

Hours of instruction. The large majority (93 percent) of students in the treatment group received at least 80 hours of intervention instruction. This represents a sustained and substantial level of exposure to intensive instruction, even though only 14 percent of intervention students received the intended dose of 100 hours. There were no significant differences in average hours of instruction across interventions, although fifth-grade students received fewer hours of intervention (88 hours) on average than did third-grade students (93 hours).

Heterogeneity of instructional groups. Due to the practical constraints imposed by the incidence and diversity of reading difficulties among third and fifth graders in the AIU schools participating in this evaluation, the instructional groups formed for the intervention were heterogeneous with regard to their beginning word-level skills. At each grade level, the average difference between the highest and lowest baseline Word Attack scores among the three students in an instructional group was about one standard deviation. Nonetheless, the program developers indicated in follow-up conversations that this amount of

within-group heterogeneity was not unusual in comparison with what they normally observe when delivering their interventions in other settings.[24]

Training of intervention teachers. Representatives of the four reading programs trained the intervention teachers. On average, teachers received almost 70 hours of professional development over the course of the intervention, starting with five days of intensive training for all teachers in August 2003. The total amount of professional development varied across the reading programs, but all of 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.

Trainer ratings of intervention teachers. The trainers from each reading program rated the teaching performance of teachers under their supervision. According to the trainers, the average instruction teacher's performance fell somewhere in the top half among similarly experienced teachers whom they had observed. In addition, the trainers' average ratings on five dimensions of program fidelity and three dimensions of general teacher quality were well above the satisfactory level for all dimensions and all programs.

Video analysis of intervention teachers. Each teacher was videotaped twice over the course of the intervention, and the videos were analyzed for adherence to program guidelines. Each program had slightly different dimensions along which fidelity was assessed, as well as different criteria for judging adequacy. Deviations from criterion were judged by members of the evaluation team as minor, moderate, or extreme. Overall, there were no extreme deviations and relatively few moderate deviations. The moderate deviations that did occur were primarily with regard to time in session (most sessions ran shorter than anticipated) and fine points of program technique.

Time by activity analysis. The videotaped instructional sessions were also analyzed to determine how teachers allocated time across instructional activities. The analysis showed that the distribution of time between word-level and vocabulary/comprehension activities did not conform to the categorization of the interventions in the original study design (which was based on the description of instructional activities from the program providers). As a consequence, the programs were regrouped for analysis, with the three programs that devoted most of their instructional time to improving word-level reading skills grouped together.

More detailed information pertaining to these and other results from our implementation analysis can be found in Torgesen et al. (2006).

B. DESCRIPTION OF THE INTERVENTIONS

A description of the essential instructional elements and procedures of each of the four instructional methods, as they were implemented in this study, is provided below, along with results from Torgesen et al. (2006) about the relative amount of time devoted to instruction in word-level skills versus vocabulary and comprehension.

[24] Furthermore, we previously found no consistent patterns in the relationship between instructional group heterogeneity and students' reading outcomes (Torgesen et al. 2006).

Interventions that focused primarily on word level skills

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 this study, includes four levels—A, B1, B2, and C. Placement testing is used to start each group at the appropriate level. However, because the instructional groups in this study were relatively heterogeneous in terms of their beginning skills, there was not always 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 study.

Estimated allocation of instructional time: 74 percent on word-level skills and 26 percent on comprehension/vocabulary.

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 these components 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.

Estimated allocation of instructional time: 94 percent on word-level skills and 6 percent on comprehension/vocabulary.

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 Phase C.

The Spell Read intervention had originally been one of the two "word-level plus comprehension" interventions, but after the time-by-activity analysis, we determined that it was more appropriately classified as a "word-level" intervention. Because the word-level instructional content in Spell Read is more structured than the instruction designed to build reading comprehension, the relatively short instructional sessions in this study led to a different balance of word-level and comprehension instruction than was anticipated. That is, to accomplish the highly specified word-level instruction contained in the program, the teachers reduced the amount of time they spent on the comprehension components. In clinical settings, Spell Read is typically provided in 70 minute sessions, whereas the sessions in this study averaged closer to 55 minutes in length.

Estimated allocation of instructional time: 83 percent on word-level skills and 17 percent on comprehension/vocabulary.

Intervention that focused on word level skills, vocabulary, and comprehension

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 one 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.

Estimated allocation of instructional time: 48 percent on word-level skills and 52 percent on comprehension/vocabulary.

C. TOTAL HOURS OF READING INSTRUCTION

In addition to the hours spent in the experimental treatment, students in the intervention group also received some reading instruction in their regular classrooms. Control students, in accordance with the study design, received the mix of reading services that would normally be provided by their schools. To better understand the treatment contrast, we examined total reading instruction provided to treatment and control students during the intervention year, and again during the following year. The intervention year findings were initially presented in Torgesen et al. (2006). They are repeated here, disaggregated by grade cohort, to facilitate comparison with results for the following year.

Hours of reading instruction were based on annual surveys, which were completed for each student in the study by his or her regular classroom teacher. The surveys included questions on the duration of reading instruction provided for that student during a typical week, the sizes of the groups in which instruction was delivered, and the types of professionals who provided the instruction. For students in the intervention group, a constant amount of 4.5 hours (per week) of "treatment" small-group instruction was added during the intervention year. No effort was made to characterize the nonintervention instruction based on types of reading activities (e.g., activities to build word-level skills versus activities to develop comprehension skills or vocabulary).

For our analysis, hours of reading instruction were categorized according to group size (large groups, small groups, and one-on-one) and also according to whether the teacher was a general education teacher, a specialist teacher, or one of the treatment (i.e., intervention) teachers. "Specialist teacher" was defined as a special education teacher, a Title I teacher, an ESL teacher, a reading specialist, or other instructor.

We analyzed the data on hours of instruction using a two-level hierarchical linear model, with students and school units making up the two levels. The analyses were similar to the analyses of impacts on reading test scores, and a more detailed explanation of those procedures is provided in Chapter IV.

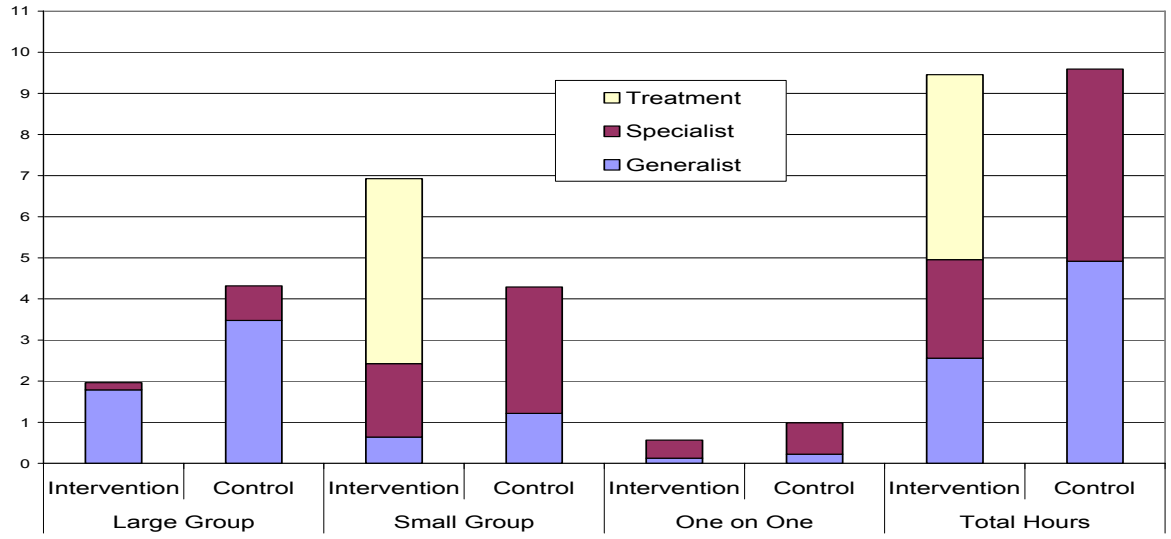
Figures III.1 and III.2 present information on the average weekly hours of reading instruction received by students in the two grade cohorts during the intervention year.

For the third-grade cohort (Figure III.1), it is noteworthy that intervention and control students received approximately the same amount of total reading instruction (9.5 and 9.6 hours, respectively) during the intervention year. This implies that, for this grade cohort, reading instruction delivered by the intervention generally substituted for, rather than added to, the students' other reading instruction. Compared to the control group, however, students in the intervention group received more hours of small group instruction [$t(280) = 2.63$, $p < 0.0001$] and fewer hours of large group instruction [$t(280) = -3.77$, $p < 0.0001$]. More specifically, students in the intervention group received 6.9 hours of small group instruction and 2.0 hours of large group instruction, while students in the control group received about 4.3 hours of each.

As can be seen in Figure III.2, reading instruction followed a different pattern in the fifth-grade cohort. At this grade level, students in the control group received only 7.8 hours of total reading instruction per week, which was significantly lower than the 9.4 hours of total reading instruction for the intervention group [$t(347) = 2.35$, $p = 0.02$]. As at grade three, the small group hours for the intervention group (6.6 hours) were significantly higher than for the control group ($t(346)=11.97$, $p = <.0001$), while the large group hours (2.5 hours) were significantly lower ($t(347)=-2.72$, $p=0.0069$).

Figure III.1

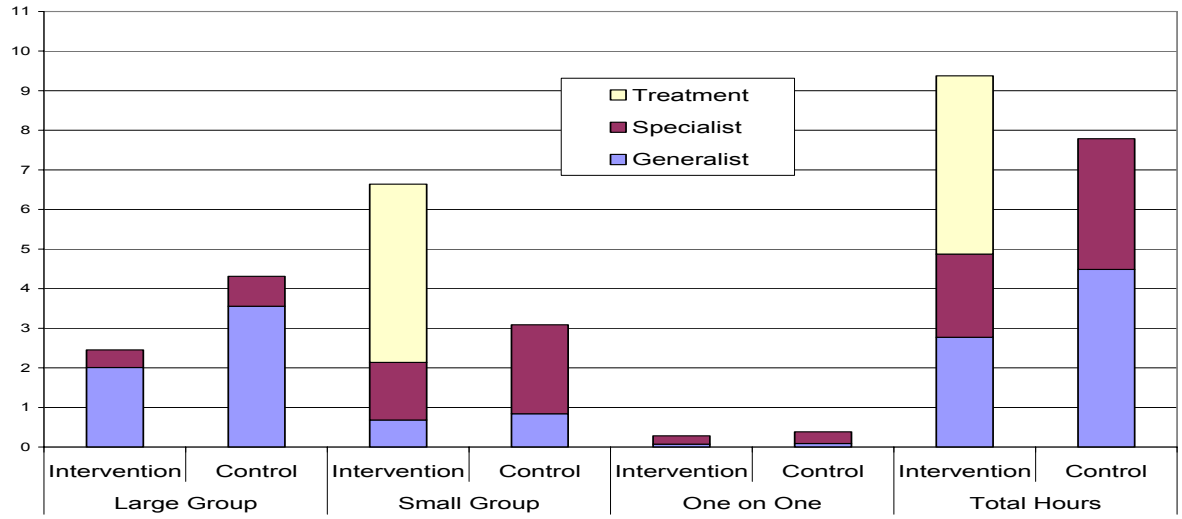
Average Hours of Reading Instruction per Week During Intervention Year: Third-Grade Cohort.



Note: A "Specialist" is defined as a teacher who is not a general education teacher. For example, specialists include special education and Title I teachers and reading specialists.

Figure III. 2

Average Hours of Reading Instruction per Week During Intervention Year: Fifth-Grade Cohort.

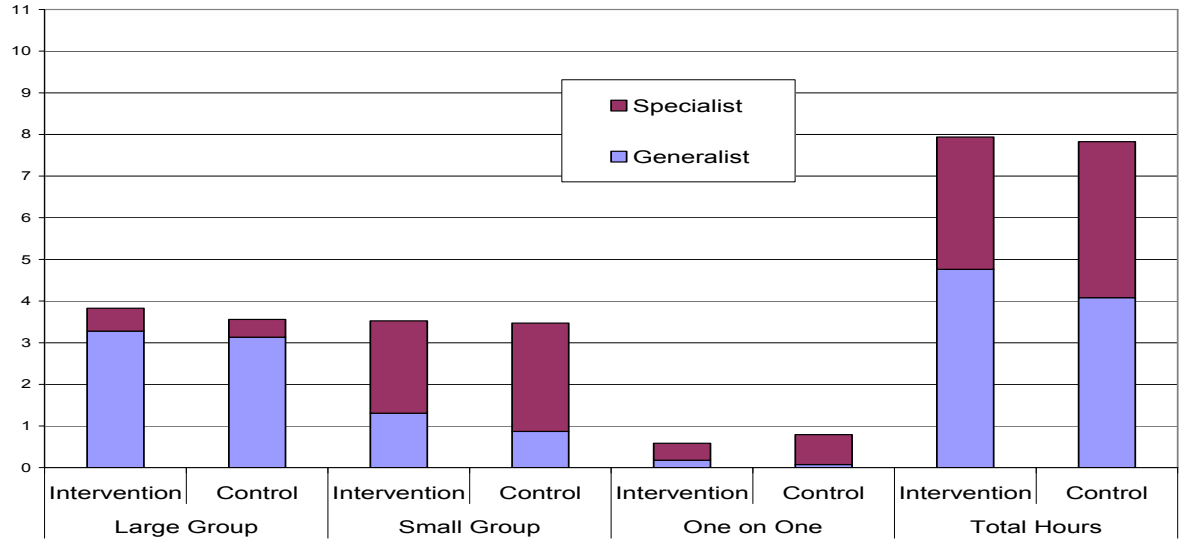


Note: A "Specialist" is defined as a teacher who is not a general education teacher. For example, specialists include special education and Title I teachers and reading specialists.

Findings for the year following the interventions are presented in Figures III.3 and III.4. By this time, most students in the third-grade cohort had transitioned into grade four, while most students in the fifth-grade cohort had transitioned into grade six. Although practices differed among the several participating districts, moving to grade six also meant moving to middle school for many students in the study.

Figure III.3

Average Hours of Reading Instruction per Week During Year Following Intervention: Third-Grade Cohort.

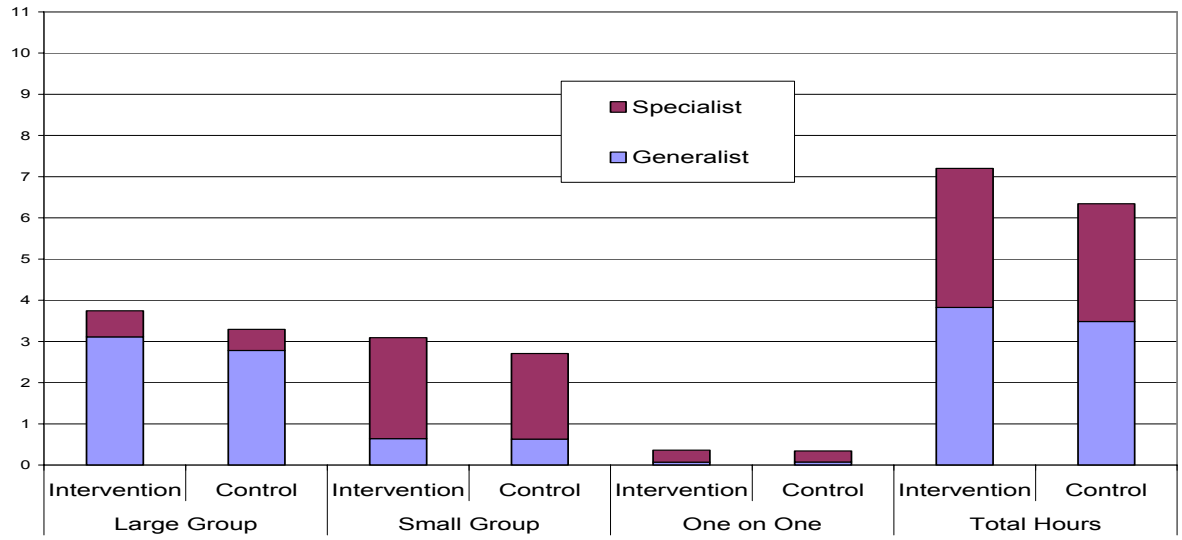


Note: A "Specialist" is defined as a teacher who is not a general education teacher. For example, specialists include special education and Title I teachers and reading specialists.

For the third-grade cohort (Figure III.3), there were no significant differences in the total hours of instruction received by the intervention and control groups during the year following the intervention (7.9 and 7.8 hours, respectively). Neither were there any significant differences by group size or by type of instructor.

Figure III.4

Average Hours of Reading Instruction per Week During Year Following Intervention: Fifth-Grade Cohort.



Note: A "Specialist" is defined as a teacher who is not a general education teacher. For example, specialists include special education and Title I teachers and reading specialists.

For the fifth-grade cohort, on the other hand, Figure III.4 shows that intervention students continued to receive significantly more total hours of reading instruction (7.2 hours) than the control students (6.3 hours) [$t(360) = 2.29$, $p = 0.02$]. There is no obvious reason for this finding, and a comparison of surveys completed for the same students across years did not exhibit much consistency at the individual student level ($r = 0.054$).

In summary, the analysis of total reading instruction indicated that, at grade three, reading instruction provided by the intervention did not increase the total hours of reading, but did shift hours between large group and small group instructional formats. At grade five, the intervention resulted in an increase in total hours of reading instruction, as well as a shift toward more hours in the small group format. During the year following the intervention, total hours of instruction decreased somewhat for all students, presumably as a function of the role of reading in the curriculum at successively higher grade levels. There was no relationship between treatment status and hours of instruction for the third-grade cohort, while intervention group students continued to receive somewhat more instruction than control students in the fifth-grade cohort. The data do not reveal why this latter relationship persists into the year after the interventions.

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 students in the third-grade cohort and all students in the fifth-grade cohort, 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.[25] 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, we estimated intervention impacts and standard errors using a weighted two-level hierarchical linear model (HLM) that allows for nested data.[26] 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). Therefore, we estimated impacts separately for the third- and fifth-grade cohorts (but test for differences between impacts for the two cohorts). The model is:

Level One: Student *i* within school unit *j*

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

[25] The sample includes 31 school units with about 730 students; one school unit dropped out of the study after random assignment, but before learning about the intervention to which it had been assigned.

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

Level Two: School unit (j)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}A_j + \gamma_{02}B_j + \gamma_{03}C_j + \sum_{l=1}^{3}\xi_{0l}P_{lj} + \mu_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}A_j + \gamma_{12}B_j + \gamma_{13}C_j + \sum_{l=1}^{3}\xi_{1l}P_{lj} + \mu_{1j} \qquad \text{(IV.2)}$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21}A_j + \gamma_{22}B_j + \gamma_{23}C_j + \sum_{l=1}^{3}\xi_{2l}P_{lj} + \mu_{2j}$$

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;

$A_j = 1$ 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:

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

where $\overline{y}_{..}$ is the weighted mean of the pretest score across all students in a given grade cohort in the evaluation sample. By mean-centering the pretest score (that is, the baseline score), 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 cohort (third or fifth) with a baseline test score equal to the average baseline test score across students in that grade cohort.

The level-one model (Equation (IV.1)) relates a student's post-intervention test score to a treatment indicator, the student's pretest score, and a residual term (unexplained variation). The level-two model (Equation (IV.2)) relates the level-one parameters (coefficients $\beta_{0j}, \beta_{1j},$ and β_{2j}) to indicators for the interventions to which the school units were randomly assigned and the blocking strata. The interventions Failure Free Reading, Spell Read, Wilson Reading, and Corrective Reading are denoted as A, B, C, and D, respectively.[27] 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.[28]

The main parameters of interest in our two-level model 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.[29,30] Such an impact 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 the most robust estimate of program impact because it involves the fewest assumptions when estimating the impact. By imposing more assumptions, which might not be valid, we could also estimate the intervention impacts on those who participated in the interventions, and on those who participated substantially, receiving at least 80 hours of instruction, for example. 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, as discussed in more detail below.

From the HLM model, we estimate impacts for each of the four interventions.[31] 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:

---

[27] 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).

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

[29] 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.

[30] Appendix D provides details on deriving the impacts from estimated model parameters.

[31] We used HLM 6 ® 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).

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)

We obtain these first two impacts as follows:

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

where the intervention impacts for Failure Free Reading (A), Spell Read (B), Wilson Reading (C), and Corrective Reading (D), respectively, are:[32]

$$\begin{aligned}I_A &= \hat{\gamma}_{10} + \hat{\gamma}_{11} + (1/4)(\hat{\xi}_{11} + \hat{\xi}_{12} + \hat{\xi}_{13})\\I_B &= \hat{\gamma}_{10} + \hat{\gamma}_{12} + (1/4)(\hat{\xi}_{11} + \hat{\xi}_{12} + \hat{\xi}_{13})\\I_C &= \hat{\gamma}_{10} + \hat{\gamma}_{13} + (1/4)(\hat{\xi}_{11} + \hat{\xi}_{12} + \hat{\xi}_{13})\\I_D &= \hat{\gamma}_{10} + (1/4)(\hat{\xi}_{11} + \hat{\xi}_{12} + \hat{\xi}_{13}).\end{aligned} \tag{IV.5}$$

When the interventions are grouped, each intervention in the group receives equal weight. We derive the impacts for the four individual interventions according to Equation (IV.5).

These impacts 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. From a policy perspective, the ITT shows the impact of being offered the opportunity to participate in an intervention. In this study, a few students assigned to the treatment group did not participate in one of the interventions. Estimates of the impact of the treatment on the treated (TOT) adjust for students not participating in the intervention.

If there is interest in estimating the impact of the treatment on those who participated—the TOT impact—there might also be interest in estimating the impact of the treatment on those who received a "full dose" of the treatment, which we might define in this evaluation as receiving at least 80 hours of instruction. A crude, but very simple approach to deriving such an impact is to apply the same methods used to obtain TOT estimates.

A TOT impact takes into account the treatment received by students in the study, but requires additional assumptions that are untestable.[33] In this evaluation, a small number of students assigned to the

[32] 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 $\frac{1}{4}(\hat{\xi}_{11} + \hat{\xi}_{12} + \hat{\xi}_{13} + 0)$.

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, the threshold for receiving a full dose of the intervention.

In this study, with no control group students who received 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, as documented in Torgesen et al. (2006), 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. Because the TOT impacts rely on untestable assumptions, and are not substantially different from the ITT impacts, we present only the ITT estimates in this report.

In addition to estimating impacts for all students in the third-grade cohort and all students in the fifth-grade cohort, we estimated impacts for subgroups of students within each grade cohort. 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 a grade cohort) for a subgroup (see Appendix D).[34]

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

(continued)

[33] Two major assumptions are involved in estimating 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 would have to use caution in interpreting TOT estimates for students with a "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. See Bloom (1984); Angrist, Imbens, and Rubin (1996); or Little and Rubin (2000) for general background information on computing TOT estimates. A TOT estimate is sometimes referred to as the Complier Average Causal Effect (CACE) or the Instrumental Variables (IV) estimate.

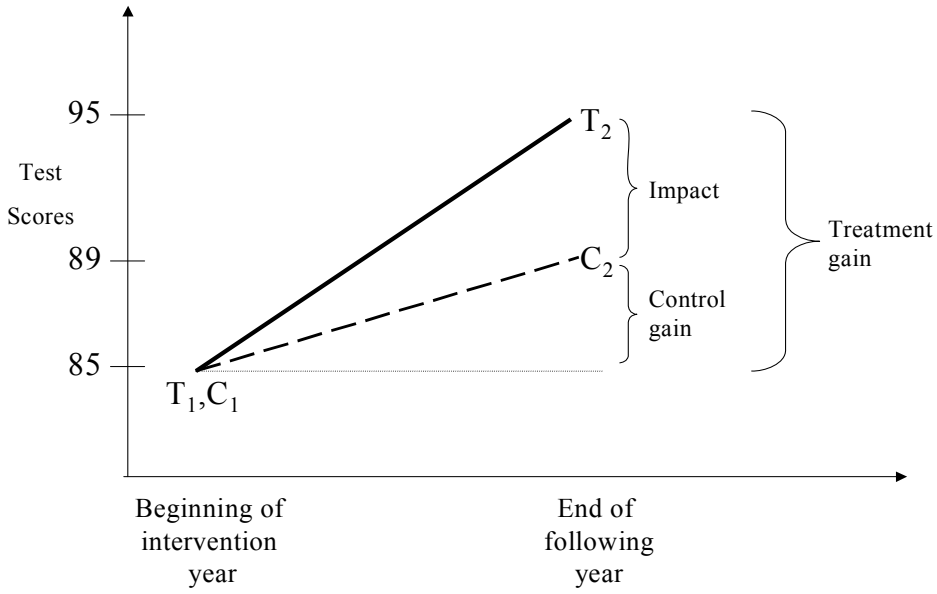
[34] Our preliminary analyses showed substantial differences in impacts by grade. Therefore, for each subgroup, we estimate separate impacts by grade. When we designed the study, our power analyses assumed that we could combine grades when conducting subgroup analyses. Because we cannot, our ability 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.

in which the interventions were operating: (1) where the students began in terms of reading ability just prior to the interventions, (2) how much improvement the students would have made in the absence of the interventions, and (3) the amount of the interventions that treatment and control students actually received.

We illustrate the first two elements using a hypothetical example, in Figure IV.1.[35] At the beginning of the intervention 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).[36] This is similar to the actual baseline test scores seen for students in this study (see Tables II.2 through II.7).

Figure IV.1

Hypothetical Example of Gains and Impact



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 of the intervention year and end of the following 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 zero if we had a national sample of students at all

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

[36] 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.

levels of reading ability. However, the students in this example (and in the actual study) began 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 either the post-test scores for the intervention and control groups or the test score gains for the intervention and control groups. Using the post-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 of the intervention year and the end of the following 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 of the intervention year and end of the following 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 during this period, 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 just prior to the start of the interventions, (2) how much improvement the students would have had in the absence of the interventions, and (3) the amount of the interventions 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 on the tests we administered for this evaluation.[37] (Tables appear at the end of this chapter.[38]) 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 cohort to a high of 93 (32nd percentile) for Word Attack and Passage Comprehension in the fifth-grade cohort.

These estimates confirm that many students in our evaluation are 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). However, based upon teacher assessment and screening and baseline test

[37] As noted above, we present separate estimates for students in the third- and fifth-grade cohorts. At baseline, these students were in the third and fifth grades, respectively. When we administered the seven reading tests for estimating impacts in this report, most of the students were in the fourth and sixth grades.

[38] In addition to seven reading tests, we administered the spelling and calculation subtests from the Woodcock-Johnson III Tests of Achievement. Estimated impacts on scores for these tests are presented in Appendix E.

performance, 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 achieved in the absence of the interventions, as measured by the tests that we administered for this evaluation, we see a mix of positive and negative gains among the control students in the third-grade cohort and mostly positive gains among students in the fifth-grade cohort, as presented in Table IV.1. In the third-grade pooled (ABCD) control group cohort, students typically showed little change, with gains between about –1 and 1 standard score points, but there was one larger positive gain and a larger negative gain. The negative gain on the Group Reading Assessment and Diagnostic Evaluation (GRADE) test suggests that the average student in the study lost ground relative to other students on this reading comprehension test between the start of the intervention year and the end of the following year. 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 the fifth-grade cohort, the gains were generally positive, ranging from 1 to 5 standard score points for the interventions combined. The exceptions for fifth-grade control students were the reading comprehension tests, which showed almost no change (Passage Comprehension) or some negative gain (GRADE).

The positive gains experienced by the control students as measured by some of these tests, indicate that these students' reading ability improved in some dimensions between the beginning of the intervention year and the end of the following 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, since 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 baseline 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 intervention year, each student in the intervention group was supposed to receive approximately 60 minutes of reading instruction per school day. However, as reported in Torgesen et al. (2006) and Chapter III of this report, we found that when the interventions were implemented, students received 54.1 minutes of instruction per day on average (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.

As discussed in Chapter II, students in the intervention and control groups of the third-grade cohort received about the same amount of reading instruction during both the intervention year and the following year, although the intervention students received more small group and less large group instruction during the intervention year. For the fifth-grade cohort, the interventions not only shifted instruction from large to small groups during the intervention year but also increased the total amount of reading instruction received by students in the interventions. This latter increase was also observed during the next school year, although the available data do not reveal why. When we examine the impacts of the interventions on test scores from the seven reading tests that we administered for our second follow-up, we are comparing the effects of one year with the interventions followed by one year without for the intervention group students to the effects of two years without the interventions for the control group students.

The impacts presented in Tables IV.1 to IV.10, which are the ITT impacts, show the effects of students being given the opportunity to receive—from November into May during one academic year—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. As discussed above, TOT impacts that take into account whether students in the intervention group received either any instruction (99 percent of students) or at least 80 hours of instruction (93 percent of students) would not be substantially different from the ITT impacts.

As noted above, preliminary analyses showed substantially different patterns of impacts by grade (see Table IV.1 and Torgesen et al. 2006). Furthermore, more significant impacts—that is, the number of impacts that are different from zero—are found for the third-grade cohort than for the fifth-grade cohort. 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 using tests that we administered for the purposes of this evaluation as well as impacts on reading and math state assessments. The seven measures of reading ability that we administered 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 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 about 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 the 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 STUDENTS IN THE THIRD-GRADE COHORT

Combined, the four interventions improved the phonemic decoding skills of the third-grade cohort one year after the intervention, raising Word Attack scores by approximately 5 standard score points (effect size 0.36)[39] and Phonemic Decoding Efficiency scores by approximately 4 points (effect size 0.26), 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.[40] The impacts of the four interventions combined are not the impacts of implementing the four interventions simultaneously. Rather, a combined impact can be interpreted as the impact of providing a struggling reader in third or fifth grade with the opportunity to receive a substantial amount of instruction in small groups with reasonably well-trained teachers, although as noted elsewhere, the content and instructional focus across the four interventions varied considerably. Such an impact is of greatest relevance to federal and state policymakers who can support broad programmatic approaches to instruction but cannot generally endorse specific products. In contrast, school district and school administrators must select specific products. For that purpose, the impact of being randomly assigned to an individual intervention—as modified or partially implemented for this study— is most relevant.

When assessing the impacts of the four interventions individually, we also found impacts on both of these measures of phonemic decoding ability, with effect sizes of approximately 0.3 to 0.4, corresponding to moving students in the interventions up the distribution of reading ability approximately 8 to 12 percentile points more than they would have gained had they not been in one of the interventions. Only Failure Free Reading had no significant impact on Phonemic Decoding Efficiency test scores.

---

[39] The impacts presented in this report are generally given 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 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. To calculate the effect sizes for impacts on AIMSweb scores, we used the standard deviations from the test administered in the fall for third- and fifth-grade students during the 2000 to 2001, 2001 to 2002, and 2002 to 2003 school years. The standard deviation was 39 for the third-grade cohort and 47 for the fifth-grade cohort. An effect size of 1 means that the intervention increased test scores by 1 standard deviation.

[40] Effect sizes can be converted into the number of percentile points by which the interventions 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. Effect sizes of 0.2, 0.4, 0.6, 0.8, and 1.0 correspond to percentile increases of approximately 5, 12, 19, 26, and 34, respectively. However, because the percentile increase depends on the baseline test scores, the percentile increase may be slightly different for students with higher or lower baseline test scores.

The four interventions combined and the three word-level interventions combined improved reading accuracy and fluency as measured by the Word Identification and Sight Word Efficiency tests, but not as measured by the number of correct words per minute read on the oral reading passages (AIMSweb). Wilson Reading, in particular, had an impact on both tests, improving scores on the Word Identification test by about 4 standard score points (effect size 0.28) and the Sight Word Efficiency test by nearly 3 standard score points (effect size 0.17). Corrective Reading improved Word Identification test scores by about 3 standard score points (effect size 0.17), but had no impact on the Sight Word Efficiency test, while Failure Free Reading had no impact on the Word Identification test but improved scores on the Sight Word Efficiency test by about 2 standard score points (effect size 0.13). Spell Read and Failure Free Reading were the only interventions to improve test scores on the AIMSweb test, by 6 and 8 words read correctly (effect sizes of 0.15 and 0.20), respectively. The impacts on the standardized tests correspond to moving students up the distribution of reading ability by approximately 3 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 2 standard score points on the third-grade cohort's reading comprehension (effect size 0.14) as measured by the Passage Comprehension test, but not as measured by the GRADE test. The improvement in the Passage Comprehension test was largely due to the Failure Free Reading and Wilson Reading interventions, which improved scores by a little more than 4 standard score points (effect size 0.29) and between 3 and 4 points (effect sizes 0.23), respectively, although the latter impact is not statistically significant.

E. IMPACTS FOR STUDENTS IN THE FIFTH-GRADE COHORT

The interventions had fewer impacts for the fifth-grade cohort than for the third-grade cohort (see Table IV.1 for impacts and Table IV.11 for effect sizes). Combined, the four interventions improved the fifth-grade cohort's 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 time of the follow-up testing, students in the control group had an average Word Attack score of approximately 95 (36th percentile), while the average score among students in the interventions was approximately 97 (43rd percentile). The three word-level interventions combined also improved scores on the Word Attack test, with an impact of about 4 points (effect size 0.26), and on the Phonemic Decoding Efficiency test by about 2 points (effect size 0.16). Across the individual interventions, only Wilson Reading had a significant impact on a measure of phonemic decoding, increasing Word Attack test scores by about 8 standard score points (effect size of 0.52).

For the fifth-grade cohort, the four interventions combined and three word-level interventions combined had no positive impacts on the three measures of word reading accuracy and fluency. However, participation in the interventions did lead to lower test scores on the AIMSweb test, with about 4 fewer words read correctly.[41] Only Spell Read had a positive impact on any of the measures of word reading accuracy and fluency, raising Sight Word Efficiency test scores by 3 points (effect size of 0.23).

The four interventions combined did not affect the fifth-grade cohort's reading comprehension skills. Similarly, neither the three word-level interventions combined nor any of the individual interventions improved the fifth-grade cohort's reading comprehension, as measured by either test.

[41] This impact is no longer significant when we adjust for multiple comparisons (see Appendix D).

F. IMPACTS FOR SUBGROUPS OF THE THIRD- AND FIFTH-GRADE COHORTS

Three of the four interventions—Spell Read, Wilson Reading, and Corrective Reading—focus on improving students' word-level reading skills. 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 to Title I 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 zero, as indicated by an asterisk. That is, within a subgroup—for example, students in the third-grade cohort with baseline Word Attack scores below the 30th percentile—an asterisk indicates that the interventions improved reading ability, as measured by a particular test, compared with the control group. The other significance test is whether the impact for the subgroup is different from the overall impact (within grade cohorts), as indicated by a pound sign (#). In the example above, a pound sign would indicate that the impact for students in the third-grade cohort with low baseline Word Attack scores was significantly different from that for all students in the third-grade cohort.[42]

[42] The estimated impacts are model-based estimates, derived from the estimated parameters of the two-level hierarchical linear models specified earlier in this chapter for the third- and fifth-grade cohorts. 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 H. 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 of the third-grade cohort or all of the fifth-grade cohort. 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

(continued)

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. Approximately half of the students in each cohort had relatively low baseline scores on this test. 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 at the same ages, there were no students in this group with average or above average scores in phonemic decoding before the interventions began.

Several of the impacts for students with low baseline Word Attack scores were similar to those for the full sample of students (see Table IV.2). Among students with low Word Attack scores in the third-grade cohort, the four interventions combined and the three word-level interventions combined had positive impacts on both measures of phonemic decoding, as was seen for the entire third-grade cohort. Likewise, the four interventions combined and the three word-level interventions combined improved scores on the measure of reading accuracy (Word Identification) and had no impact on the number of correct words read per minute on the oral reading passages (AIMSweb) or on reading comprehension according to the GRADE measure for all the students in the third-grade cohort and for those 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 fluency test for the third-grade cohort as a whole, they did not improve scores on this test for students with low Word Attack scores. Finally, the interventions combined improved scores on the Passage Comprehension test for all students in the third-grade cohort, but not for students with low Word Attack scores.

Within the fifth-grade cohort, a few more impacts are seen for students with low Word Attack scores than for all students.[43] Along with raising the Word Attack test scores and lowering AIMSweb test scores, the four interventions combined improved scores on the Sight Word Efficiency test and the GRADE test for students with low Word Attack scores. The three word-level interventions improved scores for the entire fifth-grade cohort and for the fifth-grade cohort with low Word Attack scores on the phonemic decoding tests, and negatively impacted AIMSweb test scores, while also improving scores on the GRADE test for students with low Word Attack scores but not all students.

For at least several of the reading measures, impacts might seem smaller for students with low Word Attack scores in the third-grade cohort and larger for such students in the fifth-grade cohort in comparison with all students in those cohorts. However, differences in impacts are significant in only a few instances. Thus, we cannot conclude that low scores on the Word Attack test prior to the interventions made a reliable and consistent difference in the size of impacts obtained. Likewise, relatively high scores on the Word Attack test prior to the interventions do not reliably and consistently affect the size of impacts (see Table IV.3).

(continued)

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 cohort—and not to estimate small impacts precisely.

[43] Some of the impacts, however, are no longer significant when we adjust for multiple comparisons.

2. Students with Relatively Low or High Vocabulary at Baseline

The impacts of the interventions may vary by students' broad vocabulary level. Therefore, we 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).[44] Forty-five percent of students in the third-grade cohort and 49 percent of students in the fifth-grade cohort had relatively low scores on this test. As described in Chapter II, we used this test in screening students for eligibility.

Although fewer impacts are seen with the interventions combined for the third-grade cohort with low Peabody Picture Vocabulary test scores than were seen for the entire third-grade cohort, only the difference in impacts for the AIMSweb test is statistically significant (see Table IV.4). In contrast, the four interventions combined had one more impact on third-grade students who began the year with relatively high Peabody Picture Vocabulary test scores (see Table IV.5), as compared to all of the third-grade cohort. Again, however, only the difference in the AIMSweb impact is statistically significant.

For the fifth-grade cohort, the patterns of impacts were generally similar for students with high and low Peabody Picture Vocabulary test scores. The four interventions combined increased GRADE scores for the students with low Peabody Picture Vocabulary test scores, but not for the students with high scores. However, the difference between the GRADE impacts was not significant.

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). Therefore, we examined impacts within subgroups defined by baseline Word Attack and screening Peabody Picture Vocabulary test scores. Each subgroup is approximately 25 percent of the full sample.

We did find seemingly different patterns of impacts across subgroups defined by these tests (see Tables IV.6 through IV.8), although many differences are not significant because sample sizes are small. The interventions combined improved more test scores for students in the third-grade cohort with relatively high scores on both the Word Attack test and the Peabody Picture Vocabulary test than for the other subgroups. In contrast, the interventions combined had fewer impacts for the fifth-grade cohort students with relatively high scores on both the Word Attack test and the Peabody Picture Vocabulary test than for the subgroups with lower Word Attack scores. The following is a summary of the impacts for the three groups of students of particular interest defined by these two tests:

- ***Students with Low Word Attack and Low Peabody Picture Vocabulary Test Scores.***[45] For the third-grade cohort students in this group, the four interventions combined and the three word-level interventions combined improved scores on the Word Attack test and

[44] The 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.

[45] 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).

negatively impacted scores on the AIMSweb test. This is the only group within the third-grade cohort where the combined interventions had a negative impact on AIMSweb scores, an impact that is significantly different than the results for the entire third-grade cohort. For students in the fifth-grade cohort in this group, the four interventions combined had positive impacts on scores on both of the phonemic decoding tests, the Sight Word Efficiency test, and the Passage Comprehension test, but a negative impact on scores on the AIMSweb test.

- ***Students with Low Word Attack and High Peabody Picture Vocabulary Test Scores.***[46] For students in the third-grade cohort, the four interventions combined had positive impacts on scores on the Word Identification, AIMSweb, and Passage Comprehension tests. The three word-level interventions improved only the AIMSweb test scores. For students in the fifth-grade cohort, the four interventions combined and the three word-level interventions combined had impacts on both measures of phonemic decoding, the Sight Word Efficiency test, and the GRADE test.

- ***Students with High Word Attack and High Peabody Picture Vocabulary Test Scores.***[47] For students in the third-grade cohort, the four interventions combined and the three word-level interventions combined improved scores on both measures of phonemic decoding and two of the three reading accuracy and fluency test scores. The four interventions combined also improved scores on the Passage Comprehension test. For students in the fifth-grade cohort, the four interventions combined and the three word-level interventions combined had impacts only on the Phonemic Decoding Efficiency test scores and the GRADE test scores: the interventions had a positive impact on decoding and a negative impact on comprehension.

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

Because of increased attention to 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).[48] Within the third-grade cohort, larger impacts were seen for the 42 percent of students ineligible for FRPL (with relatively high family income) than for the 58 percent of students eligible for FRPL (with relatively low family income). The four interventions combined had an impact on the two measures of phonemic decoding and the Sight Word Efficiency test for students eligible for FRPL, and had impacts on every test except the GRADE for students ineligible for FRPL. The difference between impacts for the two subgroups was significant for the Phonemic Decoding Efficiency and Passage Comprehension tests.

In contrast to our finding pertaining to the third-grade cohort, within the fifth-grade cohort, the four interventions combined had only a few impacts for either the 55 percent of students eligible for FRPL or the 45 percent of students ineligible for FRPL. No differences between impacts for the subgroups were significant.

---

[46] This group of students had low reading ability (below the 30th percentile) but relatively high vocabulary skills (above the 30th percentile) prior to the interventions.

[47] These students had relatively high reading ability and vocabulary skills (above the 30th percentile on both tests).

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

G. DO THE INTERVENTIONS CLOSE THE READING GAP?

The impact estimates show that, for five of the seven outcomes that measured word-level skills and comprehension, students in the third-grade cohort who were assigned to one of the four interventions had better reading scores one year after the end of the interventions than the control students who received their ordinary instruction. For the fifth-grade cohort, in contrast, impacts of the four interventions combined were found only for Word Attack and AIMSweb.

To assess the extent to which the interventions helped to close the reading gap, we determine how much smaller the gap is for students in the interventions than for students in the control group one year after the interventions ended. 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.21). By the end of the year following the intervention year, students in the control group were still generally between one-half and one standard deviation below the population average.

Reflecting the estimated pattern of impacts, the gaps for students in the interventions were smaller at the end of the year following the intervention year than the gap 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 following the intervention year.[49]

Table IV.21 shows that for students in the control group of the third-grade cohort the gap in passage comprehension skills on the Passage Comprehension subtest of the WRMT-R, for example, is about 9 standard score points at the end of the year following the intervention year (100 - 91). Students in the intervention group had an average standard score that was about 7 points below the population mean (100 - 93). The 2-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 fifth (2/9 = 0.22).[50] Results for the other outcomes show that the largest reduction in the reading gap for the third-grade cohort occurred on the Word Attack test (68 percent reduction). On the tests for other word-level skills, the interventions reduced the gap by about one-fifth or between one-quarter and one-third a full year after the end of the interventions. For the GRADE test, on the other hand, students in both the control group and the treatment group lost ground relative to the national average between the beginning of the intervention year and the end of the following school year. Even with the interventions, the gap between the students participating in this study and average performance was 20 standard score points.

[49] The relative gap reduction due to the intervention was computed as: RGR=[(100-Mean for Control Group at Follow-up)-(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.

[50] As shown in Table IV.21, the gap reduction calculated from means with less rounding is 0.24.

For the fifth-grade cohort, the interventions reduced the gap by 50 percent on Word Attack. For most of the other outcomes, for which impacts were not statistically significant, negligible reductions were observed. One year after the intervention year, the gap for the average intervention student was approximately 3 points for Word Attack, and 9-12 points for the other tests.

H. IMPACTS ON PENNSYLVANIA SYSTEM OF SCHOOL ASSESSMENT TEST SCORES

The tests discussed to this point were administered one year after the end of the intervention year. We also obtained student scores from the PSSA reading and math tests that were taken in late March to early April of the intervention year.

Tables IV.22 through IV.31 present the results for the PSSA tests.[51,52] Most impacts were not statistically significant.[53] Of the impacts that were significant, many were negative, indicating that the students who did not receive the treatment performed better on these tests than the students receiving the treatment.

1. Impacts for the Third-Grade cohort

While the impacts on the reading test scores are mostly negative for the third-grade cohort, it was only the Failure Free Reading intervention that showed a significant negative impact among all students, lowering scores by approximately 51 scaled points (effect size -0.26). Neither the four interventions combined nor the three word-level interventions combined had significant impacts. Considering subgroups we find that the interventions combined decreased scores for students in the third-grade cohort who were eligible for free and reduced-price lunch by 66 scaled points (effect size -0.33).

In contrast to the impacts on PSSA reading test scores, the interventions had mostly positive impacts on math scores for the third-grade cohort, although only the impact of 57 scaled points (effect size 0.27) by Wilson Reading was significant. Among subgroups, the four interventions combined and the three word-level interventions improved scores by 46 scaled points and 51 scaled points (effect sizes 0.22 and 0.24), respectively, for students with relatively low Word Attack scores at baseline and by 132 and 139 scaled points (effect sizes 0.62 and 0.66), respectively, for students with both relatively low baseline Word Attack scores and low screening vocabulary scores.

[51] To estimate these results, we used the same basic model as described in Equations (IV.1) and (IV.2). Because we do not have baseline scores for the PSSA tests, we include the baseline scores from the Passage Comprehension and GRADE tests as covariates when estimating impacts on the PSSA reading scores. We include the baseline scores from the Woodcock Johnson Calculation test as covariates when estimating impacts on the PSSA math scores.

[52] The impacts we present are based on the scaled scores. The population mean for third graders in 2004 for was 1303 for reading and 1349 for math. The standard deviations were 198 and 213, respectively. The population mean for fifth graders in 2004 was 1370 for reading and 1380 for math. The standard deviations were 242 and 239, respectively.

[53] The standard errors and test statistics are reported in Appendix H.

2. Impacts for the Fifth-Grade cohort

Similar to the results for the third-grade cohort, most impacts on reading scores for the fifth-grade cohort are negative. The four interventions combined lowered scores by 27 scaled points (effect size -0.11), although Failure Free Reading was, again the only individual intervention to show a significant impact. Among fifth graders with relatively high baseline Word Attack or relatively high screening vocabulary scores or both relatively high baseline Word Attack and screening vocabulary scores, the four interventions combined lowered reading scores. In addition, the four interventions combined lowered scores for students eligible for free or reduced-price lunch by 74 scaled points (effect size -0.31).

The four interventions combined and the three word-level interventions combined lowered math scores for students in the fifth-grade cohort by 29 and 34 scaled points (effect sizes -0.12 and -0.14), respectively, primarily due to the negative impacts of Wilson Reading and Corrective Reading. The four interventions combined also lowered scores for students with relatively high baseline Word Attack scores and for students with low screening vocabulary scores.

I. SUMMARY OF KEY FINDINGS

Although many estimates have been provided in this report, our key findings are as follows:[54]

1. ***The interventions improved some reading skills.*** For students in the third-grade cohort, the four interventions combined had impacts on phonemic decoding, word reading accuracy and fluency, and reading comprehension, although impacts were not detected for all measures of accuracy and fluency or comprehension. For students in the fifth-grade cohort, the four interventions combined improved phonemic decoding on one measure, but led to a small reduction in oral reading fluency. The three word-level interventions combined had similar impacts to those for all four interventions combined, although they did not have an impact on either measure of comprehension for students in the third-grade cohort, and they did have impacts on both measures of phonemic decoding for students in the fifth-grade cohort. For students in the third-grade cohort, Failure Free Reading (the only word level plus comprehension program) had impacts on one measure of phonemic decoding, two of the three measures of word reading accuracy and fluency, and one measure of comprehension. However, this intervention did not have any impacts for students in the fifth-grade cohort.

2. ***The interventions did not improve PSSA scores.*** For students in the third-grade cohort, we did not detect significant impacts of the four interventions combined on reading and mathematics test scores from the Pennsylvania System of School Assessment that were taken in late March and early April of the intervention year. For students in the fifth-grade cohort, the four interventions combined lowered the reading and mathematics scores.

3. ***Younger students benefited more.*** The interventions generally helped students in the third-grade cohort more than students in the fifth-grade cohort. However, the interventions did not consistently benefit any one subgroup more than another.

[54] To identify and test possible explanations for the findings presented in this report are beyond the purpose and design of the study.

4. ***The interventions narrowed some reading gaps.*** The four interventions combined generally narrowed the reading gap for students in the intervention groups compared with students in the control group for the third-grade cohort. Being in one of the interventions reduced the reading gap in Word Attack skills by about two-thirds for students in the third-grade cohort. On other word-level tests and a measure of reading comprehension, the interventions reduced the gap for students in the third-grade cohort by about one-sixth to one-third. For students in the fifth-grade cohort, the interventions reduced the gap in Word Attack skills by one-half.

The key findings presented in this report for the seven tests administered for this study one year after the interventions ended are similar to the findings from the end of the intervention year. In our earlier report (Torgesen et al. 2006) we found that the four interventions combined and the three word-level interventions had impacts for students in the third-grade cohort on phonemic decoding, word reading accuracy and fluency, and reading comprehension. We found fewer significant impacts for students in the fifth-grade cohort than for students in third-grade cohort. Also, for the four interventions combined, the reading gaps for students in the intervention group were generally smaller than the gaps for students in the control group.

Figure IV.2

Gap Reduction for Third-Grade Cohort: Word Attack

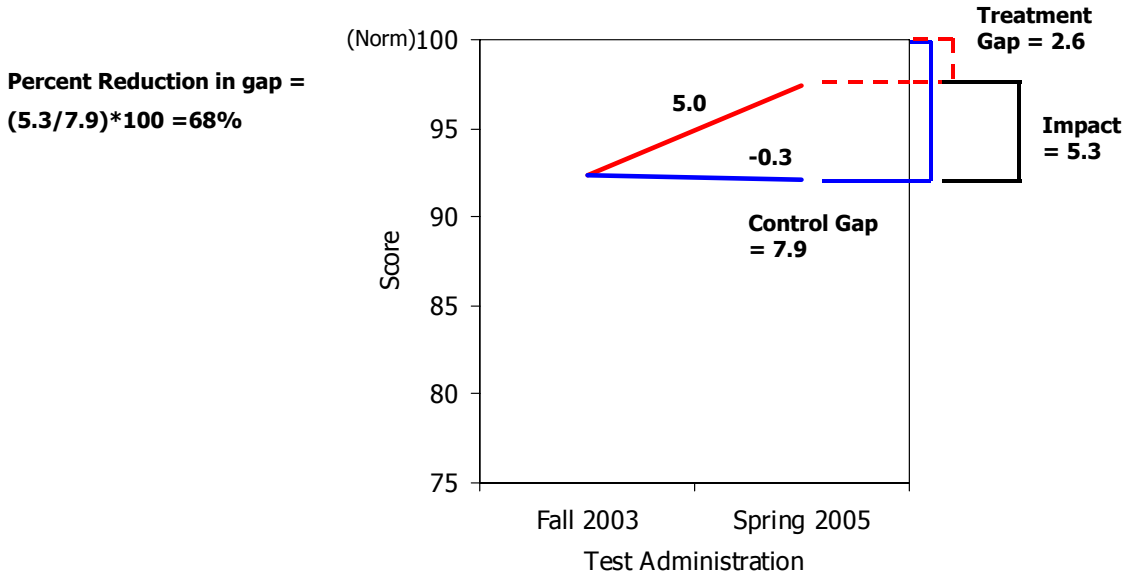


Figure IV.3

Gap Reduction for Third-Grade Cohort: TOWRE PDE

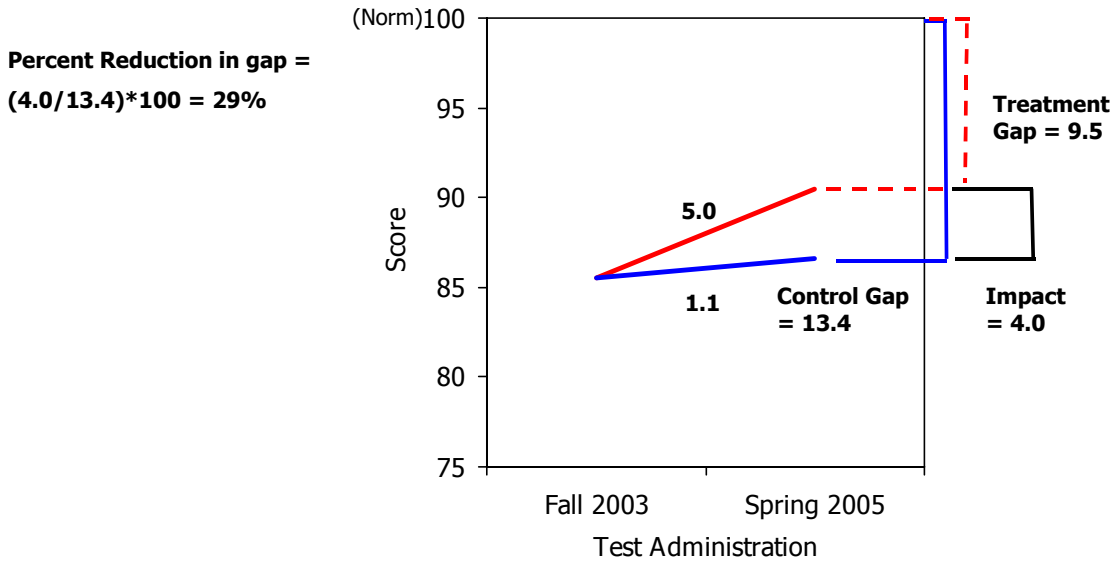


Figure IV.4

Gap Reduction for Third-Grade Cohort: Word Identification

Percent Reduction in gap =
(2.3/11.4)*100 = 20%

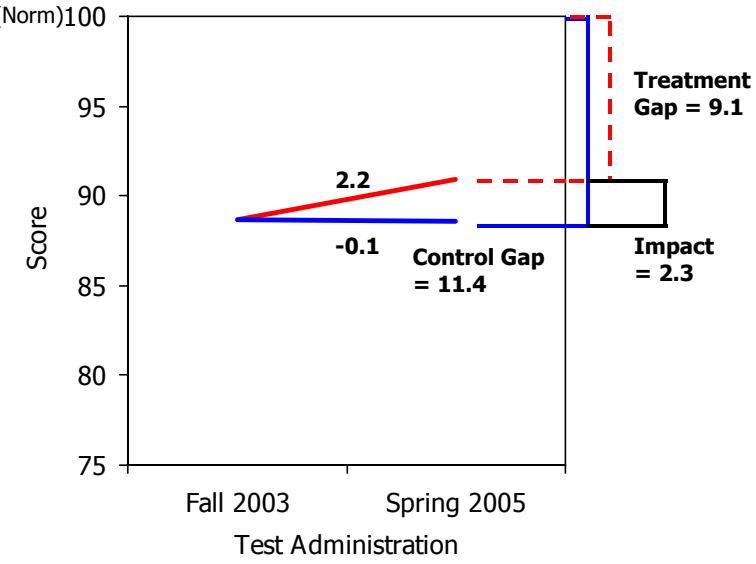


Figure IV.5

Gap Reduction for Third-Grade Cohort: TOWRE SWE

Percent Reduction in gap =
(1.7/10.0)*100 = 17%

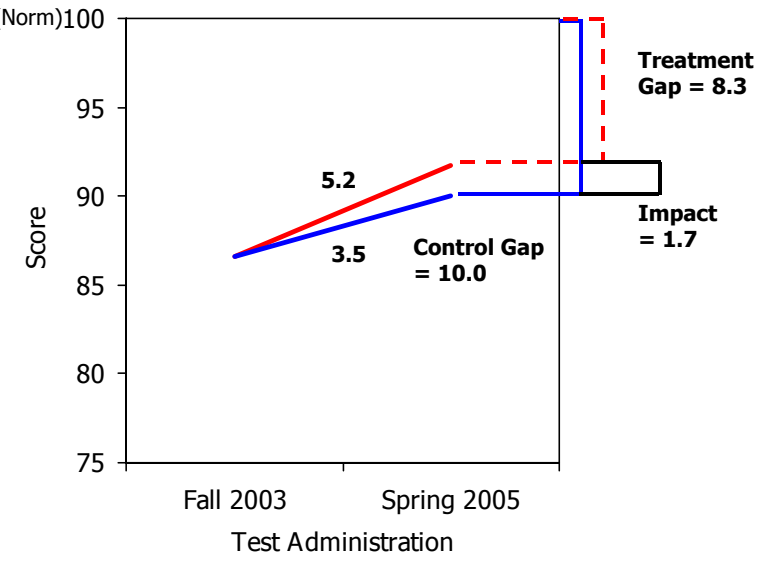


Figure IV.6

Gap Reduction for Third-Grade Cohort: Passage Comprehension

Percent Reduction in gap =
(2.1/8.7)*100 =24%

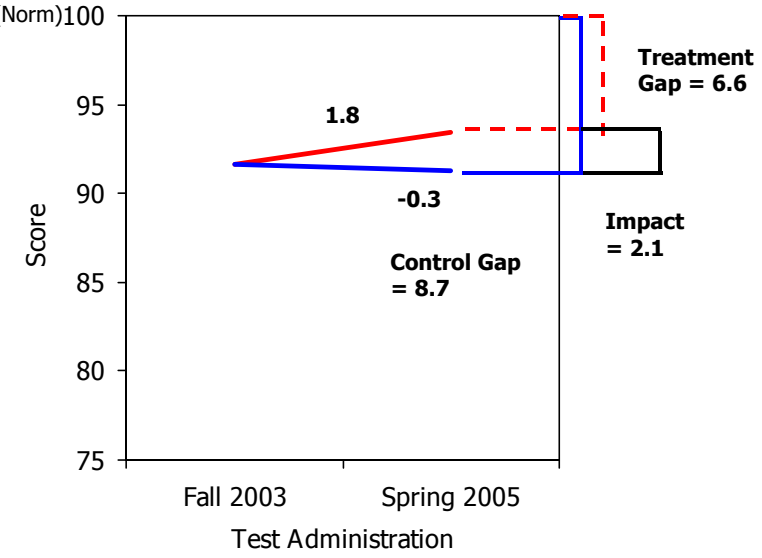


Figure IV.7

Gap Reduction for Third-Grade Cohort: GRADE

Percent Reduction in gap =
(1.0/21.4)*100 = 5%

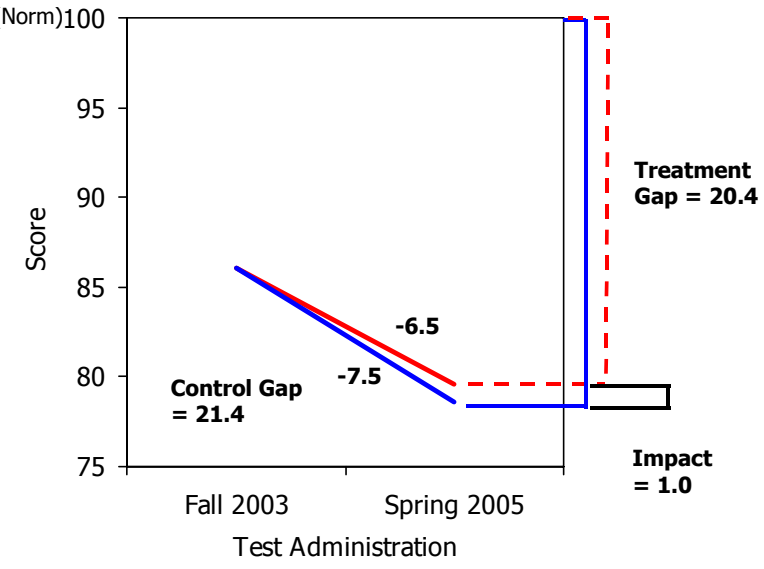


Figure IV.8

Gap Reduction for Fifth-Grade Cohort: Word Attack

Percent Reduction in gap =
(2.7/5.3)*100 = 50%

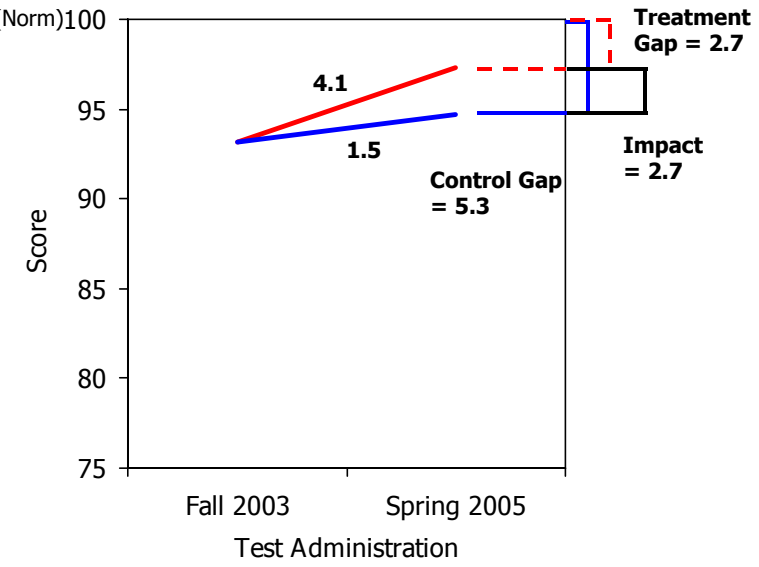


Figure IV.9

Gap Reduction for Fifth-Grade Cohort: TOWRE PDE

Percent Reduction in gap =
(1.7/13.7)*100 = 13%

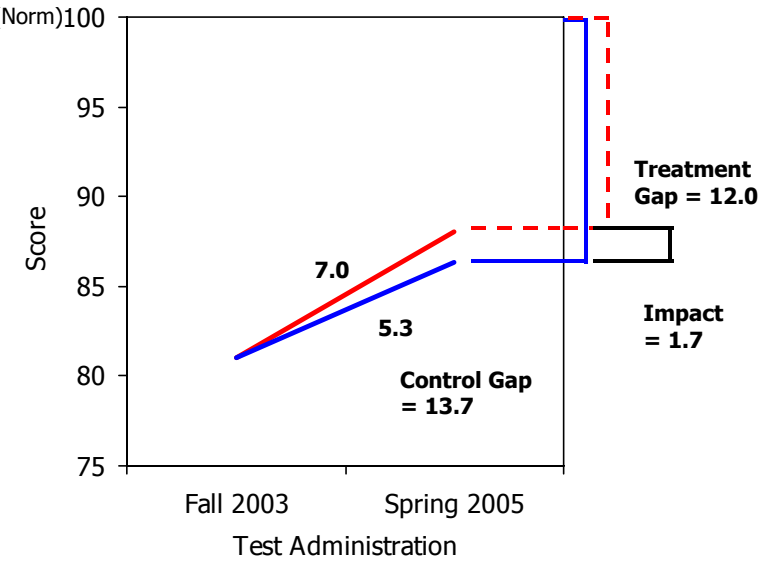


Figure IV.10

Gap Reduction for Fifth-Grade Cohort: Word Identification

Percent Reduction in gap =
(-0.6/8.3)*100 = -7%

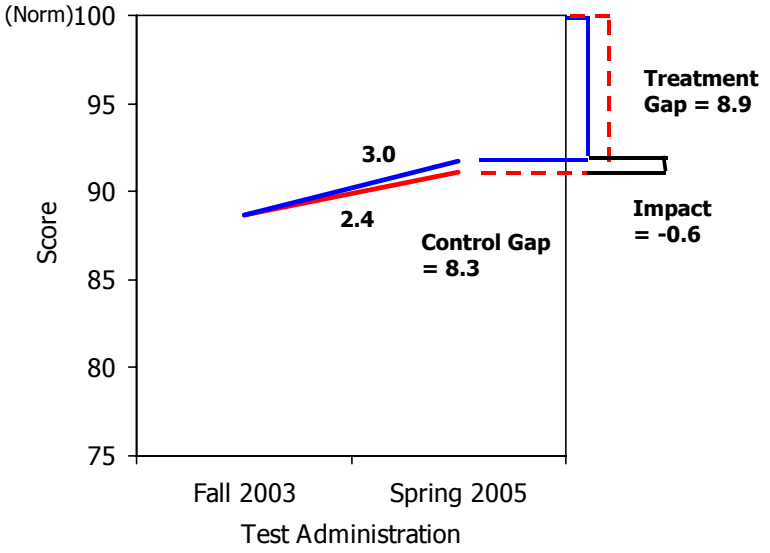


Figure IV.11

Gap Reduction for Fifth-Grade Cohort: TOWRE SWE

Percent Reduction in gap =
(1.4/12.8)*100 = 11%

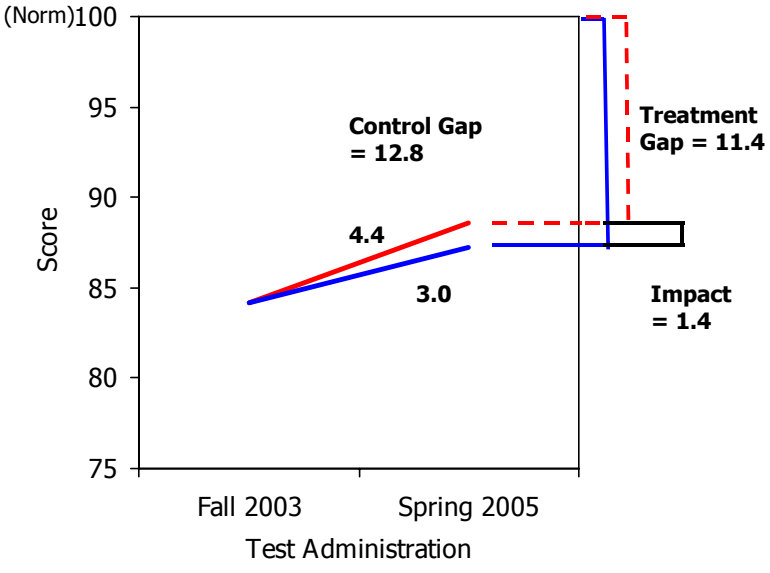


Figure IV.12

Gap Reduction for Fifth-Grade Cohort: Passage Comprehension

Percent Reduction in gap =
(-1.1/7.8)*100 = -15%

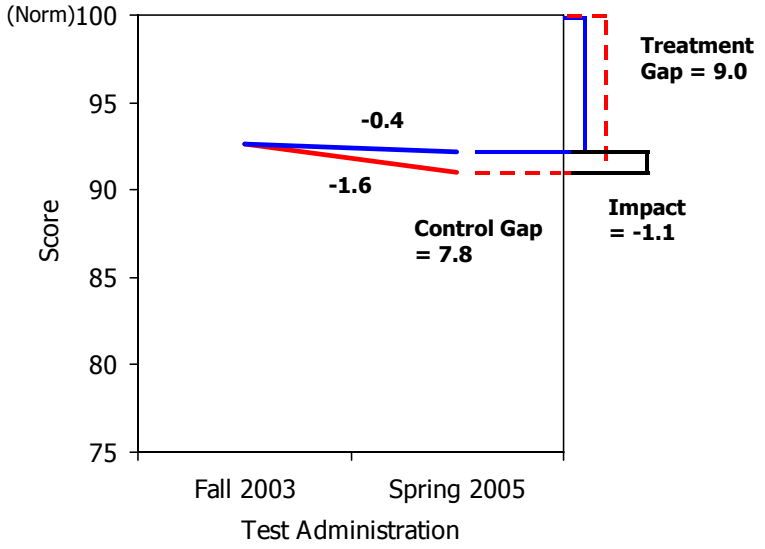


Figure IV.13

Gap Reduction for Fifth-Grade Cohort: GRADE

Percent Reduction in gap =
(0.7/12.3)*100 = 6%

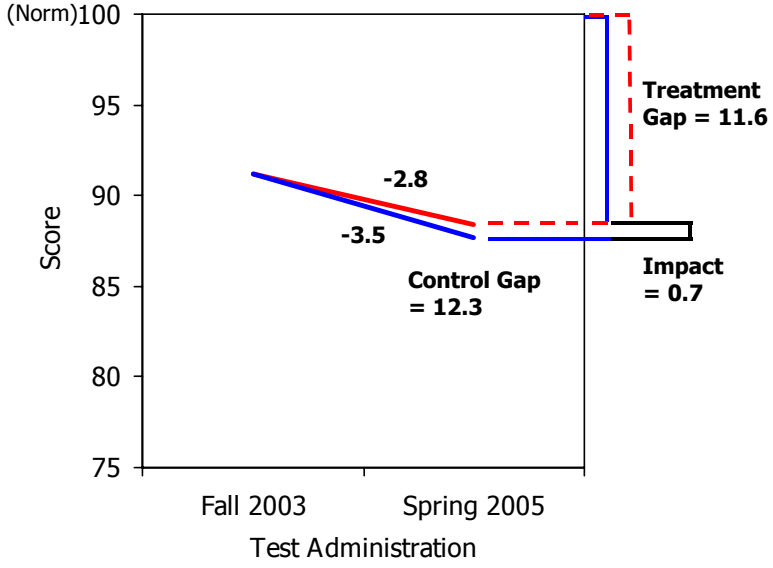


Table IV.1

Impacts for 3rd and 5th Grade Cohorts
One Year After the Intervention Year

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 3 Cohort	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	92.4	-0.3	5.3 *	0.3	5.5 *	-2.2	4.9 *	-0.3	5.4 *	0.5	5.8 *	0.7	5.2 *
TOWRE PDE	85.5	1.1	4.0 *	1.0	4.9 *	1.3	1.3	3.4	4.9 *	1.8	4.1 *	-2.3	5.5 *
Word Identification	88.7	-0.1	2.3 *	0.0	2.5 *	-0.4	1.8	1.8	0.7	-2.2	4.1 *	0.6	2.6 *
TOWRE SWE	86.6	3.5	1.7 *	3.7	1.6 *	2.7	2.0 *	4.0	0.9	2.8	2.6 *	4.3	1.4
AIMSweb	40.9	33.7	5.3	33.4	4.5	34.4	7.9 *	30.1	6.0 *	31.9	3.6	38.3	3.9
Passage Comprehension	91.6	-0.3	2.1 *	0.3	1.3	-2.1	4.4 *	1.2	0.1	-2.5	3.5	2.3	0.3
GRADE	86.1	-7.5	1.0	-6.9	0.4	-9.3	2.8	-10.0	2.1	-10.4	0.1	-0.1	-1.1
Sample Size	329		329		240		89		91		70		79

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 5 Cohort	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	93.2	1.5	2.7 * #	1.8	3.8 *	0.4	-0.8 #	0.0	3.5	2.3	7.8 *	3.1	0.2 #
TOWRE PDE	81.0	5.3	1.7	5.2	2.4 *	5.4	-0.3	4.9	3.2	4.2	2.6	6.6	1.4
Word Identification	88.7	3.0	-0.6 #	3.4	-0.6 #	1.7	-0.6 #	1.5	0.1	4.3	0.0 #	4.3	-1.9 #
TOWRE SWE	84.2	3.0	1.4	3.1	1.4	3.0	1.5	1.5	3.4 *	2.7	1.1	5.0	-0.4
AIMSweb	77.0	30.9	-3.9 * #	30.7	-3.9 * #	31.6	-4.1 #	26.5	-3.3 #	29.7	-3.0	35.9	-5.3
Passage Comprehension	92.6	-0.4	-1.1 #	-0.8	-0.7	0.8	-2.5 #	-2.8	-0.9	-1.6	0.9	1.9	-2.1
GRADE	91.2	-3.5	0.7	-4.7	1.2	0.2	-0.9	-4.8	-1.1	-8.9	4.7	-0.4	0.0
Sample Size	400		400		272		128		100		88		84

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

* Impact is statistically significant at the 0.05 level.
# Impact is statistically different from the 3rd grade cohort impact at the 0.05 level.

Table IV.2

Impacts for 3rd and 5th Grade Cohorts with Low Baseline Word Attack Scores
One Year After the Intervention Year

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 3 Cohort	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	4.2 *	3.2	4.7 *	1.7	2.7	3.2	5.1 *	5.3	2.7	1.1	6.3
TOWRE PDE	82.0	0.6	3.9 *	0.7	4.7 *	0.6	1.4	2.6	5.4 *	4.6	0.0 #	-5.2	8.6
Word Identification	85.3	0.3	1.7 *	0.5	1.6 *	-0.5	1.9	2.2	0.3	-1.1	3.0 *	0.6	1.6
TOWRE SWE	82.9	5.0	0.9 #	5.2	0.6 #	4.6	2.0	4.7	0.8	4.0	2.5	6.8	-1.6 #
AIMSweb	32.2	36.2	2.4	37.1	0.0	33.6	9.7 *	28.5	6.4 *	35.8	-3.2	47.0	-3.1
Passage Comprehension	86.7	2.3	1.5	2.5	1.4	1.6	2.1	5.1	-2.1	-0.4	3.9	2.8	2.2
GRADE	83.0	-8.7	0.9	-8.0	1.1	-10.6	0.4	-8.1	-2.6 #	-16.9	8.2 #	0.9	-2.3
Sample Size	170		170		114		56		47		30		37

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 5 Cohort	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	84.7	3.2	4.3 * #	2.9	6.0 * #	4.1	-1.0	-0.6	6.3 * #	5.7	7.8 *	3.5	4.0 *
TOWRE PDE	75.5	5.4	1.9	5.0	3.1 *	6.6	-1.7	5.1	3.5	4.6	2.9	5.2	3.0
Word Identification	84.0	1.4	0.0	1.6	-0.2	0.7	0.3	0.7	0.3	3.1	-1.4	1.0	0.6
TOWRE SWE	81.2	2.5	2.2 *	2.5	2.1	2.6	2.2 *	1.6	3.8 *	2.4	1.4	3.6	1.1
AIMSweb	67.0	31.1	-5.8 *	30.6	-5.5 *	32.5	-6.5	26.3	-7.9	32.0	-6.9 *	33.6	-1.7
Passage Comprehension	89.2	-1.2	0.7	-2.2	1.4	1.8	-1.5	-2.0	0.3	-3.7	3.3	-0.8	0.4
GRADE	88.0	-4.1	3.7 * #	-5.6	4.7 * #	0.5	0.9	-3.0	-1.2	-11.9	10.6 * #	-1.9	4.7 #
Sample Size	195		195		137		58		59		42		36

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

* 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 Grade Cohorts with High Baseline Word Attack Scores
One Year After the Intervention Year

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 3 Cohort	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	100.9	-3.5	6.5 *	-2.2	5.6 *	-7.4	9.3 *	-2.5	3.7	-1.3	4.8 *	-2.9	8.3 *
TOWRE PDE	89.2	2.2	3.6 *	1.8	4.8 *	3.3	0.1	3.2	5.4 *	0.6	6.1 * #	1.5	2.8 *
Word Identification	92.2	0.4	3.0 *	0.6	3.0 *	-0.1	3.0 *	2.0	0.9	-2.1	4.4 *	1.8	3.8
TOWRE SWE	90.5	1.6	2.8 * #	1.7	2.7 * #	1.3	3.1 *	3.0	0.9	0.9	3.1 *	1.2	4.2 * #
AIMSweb	50.2	33.1	4.9 *	31.2	6.3 *	38.8	0.7	32.0	3.9	27.4	10.8 *	34.1	4.2
Passage Comprehension	96.8	-2.5	2.0 *	-1.7	0.6	-5.0	6.0 *	-2.1	1.3	-1.7	-0.8	-1.2	1.4
GRADE	89.4	-4.9	-0.5	-4.6	-2.4	-5.8	5.2 *	-10.9	4.9 #	-3.6	-10.5 * #	0.7	-1.4
Sample Size	159		159		126		33		44		40		42

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 5 Cohort	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	0.1	0.8 #	1.3	1.2 #	-3.5	-0.2	2.1	-1.8 #	1.1	6.0 *	0.8	-0.8
TOWRE PDE	86.4	5.4	1.2	6.0	1.1	3.7	1.5	5.5	2.0	5.3	1.0	7.3	0.2
Word Identification	93.2	4.7	-1.8 *	5.1	-1.7	3.7	-2.2	4.4	-2.2	4.5	0.6	6.3	-3.3
TOWRE SWE	87.1	3.1	1.1	3.1	1.2	3.1	0.8	1.5	3.3 *	2.8	0.6	5.0	-0.4
AIMSweb	86.8	30.0	-1.5	29.6	-1.2	31.0	-2.2	25.7	3.1	26.3	0.5	37.0	-7.3
Passage Comprehension	96.0	-0.8	-2.1 *	-0.8	-1.9	-0.9	-2.8	-3.7	-2.5	-1.0	0.1	2.3	-3.2 *
GRADE	94.3	-3.4	-1.7 #	-4.7	-1.1 #	0.6	-3.4 *	-7.0	0.2	-6.2	-1.0 #	-0.9	-2.3 * #
Sample Size	205		205		135		70		41		46		48

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

* 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 Grade Cohorts with Low Screening Peabody Picture Vocabulary Test Scores
One Year After the Intervention Year

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 3 Cohort	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	92.0	0.0	4.0 *	1.9	3.0 *	-5.8	6.9 *	6.0	-0.7 #	-0.4	6.2 *	0.2	3.6 *
TOWRE PDE	85.3	1.6	3.6 *	2.4	3.9 *	-0.6	2.8	6.9	2.1	2.8	3.6 *	-2.5	6.1 * #
Word Identification	87.9	0.1	1.7	0.4	1.7	-0.9	1.7	2.8	-0.7	-1.7	3.9 *	0.1	2.0
TOWRE SWE	86.0	3.6	1.0	3.8	1.0	3.0	0.8	3.5	0.8	3.3	2.1	4.6	0.2
AIMSweb	39.0	38.0	-2.1 #	40.1	-6.3 #	31.7	10.4	36.0	-1.9	47.4	-16.2 #	36.8	-0.9
Passage Comprehension	89.9	-0.1	0.6	0.6	0.1	-2.4	2.1	0.3	0.5	1.4	-1.2	0.2	1.2
GRADE	83.4	-8.7	1.1	-9.2	1.9	-7.1	-1.5 #	-14.6	5.9	-11.2	3.3	-1.9	-3.3
Sample Size	147		147		110		37		37		42		31

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 5 Cohort	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	91.4	1.9	2.4 *	2.0	3.9 *	1.7	-1.9	2.4	1.2	4.4	7.2 *	-0.8	3.2
TOWRE PDE	79.9	5.3	2.2 *	5.5	2.6 *	4.7	1.2	6.6	1.0 #	5.7	2.1	4.2	4.7
Word Identification	86.7	1.6	-0.2	2.3	-0.4	-0.7	0.5	1.7	-1.2 #	2.8	1.3	2.4	-1.5
TOWRE SWE	83.5	1.9	2.4 *	1.4	3.0 *	3.3	0.7	1.4	2.7 *	1.0	2.1	1.8	4.0 * #
AIMSweb	72.3	34.7	-6.6	32.6	-5.0	40.9	-11.4	28.1	-6.7	28.0	0.0	41.9	-8.3
Passage Comprehension	89.4	-2.5	1.6 #	-2.8	2.2 * #	-1.5	-0.4	-3.0	1.3	-3.0	3.3 * #	-2.5	2.0
GRADE	87.0	-3.9	3.5 *	-4.1	3.3 *	-3.4	4.0 * #	-3.5	-0.3	-8.1	5.1 *	-0.7	5.2 *
Sample Size	195		195		140		55		53		52		35

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

* 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 Grade Cohorts with High Screening Peabody Picture Vocabulary Test Scores
One Year After the Intervention Year

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 3 Cohort	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	92.8	0.9	4.4 *	0.9	5.2 *	0.8	1.9	-2.5	7.7 * #	-0.4	5.7	5.6	2.1 *
TOWRE PDE	85.7	2.4	2.8 *	2.9	3.0 *	1.1	1.9	1.8	6.6 *	3.1	2.6	3.8	-0.1 #
Word Identification	89.3	0.1	2.3 *	0.1	2.5 *	0.2	1.7 *	1.2	1.5	-2.6	4.2 *	1.7	1.9
TOWRE SWE	87.0	3.3	2.0 *	3.5	1.8	2.5	2.7 *	3.9	1.1	2.7	2.4	4.0	1.8
AIMSweb	42.5	30.1	9.9 * #	28.4	10.9 * #	35.0	6.8 *	26.9	9.6 *	18.8	17.9 * #	39.6	5.1
Passage Comprehension	92.9	-0.8	3.0 *	-0.4	2.2	-1.9	5.4 *	1.7	-0.6	-5.8	6.5	2.9	0.7
GRADE	88.3	-8.1	3.2	-7.3	2.1	-10.7	6.3 #	-11.1	4.5	-8.1	-2.6	-2.5	4.5
Sample Size	182		182		130		52		54		28		48

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 5 Cohort	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	94.7	1.1	2.2 *	1.6	3.1 *	-0.4	-0.6	-2.2	5.4 *	2.5	5.1	4.4	-1.2
TOWRE PDE	82.0	4.5	2.4 *	4.0	3.7 *	5.9	-1.5	2.9	6.7 * #	2.0	3.4	7.2	1.1
Word Identification	90.3	3.5	-0.6	3.7	-0.6	2.8	-0.6	0.7	1.8 #	5.6	-2.1	4.8	-1.6
TOWRE SWE	84.7	3.5	1.0	3.7	0.9	2.8	1.6	1.7	3.8 *	3.0	1.5	6.3	-2.6 #
AIMSweb	80.9	31.3	-5.9 *	31.5	-6.3 *	30.6	-4.5	26.9	-3.0	31.4	-6.7	36.2	-9.2 *
Passage Comprehension	95.3	1.1	-2.3 #	0.7	-1.9 #	2.2	-3.5	-1.6	-2.3	1.3	-1.7 #	2.5	-1.6
GRADE	94.7	-3.9	-0.3	-5.6	0.5	1.4	-2.9 #	-6.1	-1.3	-8.6	4.9	-2.3	-1.9
Sample Size	205		205		132		73		47		36		49

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

* 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 Grade Cohorts with Low Baseline Word Attack and Low Screening Peabody Picture Vocabulary Test Scores
One Year After the Intervention Year

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 3 Cohort	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	85.3	1.6	3.7 *	2.1	4.5 *	-0.1	1.0	3.6	4.2	3.2	3.5	-0.5	5.9 *
TOWRE PDE	82.0	3.0	1.3	4.1	1.1	-0.2	1.8	11.5	-3.7 #	-0.3	3.7	1.1	3.3
Word Identification	85.1	0.3	0.7	0.9	0.4	-1.6	1.4	3.6	-1.4	-0.3	1.3	-0.5	1.3
TOWRE SWE	82.3	5.5	-0.6 #	5.8	-0.6	4.4	-0.5	4.4	2.8	8.3	-4.2	4.8	-0.3
AIMSweb	31.5	40.9	-6.9 * #	42.8	-10.7 * #	35.0	4.8	36.8	-3.1	60.0	-35.3 * #	31.7	6.1
Passage Comprehension	85.4	2.1	0.6	3.0	0.0	-0.6	2.6	4.4	-1.9	2.0	1.1	2.6	0.7
GRADE	81.4	-8.1	-0.7	-9.4	1.4	-4.3	-6.8 #	-13.8	4.1	-9.7	4.2	-4.8	-4.2
Sample Size	81		81		55		26		24		14		17

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 5 Cohort	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	83.4	4.7	2.7 *	4.3	4.2 *	5.9	-1.9	0.6	4.3 *	10.1	3.4 #	2.1	4.9 * #
TOWRE PDE	74.6	5.3	2.7 *	5.5	3.2 *	4.5	1.1	6.9	0.9	5.7	2.9	3.9	5.8
Word Identification	82.6	0.9	0.7	1.5	0.4	-1.0	1.4	4.4	-1.4	-0.2	1.3	0.1	1.3
TOWRE SWE	80.0	2.2	3.0 *	1.7	3.3	3.9	2.1	1.8	3.4 *	1.7	2.2	1.6	4.4
AIMSweb	63.1	34.3	-10.8 *	34.1	-12.2 *	34.7	-6.6	25.5	-11.2	31.5	-7.6	45.4	-17.7
Passage Comprehension	86.1	-2.6	3.9 * #	-3.3	5.1 * #	-0.5	0.1	-5.3	5.7 * #	-1.4	2.9	-3.3	6.7 * #
GRADE	84.3	-4.4	2.7	-4.1	1.8	-5.2	5.7 #	-3.6	-2.6	-8.9	4.5	0.1	3.3
Sample Size	107		107		77		30		34		25		18

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

* 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 Grade Cohorts With Low Baseline Word Attack and High Peabody Picture Vocabulary Test Scores
One Year After the Intervention Year

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 3 Cohort	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	83.9	7.1	0.1 #	8.0	0.1	4.1	-0.2	5.7	3.0	4.3	3.5	14.2	-6.1 * #
TOWRE PDE	82.1	2.7	1.8	3.3	1.9 #	0.9	1.5	2.9	5.2 *	2.4	3.2	4.7	-2.7 #
Word Identification	85.5	0.5	2.1 *	0.4	2.3	0.7	1.4	1.4	2.2	-3.5	5.7 *	3.2	-0.9
TOWRE SWE	83.3	4.3	1.3	4.3	1.2	4.4	1.5	4.0	0.1	0.9	5.6	7.9	-1.9
AIMSweb	32.9	31.1	9.5 *	29.9	10.4 *	34.8	6.8 *	27.1	8.6	17.5	18.6	45.1	4.1
Passage Comprehension	87.9	1.4	3.6 *	1.4	3.5	1.3	3.9 *	5.0	-2.0	-5.1	9.4	4.3	3.2
GRADE	84.5	-7.4	2.7	-5.8	0.9	-12.3	7.9 * #	-5.3	-3.7	-14.7	4.8	2.7	1.6
Sample Size	89		89		59		30		23		16		20

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 5 Cohort	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	86.0	2.5	4.6 *	2.3	6.5 * #	3.2	-1.2	-0.1	7.2 *	1.9	10.7 *	5.1	1.7
TOWRE PDE	76.5	3.6	3.5 *	2.5	5.5 * #	6.8	-2.5	3.1	7.3 * #	0.3	6.0 *	4.2	3.1 *
Word Identification	85.4	2.1	-0.3	1.9	0.2	2.4	-1.8 * #	0.4	2.2 #	2.2	-0.1	3.3	-1.4
TOWRE SWE	82.3	1.3	2.6 *	1.4	2.8 *	0.9	2.1	0.6	4.4 *	0.5	3.3	3.0	0.7
AIMSweb	71.0	33.1	-7.5	32.4	-5.9	35.3	-12.3	26.5	-2.1	38.8	-14.6 * #	31.9	-1.2
Passage Comprehension	92.4	0.0	-0.6	-1.0	-0.5	2.9	-0.9	0.7	-3.9	-3.5	3.3	-0.2	-1.0
GRADE	91.8	-5.2	6.5 * #	-8.4	9.1 * #	4.2	-1.3	-5.3	5.0	-15.1	14.9 * #	-4.8	7.4
Sample Size	88		88		60		28		25		17		18

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

* 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 Grade Cohorts with High Baseline Word Attack and High Screening Peabody Picture Vocabulary Test Scores
One Year After the Intervention Year

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 3 Cohort	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	101.3	-3.1	5.7 *	-3.3	6.6 *	-2.8	3.1 *	-9.5	10.3 * #	-1.1	4.4 *	0.8	5.2 *
TOWRE PDE	89.1	2.5	2.5 *	2.1	3.7 *	3.7	-1.2	2.6	5.1 *	2.7	4.0 *	1.0	2.1
Word Identification	92.8	1.0	1.6 *	0.5	2.4 *	2.3	-0.7	2.0	1.0	-1.1	1.0	0.7	5.1 *
TOWRE SWE	90.5	2.4	2.5 *	2.5	2.6 *	2.0	2.3 *	4.9	0.1	3.0	2.6 *	-0.2	5.1 * #
AIMSweb	51.7	32.8	5.2	30.7	6.9	39.3	0.0	29.4	4.3	24.3	15.9 *	38.3	0.4
Passage Comprehension	97.7	-3.0	2.5 *	-2.0	0.8	-5.9	7.5 *	-1.8	0.2	-3.5	0.6	-0.9	1.6
GRADE	91.9	-6.8	1.0	-7.5	1.0	-4.8	1.1	-13.3	6.5	-7.5	-4.7	-1.7	1.3
Sample Size	93		93		71		22		31		12		28

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 5 Cohort	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	101.8	0.2	1.2	1.6	1.7	-3.9	-0.1	-0.9	2.1	3.0	4.3	2.7	-1.4
TOWRE PDE	86.4	5.0	2.3 *	5.5	2.7 *	3.6	1.1	5.0	3.7	3.8	3.1	7.6	1.3
Word Identification	94.3	4.6	-1.2	5.0	-1.4	3.5	-0.5	4.1	-2.6 *	6.4	-0.2	4.4	-1.5
TOWRE SWE	86.7	4.1	1.5	4.5	1.4	3.0	1.9	4.2	1.2	3.6	4.5	5.7	-1.6
AIMSweb	89.1	29.1	-2.6	29.5	-3.2	28.2	-0.7	23.6	1.0	24.0	4.0	40.8	-14.5 * #
Passage Comprehension	97.7	-1.1	-0.9	-1.1	-0.3	-1.1	-2.8	-5.1	-0.3	0.0	0.5	1.9	-1.0
GRADE	97.1	-3.8	-4.3 * #	-4.8	-4.4 * #	-1.0	-4.0 *	-7.5	-5.4	-4.7	-2.6	-2.1	-5.4 * #
Sample Size	117		117		72		45		22		19		31

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

* 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 Grade Cohorts Eligible for Free or Reduced-Price School Lunch
One Year After the Intervention Year

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 3 Cohort	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	2.0	3.7 *	3.3	2.7 *	-2.0	6.7 *	4.4	1.5	2.5	4.2 *	2.9	2.3 *
TOWRE PDE	85.1	3.3	2.8 * #	3.9	3.1 * #	1.3	2.0	4.5	5.4 *	3.8	3.2 * #	3.5	0.7 #
Word Identification	87.9	0.6	3.2	0.9	3.4	-0.1	2.5	4.1	-0.3	-1.9	5.6 *	0.5	5.1 *
TOWRE SWE	85.7	3.1	2.1 *	3.6	1.5	1.7	3.6 *	2.7	1.6	3.3	2.8 *	4.9	0.1 *
AIMSweb	38.7	36.6	2.8	36.0	1.7	38.5	5.9	32.6	3.6	34.1	0.3	41.4	1.2
Passage Comprehension	90.2	0.9	-0.1 #	1.0	-0.1	0.6	-0.2 #	1.7	-0.4	0.1	0.2	1.1	0.0
GRADE	84.4	-7.0	1.0	-5.8	-0.5	-10.6	5.6	-8.4	0.1	-10.3	1.5	1.5	-3.1
Sample Size	190		190		144		46		53		46		45

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 5 Cohort	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	2.2	2.9 *	3.0	3.8 *	0.0	0.0	1.5	3.5	5.2	6.1 *	2.3	1.8
TOWRE PDE	80.1	5.8	1.2	6.1	1.6	5.0	0.0	6.9	0.6	6.9	0.2	4.5	3.9 *
Word Identification	87.7	2.0	-0.1	2.4	0.1	0.6	-0.9	1.0	0.6	4.0	0.4	2.3	-0.7
TOWRE SWE	83.0	2.0	2.6 *	2.0	3.0 *	2.2	1.3	1.4	4.0 *	1.9	1.5	2.7	3.6 * #
AIMSweb	72.4	25.9	1.7	24.6	4.6 #	29.7	-7.0	20.9	3.7 #	23.7	3.9	29.2	6.1 #
Passage Comprehension	90.4	-0.5	-1.2	-1.1	-0.5	1.1	-3.2	-1.1	-2.1	-1.9	1.1	-0.3	-0.5
GRADE	88.0	-0.9	-3.3 *	-2.3	-2.9 *	3.3	-4.5 *	1.6	-9.4 * #	-5.6	0.4	-2.8	0.3
Sample Size	220		220		150		70		50		56		44

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

* 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 Grade Cohorts Not Eligible for Free or Reduced-Price School Lunch
One Year After the Intervention Year

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 3 Cohort	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	93.2	-2.1	6.1 *	-2.3	7.7 *	-1.8	1.0	-3.0	6.6 *	-3.6	9.7	-0.2	7.0 *
TOWRE PDE	86.0	-2.8	7.8 * #	-4.0	10.3 * #	1.0	0.4	2.2	5.3 *	-10.2	15.8 * #	-4.1	9.9 * #
Word Identification	89.8	-0.2	2.4 *	-0.1	2.2 *	-0.6	2.9 *	1.1	0.9	-2.5	3.9	1.2	1.8
TOWRE SWE	87.8	3.3	2.1 *	3.3	2.3	3.1	1.6 *	4.9	-0.1	1.4	4.1	3.7	3.0
AIMSweb	44.1	27.1	10.9 *	25.4	12.3 *	32.4	6.6 *	30.7	1.9	6.3	29.3 *	39.1	5.7
Passage Comprehension	93.5	-5.0	6.0 * #	-5.0	5.5	-4.9	7.7 * #	0.3	-0.5	-16.5	15.7	1.2	1.3
GRADE	88.6	-8.3	0.3	-8.0	0.1	-9.2	0.8	-11.7	2.5	-8.4	-4.0	-3.9	1.8
Sample Size	139		139		96		43		38		24		34

		All Interventions		Word-level Interventions		Failure Free Reading		Spell Read		Wilson Reading		Corrective Reading	
Grade 5 Cohort	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	0.5	3.1 *	0.6	4.4 *	0.2	-1.0	-1.4	4.7 *	-0.8	10.0 *	4.1	-1.4
TOWRE PDE	82.0	4.6	2.1 *	4.4	2.8 *	5.3	0.1	3.5	4.6 *	2.2	4.0 *	7.5	-0.3
Word Identification	89.8	3.5	-0.9	3.9	-1.2	2.5	0.0	1.1	0.2	4.2	-0.3	6.2	-3.4
TOWRE SWE	85.5	3.1	1.5	3.2	1.4	3.0	1.9	1.0	3.4	2.7	2.4 *	5.8	-1.7 #
AIMSweb	82.3	33.1	-4.1	33.8	-5.6 * #	30.9	0.2	29.6	-6.5 * #	33.1	-2.6	38.8	-7.6 * #
Passage Comprehension	95.1	0.0	-1.5	-0.2	-1.4	0.4	-1.8	-3.6	-0.7	0.9	-1.7	2.2	-1.8
GRADE	94.8	-5.1	0.9	-5.7	1.3	-3.1	-0.1	-9.0	2.6 #	-9.0	3.1	0.9	-1.7
Sample Size	180		180		122		58		50		32		40

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

* 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 Grade Cohorts
One Year After the Intervention Year

Grade 3 Cohort	All Interventions ABCD Effect Size	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
Word Attack	0.36 *	0.37 *	0.33 *	0.36 *	0.38 *	0.35 *
TOWRE PDE	0.26 *	0.32 *	0.08	0.33 *	0.28 *	0.37 *
Word Identification	0.15 *	0.17 *	0.12	0.05	0.28 *	0.17 *
TOWRE SWE	0.11 *	0.11 *	0.13 *	0.06	0.17 *	0.09
AIMSweb	0.14	0.11	0.20 *	0.15 *	0.09	0.10
Passage Comprehension	0.14 *	0.09	0.29 *	0.01	0.23	0.02
GRADE	0.06	0.02	0.19	0.14	0.00	-0.07

Grade 5 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
Word Attack	0.18 * #	0.26 *	-0.05 #	0.23	0.52 *	0.01 #
TOWRE PDE	0.11	0.16 *	-0.02	0.21	0.17	0.09
Word Identification	-0.04 #	-0.04 #	-0.04 #	0.01	0.00 #	-0.13 #
TOWRE SWE	0.09	0.09	0.10	0.23 *	0.08	-0.03
AIMSweb	-0.08 * #	-0.08 * #	-0.09 #	-0.07 #	-0.06	-0.11
Passage Comprehension	-0.08 #	-0.05	-0.17 #	-0.06	0.06	-0.14
GRADE	0.05	0.08	-0.06	-0.07	0.31	0.00

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

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

* Impact is statistically significant at the 0.05 level.
# Impact is statistically different from 3rd grade cohort impact at the 0.05 level.

Table IV.12

Effect Sizes for 3rd and 5th Grade Cohorts with Low Baseline Word Attack Scores
One Year After the Intervention Year

Grade 3 Cohort	All Interventions ABCD Effect Size	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
Word Attack	0.28 *	0.31 *	0.18	0.34 *	0.18	0.42
TOWRE PDE	0.26 *	0.31 *	0.09	0.36 *	0.00 #	0.57
Word Identification	0.11 *	0.11 *	0.12	0.02	0.20 *	0.11
TOWRE SWE	0.06 #	0.04 #	0.13	0.06	0.17	-0.11 #
AIMSweb	0.06	0.00	0.25 *	0.16 *	-0.08	-0.08
Passage Comprehension	0.10	0.09	0.14	-0.14	0.26	0.15
GRADE	0.06	0.07	0.03	-0.17 #	0.54 #	-0.15

Grade 5 Cohort	All Interventions ABCD Effect Size	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
Word Attack	0.29 * #	0.40 * #	-0.06	0.42 * #	0.52 *	0.02 *
TOWRE PDE	0.13	0.21 *	-0.11	0.23	0.19	0.20
Word Identification	0.00	-0.01	0.02	0.02	-0.09	0.04
TOWRE SWE	0.14 *	0.14	0.15 *	0.26 *	0.09	0.08
AIMSweb	-0.12 *	-0.12 *	-0.14	-0.17	-0.15 *	-0.04
Passage Comprehension	0.04	0.09	-0.10	0.02	0.22	0.03
GRADE	0.25 * #	0.31 * #	0.06	-0.08	0.70 * #	0.31 #

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

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

* 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.13

Effect Sizes for 3rd and 5th Grade Cohorts with High Baseline Word Attack Scores
One Year After the Intervention Year

Grade 3 Cohort	All Interventions ABCD Effect Size	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
Word Attack	0.03 *	0.37 *	0.62 *	0.25	0.32 *	0.56 *
TOWRE PDE	0.24 *	0.32 *	0.00	0.36 *	0.40 * #	0.19 *
Word Identification	0.20 *	0.20 *	0.20 *	0.06	0.29 *	0.25
TOWRE SWE	0.19 * #	0.18 * #	0.21 *	0.06	0.20 *	0.28 * #
AIMSweb	0.12 *	0.16 *	0.02	0.10	0.28 *	0.11
Passage Comprehension	0.13 *	0.04	0.40 *	0.08	-0.05	0.09
GRADE	-0.03	-0.16	0.35 *	0.33 #	-0.70 * #	-0.10

Grade 5 Cohort	All Interventions ABCD Effect Size	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
Word Attack	0.05 #	0.08 #	-0.01	-0.12 #	0.40 *	-0.05
TOWRE PDE	0.08	0.07	0.10	0.13	0.07	0.02
Word Identification	-0.12 *	-0.11	-0.15	-0.15	0.04	-0.22
TOWRE SWE	0.07	0.08	0.05	0.22 *	0.04	-0.02
AIMSweb	-0.03	-0.03	-0.05	0.07	0.01	-0.16
Passage Comprehension	-0.14 *	-0.12	-0.19	-0.16	0.00	-0.21 *
GRADE	-0.11 #	-0.07 #	-0.23 *	0.01	-0.07 #	-0.15 * #

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

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

* 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.14

Effect Sizes for 3rd and 5th Grade Cohorts with Low Screening Peabody Picture Vocabulary Test Scores
One Year After the Intervention Year

Grade 3 Cohort	All Interventions ABCD Effect Size	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
Word Attack	0.27 *	0.20 *	0.46 *	-0.05 #	0.41 *	0.24 *
TOWRE PDE	0.24 *	0.26 *	0.19	0.14	0.24 *	0.40 * #
Word Identification	0.11	0.11	0.11	-0.05	0.26 *	0.13
TOWRE SWE	0.07	0.07	0.06	0.05	0.14	0.01
AIMSweb	-0.05 #	-0.16 #	0.27	-0.05	-0.41 #	-0.02
Passage Comprehension	0.04	0.01	0.14	0.03	-0.08	0.08
GRADE	0.07	0.13	-0.10 #	0.39	0.22	-0.22

Grade 5 Cohort	All Interventions ABCD Effect Size	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
Word Attack	0.16 *	0.26 *	-0.13	0.08	0.48 *	0.21
TOWRE PDE	0.15 *	0.17 *	0.08	0.06 #	0.14	0.31
Word Identification	-0.01	-0.03	0.03	-0.08 #	0.09	-0.10
TOWRE SWE	0.16 *	0.20 *	0.04	0.18 *	0.14	0.27 * #
AIMSweb	-0.14	-0.11	-0.24	-0.14	0.00	-0.18
Passage Comprehension	0.10 #	0.15 * #	-0.02	0.09	0.22 * #	0.14
GRADE	0.23 *	0.22 *	0.27 * #	-0.02	0.34 *	0.35 *

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

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

* 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.15

Effect Sizes for 3rd and 5th Grade Cohorts with High Screening Peabody Picture Vocabulary Test Scores
One Year After the Intervention Year

Grade 3 Cohort	All Interventions ABCD Effect Size	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
Word Attack	0.29 *	0.34 *	0.13	0.51 * #	0.38	0.14 *
TOWRE PDE	0.18 *	0.20 *	0.13	0.44 *	0.18	-0.01 #
Word Identification	0.15 *	0.17 *	0.11 *	0.10	0.28 *	0.13
TOWRE SWE	0.13 *	0.12	0.18 *	0.07	0.16	0.12
AIMSweb	0.25 * #	0.28 * #	0.17 *	0.25 *	0.46 * #	0.13
Passage Comprehension	0.20 *	0.15	0.36 *	-0.04	0.44	0.04
GRADE	0.21	0.14	0.42 #	0.30	-0.17	0.30

Grade 5 Cohort	All Interventions ABCD Effect Size	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
Word Attack	0.15 *	0.21 *	-0.04	0.36 *	0.34	-0.08
TOWRE PDE	0.16 *	0.25 *	-0.10	0.45 * #	0.23	0.07
Word Identification	-0.04	-0.04	-0.04	0.12 #	-0.14	-0.11
TOWRE SWE	0.07	0.06	0.11	0.25 *	0.10	-0.17 #
AIMSweb	-0.12 *	-0.13 *	-0.10	-0.06	-0.14	-0.20 *
Passage Comprehension	-0.15 #	-0.12 #	-0.23	-0.15	-0.11 #	-0.11
GRADE	-0.02	0.04	-0.19 #	-0.09	0.33	-0.13

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

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

* 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.16

Effect Sizes for 3rd and 5th Grade Cohorts with Low Baseline Word Attack and Low Screening Peabody Picture Vocabulary Test Scores One Year After the Intervention Year

Grade 3 Cohort	All Interventions ABCD Effect Size	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
Word Attack	0.24 *	0.30 *	0.07	0.28	0.24	0.39 *
TOWRE PDE	0.09	0.08	0.12	-0.24 #	0.25	0.22
Word Identification	0.04	0.03	0.10	-0.09	0.09	0.09
TOWRE SWE	-0.04 #	-0.04	-0.03	0.18	-0.28	-0.02
AIMSweb	-0.17 * #	-0.27 * #	0.12	-0.08	-0.90 * #	0.16
Passage Comprehension	0.04	0.00	0.17	-0.13	0.08	0.04
GRADE	-0.04	0.09	-0.45 #	0.27	0.28	-0.28

Grade 5 Cohort	All Interventions ABCD Effect Size	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
Word Attack	0.18 *	0.28 *	-0.13	0.29 *	0.23 #	0.32 * #
TOWRE PDE	0.18 *	0.21 *	0.07	0.06	0.20	0.39
Word Identification	0.04	0.03	0.10	-0.09	0.09	0.09
TOWRE SWE	0.20 *	0.22	0.14	0.22 *	0.15	0.29
AIMSweb	-0.23 *	-0.26 *	-0.14	-0.24	-0.16	-0.38
Passage Comprehension	0.26 * #	0.34 * #	0.01	0.38 * #	0.19	0.45 * #
GRADE	0.18	0.12	0.38 #	-0.17	0.30	0.22

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

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

* 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.17

Effect Sizes for 3rd and 5th Grade Cohorts With Low Baseline Word Attack and High Peabody Picture Vocabulary Test Scores
One Year After the Intervention Year

Grade 3 Cohort	All Interventions ABCD Effect Size	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
Word Attack	0.00 #	0.01	-0.02	0.20	0.23	-0.40 * #
TOWRE PDE	0.12	0.13 #	0.10	0.35 *	0.21	-0.18 #
Word Identification	0.14 *	0.15	0.10	0.14	0.38 *	-0.06
TOWRE SWE	0.09	0.08	0.10	0.01	0.37	-0.13
AIMSweb	0.24 *	0.27 *	0.17 *	0.22	0.47	0.11
Passage Comprehension	0.24 *	0.24	0.26 *	-0.13	0.63	0.21
GRADE	0.18	0.06	0.53 * #	-0.24	0.32	0.11

Grade 5 Cohort	All Interventions ABCD Effect Size	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
Word Attack	0.31 *	0.43 * #	-0.08	0.48 *	0.71 *	0.11
TOWRE PDE	0.23 *	0.36 * #	-0.16	0.48 * #	0.40 *	0.21 *
Word Identification	-0.02	0.02	-0.12 * #	0.15 #	0.00	-0.09
TOWRE SWE	0.17 *	0.19 *	0.14	0.29 *	0.22	0.05
AIMSweb	-0.16	-0.13	-0.26	-0.04	-0.31 * #	-0.03
Passage Comprehension	-0.04	-0.04	-0.06	-0.26	0.22	-0.07
GRADE	0.43 * #	0.61 * #	-0.09	0.34	0.99 * #	0.49

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

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

* 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.18

Effect Sizes for 3rd and 5th Grade Cohorts with High Baseline Word Attack and High Screening Peabody Picture Vocabulary Test Scores One Year After the Intervention Year

Grade 3 Cohort	All Interventions ABCD Effect Size	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
Word Attack	0.38 *	0.44 *	0.20 *	0.69 * #	0.29 *	0.35 *
TOWRE PDE	0.17 *	0.25 *	-0.08	0.34 *	0.27 *	0.14
Word Identification	0.11 *	0.16 *	-0.05	0.06	0.07	0.34 *
TOWRE SWE	0.17 *	0.17 *	0.15 *	0.01	0.17 *	0.34 * #
AIMSweb	0.13	0.18	0.00	0.11	0.40 *	0.01
Passage Comprehension	0.17 *	0.06	0.50 *	0.02	0.04	0.11
GRADE	0.07	0.07	0.07	0.43	-0.31	0.08

Grade 5 Cohort	All Interventions ABCD Effect Size	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
Word Attack	0.08	0.11	0.00	0.14	0.29	-0.09
TOWRE PDE	0.15 *	0.18 *	0.07	0.25	0.21	0.08
Word Identification	-0.08	-0.10	-0.03	-0.17 *	-0.01	-0.10
TOWRE SWE	0.10	0.09	0.13	0.08	0.30	-0.11
AIMSweb	-0.05	-0.07	-0.02	0.02	0.09	-0.31 * #
Passage Comprehension	-0.06	-0.02	-0.19	-0.02	0.03	-0.07
GRADE	-0.29 * #	-0.30 * #	-0.27 *	-0.36	-0.17	-0.36 * #

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

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

* 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.19

Effect Sizes for 3rd and 5th Grade Cohorts Eligible for Free or Reduced-Price School Lunch
One Year After the Intervention Year

Grade 3 Cohort	All Interventions ABCD Effect Size	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
Word Attack	0.24 *	0.18 *	0.45 *	0.10	0.28 *	0.16 *
TOWRE PDE	0.19 * #	0.21 * #	0.13	0.36 *	0.21 * #	0.05 #
Word Identification	0.21	0.23	0.17	-0.02	0.37 *	0.34 *
TOWRE SWE	0.14 *	0.10	0.24 *	0.11	0.19 *	0.01 *
AIMSweb	0.07	0.04	0.15	0.09	0.01	0.03
Passage Comprehension	-0.01 #	0.00	-0.02 #	-0.03	0.01	0.00
GRADE	0.07	-0.03	0.37	0.01	0.10	-0.21

Grade 5 Cohort	All Interventions ABCD Effect Size	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
Word Attack	0.19 *	0.25 *	0.00	0.23	0.41 *	0.12
TOWRE PDE	0.08	0.10	0.00	0.04	0.01	0.26 *
Word Identification	-0.01	0.01	-0.06	0.04	0.03	-0.05
TOWRE SWE	0.17 *	0.20 *	0.09	0.27 *	0.10	0.24 * #
AIMSweb	0.04	0.10 #	-0.15	0.08 #	0.08	0.13 #
Passage Comprehension	-0.08	-0.03	-0.21	-0.14	0.07	-0.03
GRADE	-0.22 *	-0.19 *	-0.30 *	-0.63 * #	0.03	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

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

* 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.20

Effect Sizes for 3rd and 5th Grade Cohorts Not Eligible for Free or Reduced-Price School Lunch
One Year After the Intervention Year

Grade 3 Cohort	All Interventions ABCD Effect Size	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
Word Attack	0.40 *	0.52 *	0.07	0.44 *	0.65	0.47 *
TOWRE PDE	0.52 * #	0.69 * #	0.03	0.35 *	1.05 * #	0.66 * #
Word Identification	0.16 *	0.15 *	0.19 *	0.06	0.26	0.12
TOWRE SWE	0.14 *	0.16	0.10 *	0.00	0.27	0.20
AIMSweb	0.28 *	0.31 *	0.17 *	0.05	0.75 *	0.14
Passage Comprehension	0.40 * #	0.37	0.51 * #	-0.03	1.04	0.09
GRADE	0.02	0.01	0.05	0.17	-0.27	0.12

Grade 5 Cohort	All Interventions ABCD Effect Size	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
Word Attack	0.20 *	0.30 *	-0.07	0.31 *	0.67 *	-0.09
TOWRE PDE	0.14 *	0.18 *	0.01	0.30 *	0.27 *	-0.02
Word Identification	-0.06	-0.08	0.00	0.01	-0.02	-0.23
TOWRE SWE	0.10	0.09	0.13	0.23	0.16 *	-0.11 #
AIMSweb	-0.09	-0.12 * #	0.00	-0.14 * #	-0.06	-0.16 * #
Passage Comprehension	-0.10	-0.09	-0.12	-0.05	-0.11	-0.12
GRADE	0.06	0.09	-0.01	0.17 #	0.20	-0.12

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

Note: Raw scores were analyzed for the AIMSweb, and standard scores were analyzed for all other tests.

* 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.21

Relative Gap Reduction: All Interventions Combined
One Year After the Intervention Year

Grade 3 Cohort	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	92.4	0.50	97.4	92.1	0.17	0.53	5.3 *	0.68
TOWRE PDE	85.5	0.97	90.5	86.6	0.63	0.90	4.0 *	0.29
Word Identification	88.7	0.76	90.9	88.6	0.61	0.76	2.3 *	0.20
TOWRE SWE	86.6	0.90	91.7	90.0	0.55	0.67	1.7 *	0.17
AIMSweb	NA	NA	NA	NA	NA	NA	NA	NA
Passage Comprehension	91.6	0.56	93.4	91.3	0.44	0.58	2.1 *	0.24
GRADE	86.1	0.93	79.6	78.6	1.36	1.42	1.0	0.05

Grade 5 Cohort	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	93.2	0.45	97.3	94.7	0.18	0.36	2.7 *	0.50
TOWRE PDE	81.0	1.27	88.0	86.3	0.80	0.91	1.7	0.13
Word Identification	88.7	0.75	91.1	91.7	0.60	0.56	-0.6	-0.07
TOWRE SWE	84.2	1.05	88.6	87.2	0.76	0.85	1.4	0.11
AIMSweb	NA	NA	NA	NA	NA	NA	NA	NA
Passage Comprehension	92.6	0.49	91.0	92.2	0.60	0.52	-1.1	-0.15
GRADE	91.2	0.59	88.4	87.7	0.77	0.82	0.7	0.06

Note: RGR deined 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.

* Impact is statistically significant at the 0.05 level.

Table IV.22

Impacts for 3rd and 5th Grade Cohorts
Late March/Early April of the Intervention Year

Grade 3 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
PSSA Reading	-15.6	-3.8	-51.1 *	-39.9	52.5	-23.8
PSSA Math	20.2	14.2	38.4	-15.5	56.6 *	1.4
Sample Size	329	240	89	92	71	77

Grade 5 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
PSSA Reading	-27.3 *	-25.3	-33.4 *	-30.0	-23.8	-22.1
PSSA Math	-28.8 * #	-34.0 * #	-13.4	-20.1	-56.4 * #	-25.4 *
Sample Size	408	280	128	102	92	86

* Impact is statistically significant at the 0.05 level.
# Impact is statistically different from the 3rd grade cohort impact at the 0.05 level.

Table IV.23

Impacts for 3rd and 5th Grade Cohorts with Low Baseline Word Attack Scores
Late March/Early April of the Intervention Year

Grade 3 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
PSSA Reading	-1.9	25.9 #	-85.3 *	-13.6	59.7	31.4
PSSA Math	46.1 * #	51.3 * #	30.7	-32.9	120.7 *	65.9 * #
Sample Size	168	113	55	47	31	35

Grade 5 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
PSSA Reading	-29.3	-23.8	-45.8	-39.5	-43.7	11.7
PSSA Math	-10.8	-19.3	14.8	-16.7	-28.0	-13.2
Sample Size	200	142	58	60	44	38

* 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.24

Impacts for 3rd and 5th Grade Cohorts with High Baseline Word Attack Scores
Late March/Early April of the Intervention Year

Grade 3 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
PSSA Reading	-21.2	-23.1 #	-15.6	-63.7 *	25.0	-30.5
PSSA Math	-3.1 #	-13.7 #	28.4	-7.5	32.3	-65.8 * #
Sample Size	208	138	70	42	48	48

Grade 5 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
PSSA Reading	-36.3 *	-40.6	-23.3	-37.9	-17.5	-66.3
PSSA Math	-39.0 *	-42.8 *	-27.7	-12.2	-53.8	-62.2
Sample Size	203	148	55	57	55	36

* 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.25

Impacts for 3rd and 5th Grade Cohorts with Low Screening Peabody Picture Vocabulary Test Scores
Late March/Early April of the Intervention Year

Grade 3 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
PSSA Reading	25.1	37.3	-11.4	21.0	59.3	31.7
PSSA Math	83.9	81.3	91.7	76.4 #	126.9 *	40.8
Sample Size	143	108	35	36	43	29

Grade 5 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
PSSA Reading	2.0	2.2	1.6	-30.2	-56.4 *	92.9 * #
PSSA Math	-33.8 *	-40.2	-14.6	-42.9	-43.8	-34.0
Sample Size	203	148	55	57	55	36

* 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.26

Impacts for 3rd and 5th Grade Cohorts with High Screening Peabody Picture Vocabulary Test Scores
Late March/Early April of the Intervention Year

Grade 3 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
PSSA Reading	0.9	5.6	-13.0	-25.4	38.6	3.6
PSSA Math	6.7	13.5	-13.9	-41.3 #	75.1	6.8
Sample Size	186	132	54	56	28	48

Grade 5 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
PSSA Reading	-32.6 *	-26.7	-50.5 *	-3.6	-9.6	-66.8 * #
PSSA Math	-17.8	-22.4	-4.2	26.7	-82.0 *	-11.8
Sample Size	205	132	73	45	37	50

* 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.27

Impacts for 3rd and 5th Grade Cohorts with Low Baseline Word Attack and Low Screening Peabody Picture Vocabulary Test Scores
Late March/Early April of the Intervention Year

Grade 3 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
PSSA Reading	57.5	82.9 #	-18.8	34.8	101.3	112.6 * #
PSSA Math	131.9 * #	139.0 * #	110.8	125.0	172.1 *	119.8
Sample Size	77	53	24	23	15	15

Grade 5 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
PSSA Reading	-0.6	8.3	-26.8	-28.6	-94.3 *	147.4 * #
PSSA Math	-22.7	-26.2	-11.9	-32.6	-61.6	15.5
Sample Size	111	81	30	36	26	19

* 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.28

Impacts for 3rd and 5th Grade Cohorts with Low Baseline Word Attack and High Peabody Picture Vocabulary Test Scores
Late March/Early April of the Intervention Year

Grade 3 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
PSSA Reading	-4.8	8.3	-44.3	8.1	20.5	-3.6
PSSA Math	21.9	37.5	-24.6	-31.6	67.2	76.7 * #
Sample Size	91	60	31	24	16	20

Grade 5 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
PSSA Reading	-31.9	-25.3	-51.7	-16.0	-20.6	-39.3
PSSA Math	-3.0	-13.1	27.6	16.5	-61.7 *	5.8
Sample Size	89	61	28	24	18	19

* 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.29

Impacts for 3rd and 5th Grade Cohorts with High Baseline Word Attack and High Screening Peabody Picture Vocabulary Test Scores
Late March/Early April of the Intervention Year

Grade 3 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
PSSA Reading	-2.2	0.6	-10.7	-80.1 *	85.8 *	-3.8
PSSA Math	-11.1	-18.3	10.3	-60.4 *	73.1	-67.5 * #
Sample Size	95	72	23	32	12	28

Grade 5 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
PSSA Reading	-55.6 *	-61.0	-39.4	-44.1	-30.5	-108.3 * #
PSSA Math	34.7	-37.6	-26.1	-3.3	-65.4	-44.0
Sample Size	116	71	45	21	19	31

* 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.30

Impacts for 3rd and 5th Grade Cohorts Eligible for Free or Reduced Price School Lunch
Late March/Early April of the Intervention Year

Grade 3 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
PSSA Reading	-66.3 * #	-45.2 #	-129.5 * #	-57.7	-25.5	-52.5 #
PSSA Math	10.1	-0.8	42.9	-31.7	45.1 * #	-15.7
Sample Size	190	144	46	52	47	45

Grade 5 Cohort	All Interventions ABCD Impact	Word-level Interventions BCD Impact	Failure Free Reading A Impact	Spell Read B Impact	Wilson Reading C Impact	Corrective Reading D Impact
PSSA Reading	-74.1 * #	-67.3 *	-94.5 * #	-110.7 * #	-69.8 *	-21.5
PSSA Math	-35.3	-36.6	-31.3	-58.7	-30.3	-20.8
Sample Size	230	160	70	53	61	46

* 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.31

Impacts for 3rd and 5th Grade Cohorts Not Eligible for Free or Reduced Price School Lunch
Late March/Early April of the Intervention Year

Grade 3 Cohort	All Interventions ABCD Impact		Word-level Interventions BCD Impact		Failure Free Reading A Impact		Spell Read B Impact	Wilson Reading C Impact		Corrective Reading D Impact	
PSSA Reading	43.4	#	47.6	#	30.7	#	9.6	90.5		42.7	#
PSSA Math	-18.3		-34.3		29.8		-12.9	-100.3	#	10.2	
Sample Size	139		96		43		40	24		32	

Grade 5 Cohort	All Interventions ABCD Impact		Word-level Interventions BCD Impact		Failure Free Reading A Impact		Spell Read B Impact		Wilson Reading C Impact		Corrective Reading D Impact	
PSSA Reading	-21.5	#	-23.7		-15.0	#	-0.7	#	-37.3		-33.2	
PSSA Math	-11.8		-29.3		40.8		13.5		-101.7	*	0.2	
Sample Size	178		120		58		49		31		40	

* 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.

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, 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* 1993*,* 56: 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.

Pressley, M. "What Should Comprehension Instruction Be the Instruction Of?" In M.L. Kamil, P.B. Mosenthal, P.D. Pearson, and R. Barr, eds., *Handbook of Reading Research, Volume III*. Mahwah, NJ: Lawrence Erlbaum Publishers, 2000: 545-561.

Rashotte, C.A., K. MacFee, 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 Associates, 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, Joseph, David Myers, Allen Schirm, Elizabeth Stuart, Sonya Vartivarian, Wendy Mansfield, Fran Stancavage, Donna Durno, Rosanne Javorsky, and Cinthia Haan. *National Assessment of Title I Interim Report to Congress: Volume III: Closing the Reading Gap, First Year Findings from a Randomized Trial of Four Reading Interventions for Striving Readers.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, 2006.

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, 2006: 156-172.

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: 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 Educational Sciences, National Center for Education Statistics. "The Nation's Report Card: Reading 2005." NCES 2006-451. Washington, DC: NCES. Available online at [http://www.nces.ed.gov/nationsreportcard/pdf/main2005/2006451.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.