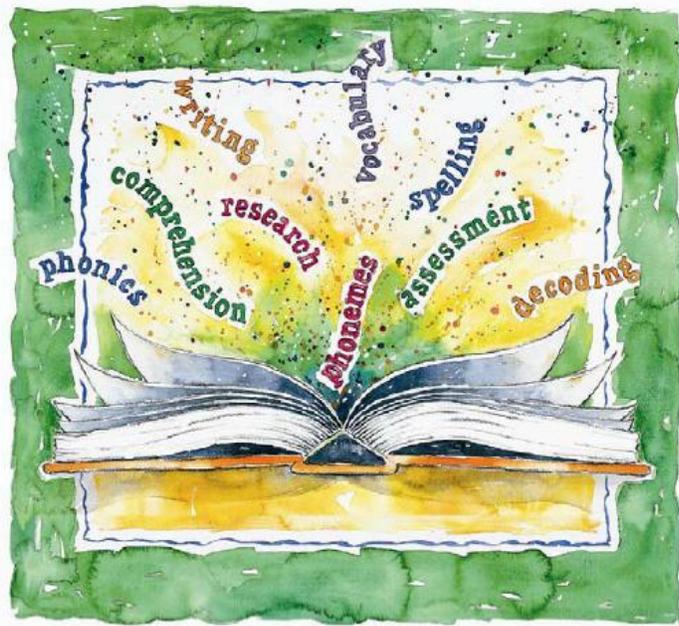# The California Reading First Year 5 Evaluation Report

November 2007

Prepared by:

Diane Haager, Ph.D.
Principal Investigator

and

Educational Data Systems
Renuka Dhar
Mark Moulton, Ph.D.
Susan McMillan, Ph.D.

EDUCATIONAL DATA SYSTEMS

Executive Summary: Chapters 1 – 3

For five years California has been participating in the Reading First program, a federal initiative aimed at improving reading instruction in the United States. This report provides an evaluation of Reading First implementation and student reading achievement in California for those five years. The key findings are summarized below, and the remainder of the report contains the detailed analysis to support these conclusions.

Finding #1:   Reading First is effective. After controlling for school demographic characteristics, Reading First implementation is a statistically significant predictor of achievement on all achievement metrics, especially those associated with grades 2 and 4. The more faithfully the program is applied, the greater the effect on achievement.

Finding #2: Growth has been significant. The Reading First Achievement Index (RFAI), a composite of K-3 achievement metrics for Reading First schools, has risen an average of 3.4 points per year, equivalent to 17 points over 5 years.

Finding #3: Reading First schools out-perform the control group. Reading First schools out-perform a statistical control group by 1.6 points per year on the RFAI, equivalent to an 8-point advantage over five years.

Finding #4: Reading First schools out-perform non-Reading First schools. While non-Reading First schools have also shown substantial growth since 2002, it is consistently and significantly less than the growth of Reading First schools, similar to the growth of the statistical control group.

Finding #5: The Reading First effect generalizes across students. Reading First effects generalize to all performance levels of the Reading First student population. On the California Standards Test (CST) metrics, the migration of students into "Proficient and Above" is matched by a comparable migration of students out of "Below Basic and Far Below Basic." These migrations are confirmed by average student CST scale score gains, on the order of 20 scale score points over a 5-year period. The rising average scale score shows that the student population, on average, experiences gains.

Finding #6: Reading First improves grade 4 performance. These findings are replicated in grade 4, even though Reading First is a K-3 program. In grade 4, Reading First schools grew 4.1 CST scale score points per year (20.5 scale score points over 5 years), versus 2.4 scale score points per year (12.0 points over 5 years) for the control group, a difference of 8.5 scale score points. Thus the program shows evidence of a sustained effect of Reading First. This is a new finding, not available in previous evaluation reports.

Finding #7: Implementation has been adequate. Most schools in the Reading First program are implementing the program adequately. The average level of implementation has risen throughout the duration of the Reading First program. The average Reading First Implementation Index (RFII) across all schools was 39 in 2006 and 2007, compared to 36 in 2004 and 2005.

Although there is ample room for improvement in program implementation and in the program itself, there are no significant negative findings to report.

Background

Reading First is a federal initiative aimed at improving reading instruction in America. Authorized in 2001 as part of the No Child Left Behind Act, Reading First promotes the use of scientifically based reading practices in grades K-3. The initiative provides a significant amount of federal funding for improving reading instruction for large proportions of students experiencing academic difficulty and socio-economic disadvantage.

The Reading First program began in California during the 2002-03 school year[1], five years ago. Its components include:

- Use of a state-adopted reading program

- Access to training programs authorized by state legislation and based on research-based reading instruction: Senate Bill (SB) 472 teacher and coach professional development and AB 75 principal professional development

- Access to assessment tools that test students' skills every six to eight weeks

- Hiring of reading coaches, expert teachers who support program implementation

Anecdotal evidence indicates that many non-Reading First schools have voluntarily been adopting some or all of these components over the same 5-year period, giving this evaluation study a relevance that extends beyond the Reading First population.

This report evaluates California's progress in implementation and achievement during the first five years of Reading First funding and provides information regarding program efficacy.

Chapter 1 provides an overview of Reading First and its history, data sources, and the research design. It also discusses demographic characteristics of four cohorts of Reading First schools and how they compare to non-Reading First schools.

---

[1] In this report, we generally refer to the "year" as that of the spring of the school year. For example, the 2003-2004 school year would be referred to as "2004."

Chapter 2 provides the achievement results for all Reading First schools (high implementing and low implementing), as well as for a statistical control group and for non-Reading First schools.

Chapter 3 provides Reading First Implementation Index (RFII) statistics. These measure fidelity of Reading First implementation and are computed for each school from data collected from surveys administered to every Reading First teacher, coach, and principal in California.

Attached are appendices (A – F), which give:

- state-level survey results for the teacher, coach and principal implementation surveys (Appendices A, B, and C, respectively)

- additional charts and graphs showing trends in achievement to supplement Chapter 2 (Appendix D)

- the RFAI calculation description and formula (Appendix E)

- listings of Reading First schools along with their RFAI and RFII scores for 2005-2007 (Appendix F)

Additional chapters will be merged into the Year 5 Report by early 2008. They will study the effects of individual program elements on achievement and implementation, the impact of coaching, the impact of Reading First on the performance of the English learner population, and the relative performance of English learners in AB 1485 "waivered classrooms" and English-only classrooms.

A Data Example from Grade 2

Our core findings are exemplified in Table ES.1.0 and Figures ES.1.0 – ES.1.2, representing the growth of Reading First schools on various grade 2 achievement metrics since their entry into the program five years ago. Similar charts for the other grades, school cohorts, and achievement metrics, as well as a summary table of gain scores for all Reading First schools, can be found in Chapter 2 of the main evaluation report. Because of their novelty and importance, we report the corresponding results for grade 4 in Table ES.2.0 and Figures ES 2.0 – ES 2.2, following the grade 2 trend-lines.

Table ES.1.0: CST Metric, Years in Program = 5, Grade = 2

Years in Program (YIP): 5 Grade: 2	Reading First Schools				All Non-Reading First Elementary Schools
	All Reading First Schools	High Implementation Schools (Avg. RFII > 41.4)	Low Implementation Schools (Avg. RFII < 36.0)	Statistical Control Group (RFII = 25.0)	
Number of Schools	259	28	101	N/A	4,053
% Proficient and Above					
2002	15.4	14.8	14.8	15.4	37.8
2007	34.2	36.7	33.0	30.4	52.3
Change Since Starting Year	**18.9**abc	**22.0**abc	**18.2**abc	**15.0**	**14.5**
% Below and Far Below Basic					
2002	54.3	53.8	55.6	54.3	30.5
2007	36.7	33.6	38.7	41.1	23.0
Change Since Starting Year	**-17.6**abc	**-20.2**abc	**-16.9**abc	**-13.2**	**-7.6**
Mean Scale Score					
2002	299.8	299.5	298.5	299.8	333.4
2007	324.7	328.6	322.3	318.8	350.9
Change Since Starting Year	**25.0**abc	**29.0**abc	**23.8**abc	**19.0**	**17.5**

[a] Significantly different (p < 0.05) relative to the "Statistical Control Group."
[b] Significantly different (p < 0.05) relative to "All Non-Reading First Elementary Schools."
[c] Significantly different (p < 0.05) relative to the starting year, i.e., significantly different from a gain of zero.
Note: Numbers reporting change since starting year were rounded and may not appear to be an exact difference between 2002 and 2007 figures.

This table pertains only to those schools that have been in the program five years (Cohort 1) and it reports only their grade 2 CST scores. Referring to the "All Reading First Schools" column, we note the following. There were 259 Reading First schools in this first cohort, which had data for grade 2. On the "% Proficient and Above" achievement metric, an average of 15.4 percent of students in these schools scored "Proficient and Above." By 2007, this percentage had increased so that 34.2 percent of students were scoring "Proficient and Above." The size of the gain was 34.2 minus 15.4, or 18.9 percentage points. Note that rounding accounts for any seeming discrepancies in computing the change from 2002 to 2007. The superscripts "abc" tell us this gain was "significantly" larger than the gains of the "statistical control group,"[2] the gain of the non-Reading First schools in California, and that the gain is significantly larger than zero. "Significant" means there is a 95% probability that a gain that large would not have occurred by chance.

---

[2] The "statistical control group" is a construct defined using multiple regression to hold the effects of school population characteristics stable while examining the independent effect of the Reading First program implementation statistic (RFII) on student achievement. For purposes of this discussion, the results of these analyses are referred to as a "statistical control group" because this approach is analogous to creating a control group of schools that are exactly like the Reading First schools, in terms of student characteristics, but without the influence of the Reading First program. See Chapter 2 for a more complete discussion.

Referring to the same column, we see the percent of students scoring "Below Basic or Far Below Basic" in 2002 and in 2007, and the subsequent change. This change is negative because it refers to students moving *out of* the bottom two performance level categories. Then we see the average student scale score (a test score ranging roughly from 250 to 450) in 2002 and in 2007, and the difference between them. Remember that these are students who were in grade 2 in 2002, and that there was another group of students who were in grade 2 in 2007. On average the 2007 students scored an average of 25 scale score points higher than their 2002 predecessors. For context, that is halfway between the "Basic" cut-point (300) and the "Proficient" cut-point (350).

The remaining columns report the same statistics for schools that have been classified as "high implementing" (using the RFII statistics, based on teacher, coach, and principal responses to the Reading First implementation survey) and "low implementing." The "Statistical Control Group" column reports the same statistics for a theoretical group of schools that are similar to the Reading First schools but not implementing the program. The last column reports the same statistics for the remaining 4,053 elementary schools in California that are not in the Reading First program. Note that this population has much higher starting scores than the Reading First schools. Therefore, for display in the trend-line charts, the starting points for "All Non-Reading First Schools" have been adjusted downward to coincide with the starting points of the Reading First groupings.

Comparing the bolded gain scores across the columns, we see that "All Reading First" schools grew faster than the "Statistical Control Group," that "High Implementation" schools grew faster than "Low Implementation" schools, and that they all grew faster than the "Non-Reading First" elementary schools in the rest of the state. All differences are statistically significant.

These findings demonstrating the efficacy of Reading First are consistent with the previous reports. They extend and confirm the findings from the Year 4 and Year 3 California Reading First Evaluation Reports.

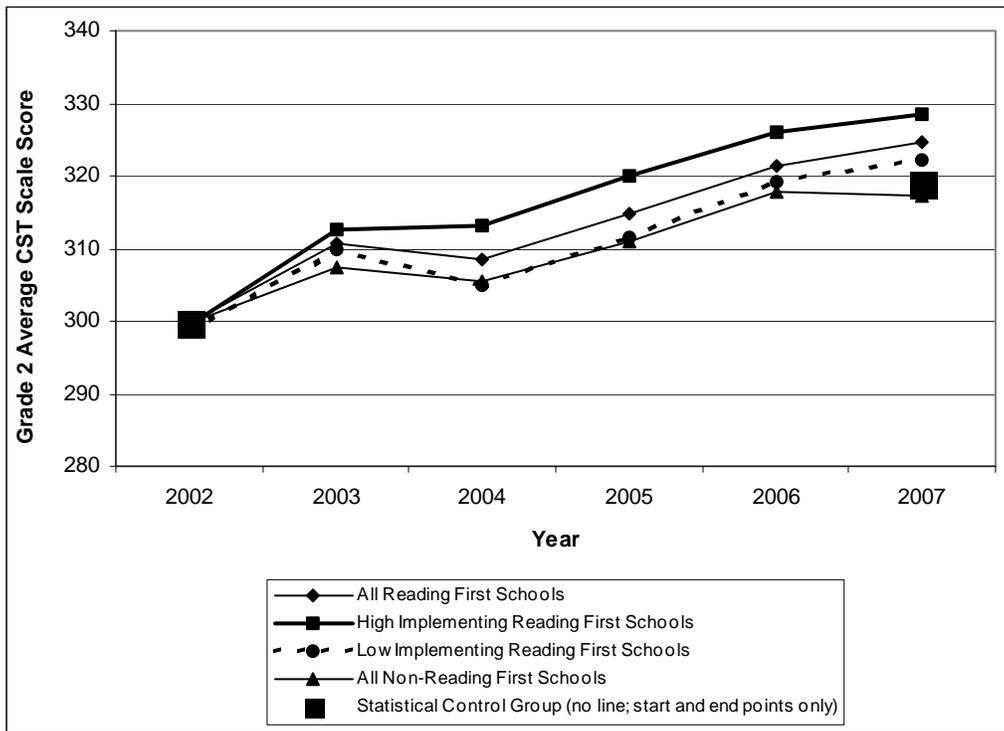The trend-lines corresponding to Table ES.1.0 are presented below.

Figure ES.1.0: CST % Proficient & Above, YIP = 5, Grade = 2
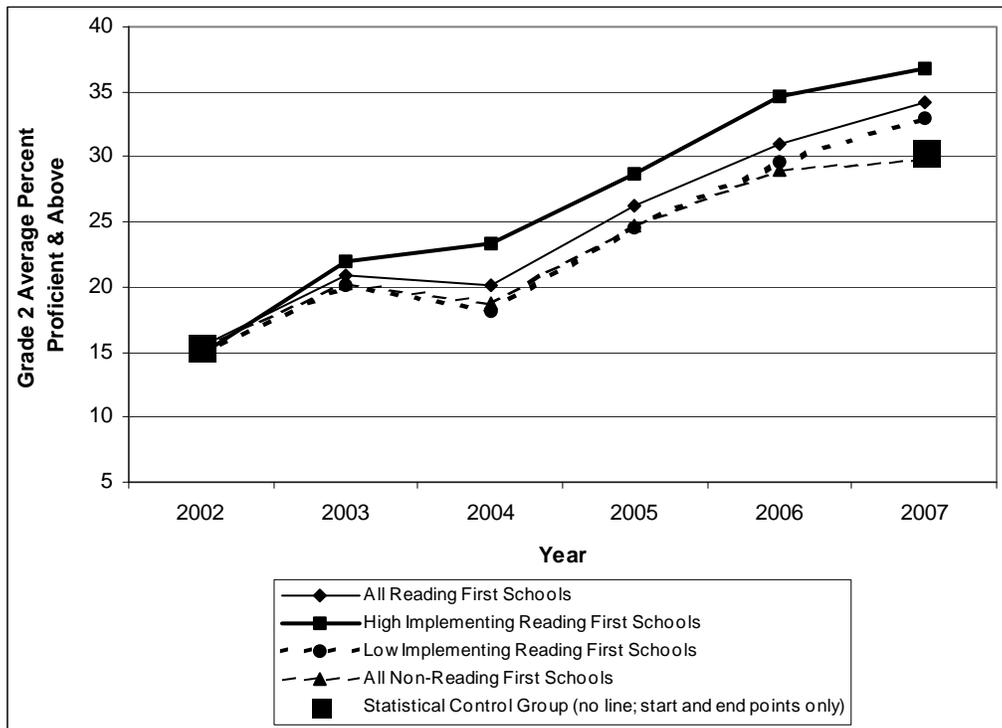
Figure ES.1.1: CST % Below Basic & Far Below Basic, YIP = 5, Grade = 2
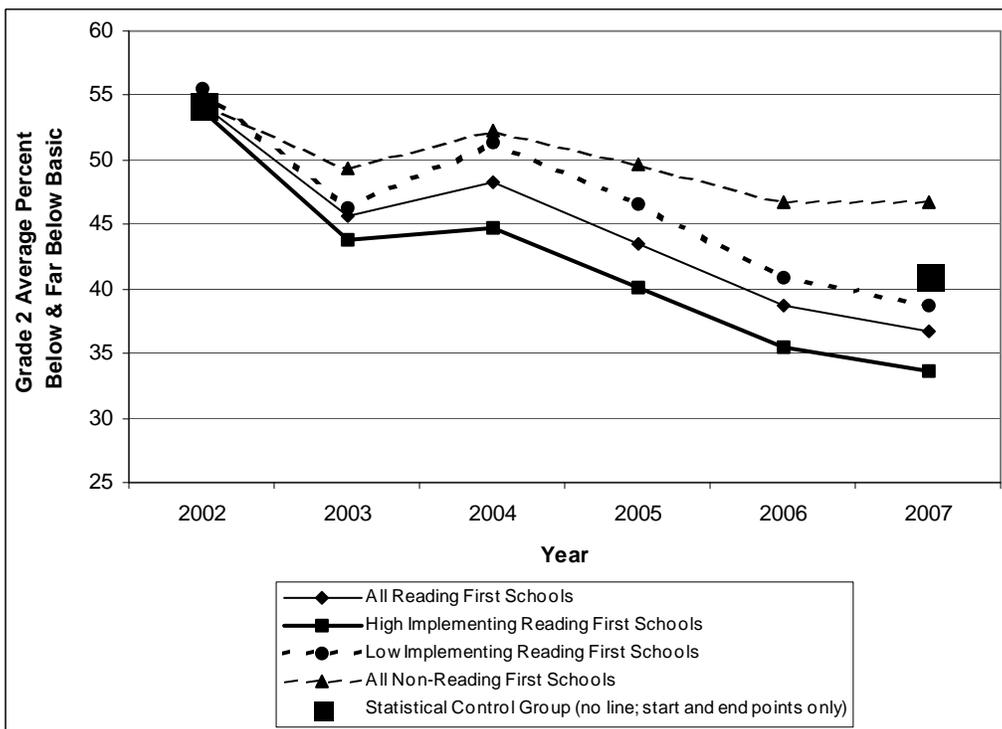
Figure ES.1.2: CST Mean Scale Score, YIP = 5, Grade = 2

A Data Example from Grade 4

This Year 5 Reading First Evaluation Report expands the scope of the evaluation by adding grade 4 CST performance as an achievement outcome. Because Reading First is administered only in grades K-3, the grade 4 results shed light on whether student exposure to Reading First in the earlier grades improves their ability to read in grades 4 and above. Table ES.2.0 and Figures ES.2.0 – ES.2.2 show that it does.

Table ES.2.0: CSTs, YIP = 5, Grade = 4

Years in Program (YIP): 5 Grade: 4	Reading First Schools				All Non-Reading First Elementary Schools
	All Reading First Schools	High Implementation Schools (Avg. RFII > 41.4)	Low Implementation Schools (Avg. RFII < 36.0)	Statistical Control Group (RFII = 25.0)	
Number of Schools	255	26	101	N/A	3,992
% Proficient and Above					
2002	15.2	14.2	15.9	15.2	42.1
2007	31.3	34.8	30.0	27.5	55.7
Change Since Starting Year	**16.1**[abc]	**20.6**[abc]	**14.1**[ac]	**12.3**	**13.6**
% Below and Far Below Basic					
2002	47.8	48.6	46.7	47.8	23.2
2007	32.2	28.4	33.4	34.9	16.9
Change Since Starting Year	**-15.6**[abc]	**-20.2**[abc]	**-13.3**[bc]	**-12.9**	**-6.3**
Mean Scale Score					
2002	306.8	305.1	307.7	306.7	340.9
2007	327.3	331.3	325.8	322.1	359.5
Change Since Starting Year	**20.5**[abc]	**26.2**[abc]	**18.1**[ac]	**15.4**	**18.6**

[a] Significantly different (p < 0.05) relative to the "Statistical Control Group."
[b] Significantly different (p < 0.05) relative to "All Non-Reading First Elementary Schools."
[c] Significantly different (p < 0.05) relative to the starting year, i.e., significantly different from a gain of zero.

The number of schools in Table ES.2.0 differs from that in Table ES.1.0 because not all schools have the same grade configurations. As noted above, any seeming discrepancies in computing the change from 2002 to 2007 are the result of rounding.

Figure ES.2.0: CST % Proficient & Above, YIP = 5, Grade = 4
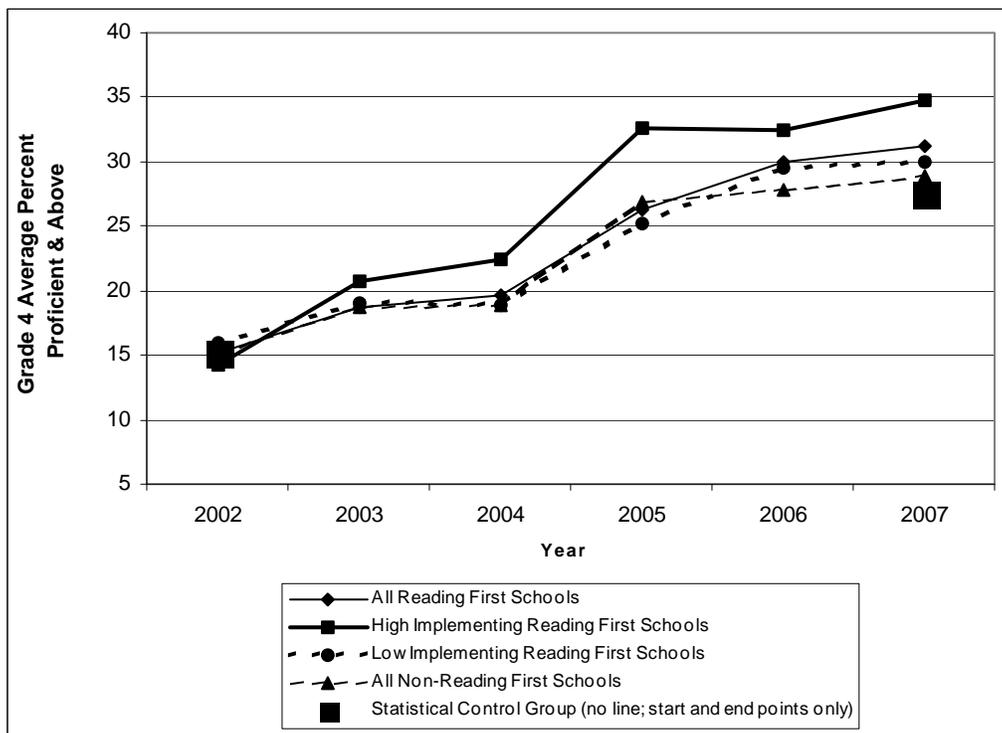
Figure ES.2.1: CST % Below Basic & Far Below Basic, YIP = 5, Grade = 4
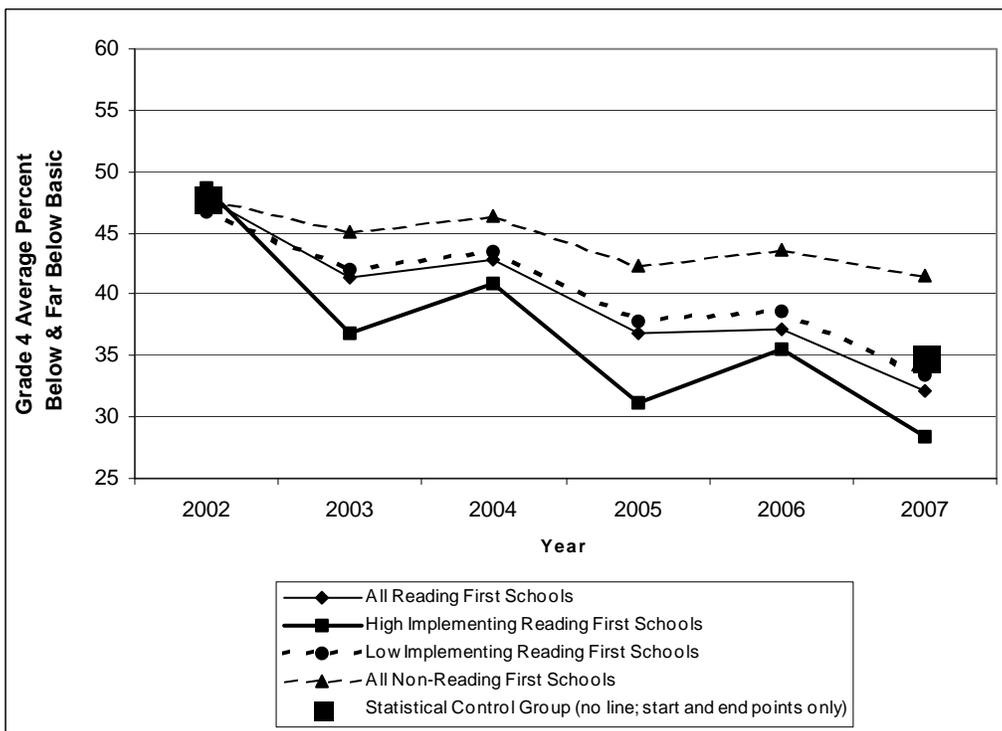
Figure ES.2.2: CST Mean Scale Score, YIP = 5, Grade = 4
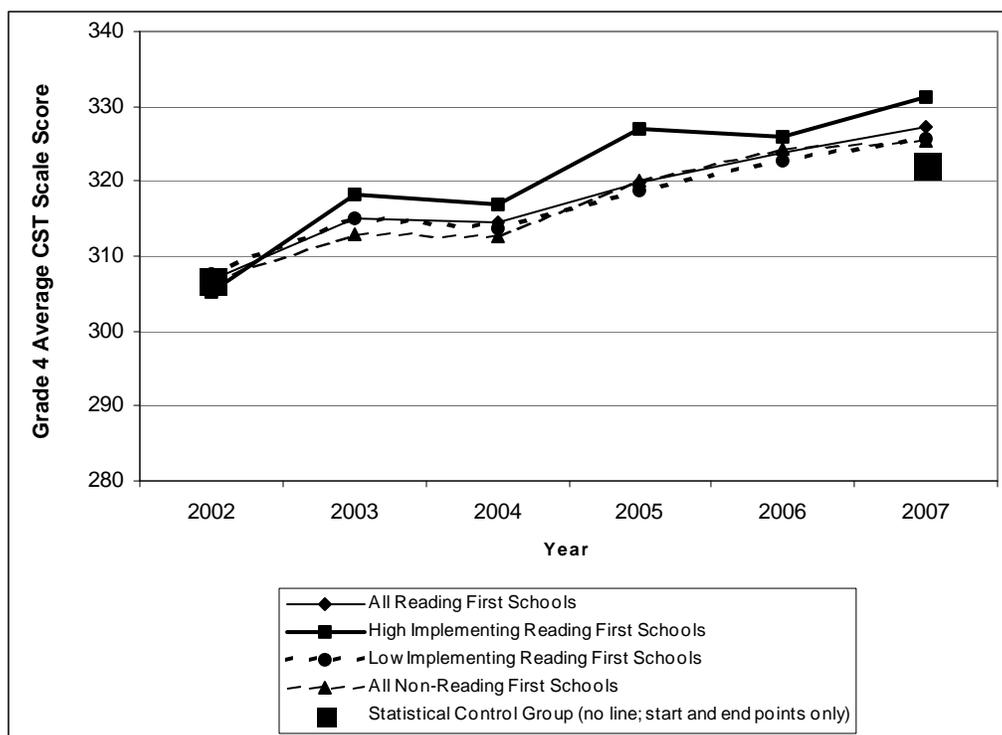
Policy Recommendations

The policy recommendations of the California Reading First Year 5 Evaluation Report are unchanged from those of the California Reading First Year 4 Evaluation Report.[3] To recap:

- Continue to focus on full implementation of Reading First.

- Support participation in Reading First over multiple years.

- Provide for a statewide data collection effort to facilitate accurate comparisons of student achievement across Reading First and non-Reading First schools, focusing on their use of the components that are required in Reading First. A statewide database of teacher and school data would confirm or refute the hypothesis that the statewide trend toward higher student proficiency is the result of the voluntary adoption of program elements that happen to be required as part of Reading First.

- Continue to support the extensive and focused professional development provided to teachers, coaches, and administrators. Continue to support the coaching model as a means for achieving instructional coherence and implementation of research-based instruction.

---

[3] *The California Reading First Year 4 Evaluation Report* can be accessed online at: www.eddata.com/resources/publications/ .

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter 1: Introduction and Demographics

Overview of California's Reading First Program

Reading First is a federal initiative that was authorized in 2001 as part of the No Child Left Behind Act (NCLB). This program, intended to improve reading outcomes in the nation, promotes the use of instructional practices and curricula based on scientifically based reading research in grades K-3. On August 23, 2002, the State of California was approved to receive approximately $900 million over a six-year period. According to federal Reading First guidelines, continued funding for states depends on demonstrating "significant progress" toward the goal that all children learn to read on grade level by the third grade. With Reading First funds, California has established a system to provide training, assist local educational agencies (LEAs) in acquiring curricular materials, monitor progress toward goals, and provide technical assistance to participating schools and school districts. This report provides an external evaluation of California's implementation of Reading First and student reading achievement for five years of implementation from academic year 2002-03 to 2006-07.

The California Reading First Plan delineates the roles and operational procedures for personnel involved at the state and local levels. The State Board of Education (SBE), Office of the Secretary of Education (OSE), and the California Department of Education (CDE) direct the Reading First program in California. The Reading and Literacy Partnership Team, with membership broadly representing the interests of reading education in the state, serves an advisory role for Reading First. A subcommittee of the Partnership, the Evaluation Advisory Group (EAG), including designees of the members, advises the external evaluator. The California Technical Assistance Center (C-TAC) has responsibility for the statewide technical assistance program and oversight of the Regional Technical Assistance Centers (R-TACs) in providing regional and local support to LEAs. It also coordinates the statewide network of professional development programs for teachers and site administrators through the Reading Implementation Centers (RICs).

The California Reading First Plan is based on a series of Assurances that are implemented by the LEAs. With these assurances, California's Reading First program is designed to ensure full implementation with fidelity to a comprehensive research-based reading program. Here, we briefly describe the assurances and program elements designed to address them.[1]

---

[1] For a complete description of the program elements, we refer the reader to previous evaluation reports, available at: http://eddata.com/resources/publications/ and the state's Reading First plan, available at: http://www.cde.ca.gov/nclb/sr/rf/.

Vision Statement

Each LEA and participating school must articulate a vision that reflects the goals and objectives of Reading First, including the belief that all children can learn to read with adequate instruction.

Curriculum

Participant LEAs are required to use one of California's two state-adopted reading curricula: SRA/McGraw-Hill's *Open Court Reading 2000* or *2002* (OCR) or the Houghton Mifflin *Reading: A Legacy of Literacy 2003* (HM). The Reading First program has provided extensive support for LEAs in the implementation of the adopted curricula. In the 2004-05 school year, California's Reading First program began offering support for LEAs with "waivered" classrooms, that is, classrooms offering a bilingual instruction model using Spanish-language versions of the adopted curricula. California law (Proposition 227) mandates instruction in English for all students unless parents sign a waiver specifically requesting bilingual instruction. The two state-adopted Spanish language reading programs are: SRA/McGraw Hill's *Foro abierto para la lectura* and Houghton Mifflin's *Lectura: Herencia y futuro*. Students receiving bilingual reading instruction in Spanish and English must transition out of bilingual instruction into English instruction, and take the English Standardized Testing and Reporting (STAR) English Language Arts Content Standards Test (CST) at the end of grade 2 and grade 3. Regardless of the LEA's selected curriculum, each LEA is required to implement fully the district's state-adopted reading/language arts program for an uninterrupted 60 minutes per day in kindergarten and 150 minutes per day in Grades 1-3, according to a district-approved pacing plan that outlines when each daily lesson is taught at each grade level in an academic year. This plan not only assures that students will complete the grade-level curriculum but also that implementation occurs systematically in every Reading First school. Also, LEAs are beginning to plan and implement extensive intervention with those K-3 students who need an additional 30 minutes of instruction. The intervention materials are approved by the SBE as scientifically research-based.

Professional Development

LEAs must assure that all K-3 teachers in Reading First schools annually participate in 40-hour training focused on the adopted core reading program. Year 1 teachers attend a state-approved training as mandated in Senate Bill (SB) 472. For Years 2-5, the LEAs must provide advanced levels of professional development, either provided through trainings developed by the C-TAC and delivered through the Reading Implementation Centers (RICs), or provided by the LEA. In addition, LEAs must provide access to these trainings for their K-12 special education teachers who are teaching K-3 reading, using either the LEAs' adopted core or intensive intervention reading program. LEAs are encouraged to provide continuous training to principals with the use of the C-TAC developed administrator modules (1-3 hours)

on implementing the adopted reading program and providing instructional leadership. Training of LEA trainers on these modules is provided by the C-TAC.

Curriculum-Embedded Assessment

For program monitoring, LEAs are required (since 2005-06) to use curriculum-embedded assessments conducted every 6 to 8 weeks. Teachers, administrators, and coaches use the data to make instructional adjustments and to identify individual students who need extra assistance. The results of the End-of-Year (EOY) tests – the curriculum-based assessment administered at the end of the school year—are required to be submitted to the State by each school. The results of these assessments are used as part of the Reading First Achievement Index (RFAI; see Chapter 2 of this report).

Collaborative Teacher Meetings

All Reading First schools are required to hold regular grade-level meetings twice a month to provide an opportunity for teachers to work together to refine their implementation of the program. School principals and reading coaches are encouraged to assist in facilitating and supporting these meetings.

District Commitment

Each LEA is required to conduct an internal evaluation on the effectiveness of its implementation of the Reading First program. This evaluation includes a district action plan for the subsequent year and each school's action plan for its first tri-semester based on student achievement data and principal, coach, and teacher recommendations. In addition, district personnel must assure that the Reading First program is well coordinated with other programs such as Title I, Language Acquisition, and Special Education. Each LEA must have a district Reading First Leadership Team that meets regularly to advise and support the program.

Coaching

LEAs may use Reading First funds to provide reading coaches, content experts, and coach coordinators and ensure that these experts are adequately trained. Coaches offer site-specific support for implementation of the LEA's adopted reading curriculum and effective instructional strategies. The C-TAC has provided these experts (1,371) two Coach Institutes annually for in-depth training and a Leadership Program for selected experts (110) in partnership with a California university. Additional training for new coaches is provided by the RICs, and support for both coach and coach coordinators is offered by the R-TACs.

Site Leadership

The site administrator's role is to support the full implementation of the school's adopted reading program and the state's Assurances. Administrators must attend the state's 40-hour AB 75/AB 430 training program to become fully knowledgeable of the reading program and participate in 40 hours of aligned activities within a two-year period. LEAs are also required to provide on-going training annually and are encouraged to use the C-TAC provided administrator modules.

Program Coherence

Reading First schools must ensure that any supplemental programs or materials are fully aligned with the adopted reading program, if using Reading First funds. LEAs are encouraged to use the SBE approved intervention and diagnostic assessment materials that offer extensive intervention. All categorical programs such as Language Acquisition, Title I, School Improvement, and Special Education programs, must be coordinated with the core program.

State Leadership

The CDE has designated key personnel to oversee and facilitate the administration of Reading First grants to LEAs, the contract with the external evaluator, and communications and legislation for the Reading First program. The SBE serves as the state educational agency for Reading First and works collaboratively with the CDE and the governor's office to develop and approve policy decisions regarding Reading First.

Technical Assistance

In addition to the statewide technical assistance programs provided by the C-TAC, the R-TACs, housed in county offices of education throughout the state, work directly with LEAs for full implementation of the Assurances. Some of their required activities include conducting classroom observations with LEAs' leadership team members; offering workshops on assessment, internal evaluation reporting, and interventions; and providing consultation on next steps to be taken by LEAs to meet goals of Reading First.

LEA Cohorts

California has now completed five years of implementation of the Reading First program. LEAs have been added to the program in cohorts. The first year, 2002-03, can be characterized as a start-up year because LEAs did not have a full year in which to implement. Cohort 1 (347 schools) has been receiving funding and implementing the program for approximately four and one-half years. LEAs in Cohort 2 (372 schools) were selected for funding in 2003-04. Cohort 3 (146 schools) was added in 2004-05. A small

number of LEAs were added in 2006-07 to make a new cohort, Cohort 4 (21 schools). A total of 886 schools in 120 LEAs are included in this Reading First Year 5 report.

California Reading First Year 5 Evaluation Study Design

The California Reading First Plan includes an annual external evaluation to study the implementation of the program and the resulting student achievement. Educational Data Systems (EDS[2]) has been the contractor for the Reading First evaluation study for each year of the program and has completed prior reports for Years 1 through 4. This current report represents the Year 5 evaluation report, and will include outcomes from the 2006-07 academic year and cumulative effects.

This report is guided by five research questions as stated in the scope of work for the external evaluation study. Two questions address program implementation:

1. How well did participating LEAs and schools implement their Reading First grants in accordance with California's Reading First plan?

2. What resources, support, and professional development activities are district-level administrative staff, school site administrators, and classroom teachers receiving in implementing the Reading First grants?

Three additional questions focus on the impact of Reading First:

3. What is the impact of the Reading First program on K-3 students in participating districts and schools?

4. What evidence is there that the Reading First program has improved the effectiveness of participating schools and districts?

5. Have any unintended consequences resulted from the implementation of the Reading First program?

The conceptual framework below provides an overview of the evaluation study design. As described in the conceptual framework, the Reading First data can be organized into three types: a) school and district characteristics; b) achievement data; and c) implementation data. The school and district characteristics are described later in this chapter, with data drawn from state databases, including the California Basic Educational Data System (CBEDS) file, the demographic sections of the California English Language Development Test (CELDT) and STAR files, and the LEA-level database compiled by C-TAC to capture LEA internal evaluation data. The achievement data consist of school-level California Standards Test

---

[2] EDS is a registered trademark of Electronic Data Systems. However, in the context of this document, EDS refers exclusively to Educational Data Systems, Inc.

(CST) scores in a performance level metric and a scale score metric, school-level standardized test scores (drawn from the California Achievement Test, CAT/6) in a percentile metric, and C-TAC End-of-Year (EOY) scores (eight subtests for kindergarten and Oral Fluency for Grades 1-3) for both English and Spanish. The implementation data will, as before, be drawn primarily from the teacher, coach, and principal surveys that are administered to all Reading First schools annually.

The conceptual framework indicates the types of analysis employed. The achievement data are analyzed according to the percentage of students in a school at a given performance level and average school performance level. An additional analysis yields the Reading First Achievement Index (RFAI), which combines the STAR and EOY data. To examine implementation, the Many-Facet Rasch models are used to combine the teacher, coach, and principal surveys into a coherent measurement framework. The variables used and the analyses have been conducted in accordance with recommendations of the Reading First EAG. Note that in the first 3 chapters of the Year 5 report, there will be no qualitative analyses of open-ended survey responses. We refer the reader to this analysis in the Year 4 report and to supplemental chapters to be published as an extension of this report in January 2008.

Figure 1.1: Conceptual Framework – Year 5

Data Extracted by EDS from CDE and C-TAC	Data Extracted by EDS from schools (RF and non-RF), LEAs, and data providers (e.g., OARS)

Waivered Classes

CBEDS (CDIF, PAIF, SIF)	C-TAC district data, prof. dev., coach data	CTAC End-of-Year (EOY) results (K-3)	STAR results (G2-4 CSTs, CAT/6)	Teacher/Coach/ Principal surveys	Open-ended Survey responses

School Characteristics	Achievement Data / Gain Scores	RF Implementation Variables

3-Facet Rasch Analysis, Factor Analysis

Preliminary Data Analysis

Derived Variables of Study

Independent Variables

- Student subgroups: English language learners, special education, socio-economically disadvantaged, ethnic/racial subgroups
- School characteristics: RF or non-RF, number of students, aggregated student characteristics, aggregated teacher characteristics, cohort membership, years in program
- Teacher characteristics: years of experience, credentials, educational level, ethnicity, area of authorization, full/part time
- School-level program implementation variables, subscale measures on 17 implementation dimensions plus RFII (Reading First Implementation Index)

Dependent Variables

- Performance level proportions on the G2-4 CSTs and CAT/6 scores, End-of-Year results, Reading First Achievement Index (RFAI), school gain scores organized by years in program, STAR results for waivered and EL non-waivered classrooms

Data Analysis

- Descriptive statistics (means, standard deviations, graphs, cross-tabs, correlations)
- Rasch IRT Models (to convert survey data into measures)
- Factor Analysis (to group implementation variables)
- Multivariate Regression (to quantify the effect of implementation on achievement, after controlling for other factors)

Reports

(By school, years in program, cohort, student subgroup, and classroom waivered status)

- School achievement over time on various scales
- Implementation over time
- Effect of implementation on achievement over time
- Demographic descriptions of Reading First cohorts
- Reading First program evaluation recommendations

Comparison Group

Past reports have included comparison groups against which to gauge the relative effects of the Reading First program. Past efforts included using "Reading First Eligible" schools, or those who would likely meet socio-economic and achievement criteria for Reading First if their LEA were included in the program. However, in the Year 3 report, it was demonstrated that these schools were too demographically dissimilar to Reading First schools to serve as a legitimate comparison group. The Year 4 report also discussed problems with creating a demographically matched group of schools due to differences in the starting place for their achievement as compared to Reading First schools. An additional difficulty with using comparison groups is the statewide effort to improve reading instruction in non-Reading First schools. It is likely that state-adopted curricula, state-funded professional development, and other elements of Reading First were present in non-Reading First schools, making it impossible to discern the true impact of the Reading First program. The reader is referred to the Year 4 report for a complete discussion of these difficulties. For this Year 5 report, no data are reported for non-Reading First comparison schools due to inherent difficulties in establishing adequate comparisons. However, analyses are conducted using a statistically derived comparison group, as described in the Year 4 report and in Chapter 2 of this report.

Demographic Characteristics of Reading First Schools

California's Reading First program began in the 2002-03 academic year. During each subsequent year except for 2005-06, additional LEAs were funded. The Year 4 report distinguished between cohort groupings based on the year the LEAs received funding and "Years in Program" (YIPs), for school-level analyses. A small number of schools included in Reading First databases do not have the same years of participation as their assigned LEA cohort, due to gaining and losing schools in cohorts for various reasons such as schools merging, closing, or replacing other schools dropped from the program. This is a relatively small number of schools, but for accuracy of school-level analyses, this report will use the YIP for achievement and implementation analyses in Chapters 2 and 3. For demographic analyses included in this chapter, we use LEA Cohorts to describe the characteristics of participants.

The following is a summary of the LEA cohorts, the typical YIP for that cohort, and the number of schools (a total of 886 in the 2006-07 academic year) from the cohort included in the current report:

(a) Cohort 1, first funded in 2002-03, with 13 LEAs (347 schools in current report); YIP 5

(b) Cohort 2, first funded in 2003-04, with 60 LEAs (372 schools in current report); YIP 4

(c) Cohort 3, first funded in 2004-05, with 27 LEAs (146 schools in current report); YIP 3

(d) Cohort 4, first funded in 2006-07, with 10 LEAs (21 schools in current report); YIP 1

The demographic data included in this chapter are extracted from the STAR research file published on the CDE website[3]. In the STAR file, student-level data have been aggregated and presented at the school level. Therefore, the smallest unit of analysis in this chapter is the school. Other sources of data include the Professional Assignment Information Form (PAIF) file, and the CBEDS file.

Socio-Economically Disadvantaged (SED) Students in Reading First

According to the Reading First legislation, funding is earmarked for schools in the state with high numbers of students of low socio-economic status and a history of low achievement. Therefore, it is not surprising that the Reading First schools have a higher number of SED students as compared to all elementary schools in the state. Table 1.1 displays the percentage of SED students in each cohort of Reading First and in all elementary schools in the state for each year of the program. It is evident that Cohort 1 had the highest percentage of SED students compared to other cohorts, with 92.2% in 2007. Cohorts 2 and 3 in 2007 both had 86.8% SED students. Cohort 4 had the lowest percentage of SED students, 73.4%.

English Learners (ELs)

In 2007, Reading First schools also had higher percentages of ELs than the category of All Elementary Schools. The percentage of ELs in Cohorts 1, 2 and 3 was 53.7%, 54.7% and 58.5% respectively. Cohort 4, with 31.2% ELs, more closely resembled the statewide figure of 29.5%.

Students with Disabilities

In 2007, the percentage of students with disabilities was reported as 8.3% for Cohort 1, 7.6% for Cohort 2, 6.7% for Cohort 3 and 7.9% for Cohort 4. This varies somewhat from the statewide percentage of 10.6%. It is interesting to note that for Cohort 1, the percentage has risen over time while the percentage has dropped slightly for Cohorts 2 and 3 since their participation in Reading First.

Ethnicity Breakdown of Reading First Schools

Table 1.1 shows the percentage of students in each ethnicity category for each cohort, by year as compared to statewide figures. As compared to the All Elementary Schools category, Reading First schools in general had significantly higher percentages of Hispanic students and significantly lower percentages of White students. Cohorts 1, 2 and 3 had significantly higher percentages of Hispanic students than Cohort 4. Additionally, it is evident that African American students were significantly over-represented in Cohort 1 compared to Cohorts 2, 3, and 4 and the All Elementary Schools category.

---

[3] The STAR research file used for the 2006-07 data was the version obtained by EDS on September 24, 2007, referred to as "P2."

Table 1.1: Student Demographic Data, 2003 to 2007

| | Reading First Schools | | | | | | | | | | | | | | All Elementary Schools[1] | | | | |
| | Cohort 1 | | | | | Cohort 2 | | | | Cohort 3 | | | Cohort 4 | | | | | | |
	2003	2004	2005	2006	2007	2004	2005	2006	2007	2005	2006	2007	2007	2003	2004	2005	2006	2007
Number of Schools	329	329	325	329	336	343	353	370	370	136	143	144	19	5823	5919	5977	5983	6057
SED (%)	91.3	92.7	89.1	91.4	92.2	82.7	86.7	83.5	86.8	85.1	85.8	86.8	73.4	51.0	51.6	53.3	53.4	54.0
EL (%)	58.5	58.6	58.8	56.5	53.7	53.0	55.5	54.9	54.7	57.5	57.2	58.5	31.2	27.2	28.2	29.3	29.3	29.5
Students with Disabilities (%)	7.5	8.4	8.5	8.4	8.3	8.0	7.7	7.9	7.6	7.1	7.7	6.7	7.9	9.8	11.0	11.1	10.8	10.6
African American (%)	17.2	16.6	15.4	14.1	13.7	8.8	8.2	8.0	7.8	6.6	6.4	6.3	14.1	7.8	7.8	7.6	7.8	7.7
American Indian (%)	0.3	0.3	0.3	0.3	0.3	1.0	0.9	0.8	0.7	0.8	0.6	0.8	7.7	1.3	1.3	1.3	1.4	1.4
Asian (%)	4.0	3.6	3.8	3.2	3.0	4.6	4.1	3.9	3.9	1.1	1.0	1.0	1.7	7.3	7.3	7.5	7.3	7.3
Filipino (%)	1.0	0.9	1.0	1.0	1.0	1.7	1.6	1.5	1.5	1.3	1.0	0.9	4.6	2.2	2.2	2.3	2.4	2.4
Hispanic (%)	71.5	73.3	74.4	76.5	77.2	72.0	74.1	75.2	76.1	77.1	78.6	79.7	50.5	40.2	41.5	42.6	43.6	44.1
Pacific Islander (%)	0.5	0.5	0.4	0.4	0.4	0.8	0.8	0.8	0.7	0.5	0.5	0.5	0.6	0.6	0.7	0.7	0.7	0.7
White (%)	3.7	3.4	3.2	2.8	2.7	9.6	8.8	8.0	7.5	11.2	10.3	9.4	19.3	36.5	35.2	33.9	33.0	32.3

[1]The group "All Elementary Schools" *includes* Reading First schools in this chapter. In Chapter 2, "All Non-Reading First Elementary Schools" *excludes* Reading First schools.

Data source: California Standardized Testing and Reporting (STAR) research file. The number of schools included on this table may differ from other tables because STAR data is obtained beginning with grade 2 and therefore does not include schools with enrollment only for grade K-1.

Urban-Rural Distribution

Table 1.2 presents the prevalence of urban and rural designations in the Reading First LEA cohorts and for all cohorts combined. In this table, it is evident that most of the LEAs in Cohort 1 were designated as large or mid-sized cities, while Cohort 2 included primarily large, mid-size and both large and mid-size fringe categories. Cohort 3 included mainly mid-size cities, urban fringe of large and mid-size cities and rural designations. Cohort 4 had the highest percentage of rural LEAs.

Table 1.2: Urban-Rural Distribution for Reading First Districts 2007

District Location	Cohort 1		Cohort 2		Cohort 3		Cohort 4		All Cohorts	
	N of Districts	Percent of Districts[1]	N of Districts	Percent of Districts	N of Districts	Percent of Districts	N of Districts	Percent of Districts	N of Districts	Percent of Districts
Large City	6	46.2	10	16.7	3	8.1	1	10.0	20	16. 7
Mid-size City	4	30.8	11	18.3	7	18.9	3	30.0	25	20.8
Urban Fringe of Large City	1	7.7	17	28.3	10	27.0	0	0.0	28	23.3
Urban Fringe of Mid-size City	1	7.7	14	23.3	9	24.3	1	10.0	25	20.8
Small Town	0	0.0	1	1.7	1	2.7	1	10.0	3	2.5
Rural	1	7.7	7	11.7	7	18.9	4	40.0	19	15.8
Total	13	100.0	60	100.0	37	100.0	10	100.0	120[2]	100.0

[1] The percent of the districts in that cohort in a particular type of location.
[2] There are 118 LEAs that are school districts and two that are independent charter schools.
Data source: National Center for Education Statistics (NCES)

Teacher Qualifications in Reading First Schools

Table 1.3 provides information about Reading First teachers' credentials and teaching experience as derived from the CBEDS and PAIF research files. This table shows the percentage of teachers falling into each educational degree category by cohort and year, as well as teachers' average years of experience.

The issue of teacher qualifications is an important one, given the focus of the NCLB on ensuring that schools are staffed with highly qualified teachers. Comparing cohorts, the teachers in Cohort 1 had somewhat lower percentages of advanced degrees than teachers in the other cohorts while Cohort 1 also had a higher proportion of teachers with bachelors degrees only. Examining the percent of teachers who were fully credentialed in each cohort, it is interesting to examine the changes over time in the percentages of fully credentialed teachers at Reading First schools. Cohort 1 had the greatest gain, moving from 77.8% to 95.55% in five years.

To more easily compare cohorts to each other, a weighted index was computed based on CBEDS data sources relative to teacher qualifications. The weighted teacher qualification is an index ranging from a low teacher qualification of 1 to a high teacher qualification of 5. Table 1.3 shows that Cohort 1 Reading First schools had lower Weighted Teacher Qualification indices (2.01 to 2.24) than the other cohorts (ranging from 2.24 to 2.36) and the non-Reading First schools.

Table 1.3: Elementary Teacher Credentials and Experience 2003 – 2007

	Reading First Schools														All Elementary Schools[2]				
	Cohort 1					Cohort 2				Cohort 3			Cohort 4						
	2003	2004	2005	2006	2007	2004	2005	2006	2007	2005	2006	2007	2007	2003	2004	2005	2006	2007
Number of Schools	329	329	325	329	347	359	344	370	372	135	143	146	20	5647	5694	5720	5837	5917
PhDs (%)	.55	.74	.75	.78	.71	.66	.66	.65	.70	.59	.51	.62	1.75	.90	.80	.80	.71	.69
Masters plus 30 or more semester units (%)	9.39	11.69	12.62	14.20	16.03	13.73	13.57	13.43	14.42	16.28	14.31	15.86	14.06	14.00	14.50	14.30	14.09	15.52
Masters (%)	10.95	11.80	12.23	12.43	13.02	16.86	18.53	20.01	20.18	16.63	13.55	13.77	17.40	15.50	16.90	18.10	18.97	19.32
Bachelors plus 30 or more semester units (%)	41.25	44.22	45.26	45.67	47.30	49.36	49.58	48.98	49.96	47.05	53.04	53.55	52.09	51.30	51.70	50.60	49.77	50.27
Total Advanced Degrees (%)	**62.14**	**68.45**	**70.87**	**73.07**	**77.06**	**80.61**	**82.35**	**83.06**	**85.25**	**80.56**	**81.41**	**83.81**	**85.30**	**81.70**	**83.90**	**83.80**	**83.54**	**85.80**
Bachelors (%)	35.06	30.92	28.38	26.37	22.74	19.30	16.89	16.82	14.67	19.33	18.34	16.11	14.55	16.40	15.80	15.90	16.22	14.05
Less than Bachelors (%)	.74	.58	.81	.53	.09	.10	.82	.08	.06	.22	.17	.08	.15	.20	.20	.40	.20	.10
Total Bachelors or less (%)	**35.80**	**31.50**	**29.19**	**26.90**	**22.83**	**19.40**	**17.71**	**16.90**	**14.74**	**19.55**	**18.52**	**16.19**	**14.69**	**16.60**	**16.00**	**16.30**	**16.42**	**14.15**
Weighted Teacher Qualification[1]	2.01	2.05	2.10	2.15	2.24	2.26	2.29	2.31	2.36	2.31	2.24	2.31	2.36	2.20	2.30	2.32	2.32	2.38
Fully Credentialed Teachers (%)	77.80	82.12	91.29	94.85	95.55	93.73	96.00	97.22	97.57	92.05	93.77	95.35	98.03	90.90	93.70	95.80	96.55	97.20
Average years teaching	10.75	10.91	11.27	11.47	11.57	11.25	11.60	11.87	11.96	11.40	11.88	11.95	13.06	12.70	12.80	12.80	12.89	13.00

[1]The Weighted Teacher Qualification is computed as follows: The percentage of teachers with PhDs is given a weight of 5; the percentage of teachers with Masters plus 30 or more semester units is given a weight of 4; the percentage of teachers with Masters is given a weight of 3; the percentage of teachers with Bachelors plus 30 or more semester units is given a weight of 2; and the percentage of teachers with Bachelors is given a weight of 1. The weighted degree percentages are summed, and then divided by 100, to reach the Weighted Teacher Qualification. This index spans from 1 (lowest qualification) to 5 (highest qualification).
[2]In this chapter, the group "All Elementary Schools" *includes* Reading First schools. In Chapter 2, "All Non-Reading First Elementary Schools" *excludes* Reading First schools.
Data source: California Basic Educational Data System (CBEDS) file.

Conclusions

This chapter yields the following:

- For this Year 5 report, no data are reported for comparison schools due to inherent difficulties in establishing adequate comparisons; however, a statistically derived comparison group is used in the achievement analyses in chapter 2.

- The term "Cohorts" refers to the year a Reading First LEA (district) accepted funding. The term "Years in Program," (YIP), indicates the number of years a school within an LEA cohort has actually been implementing the program. For demographic analyses, this report uses cohorts. For achievement and implementation analyses, this report uses YIPs.

- Cohort 1 had the highest percentage of socio-economic disadvantage (SED) students at 92.2% in 2007 demographic files. Other cohorts ranged from 73.4% to 86.8%. The figure for All Elementary Schools was 54.0%.

- Reading First schools had higher percentages of ELs than the figure for All Elementary Schools (29.5%). Percentages of ELs in cohorts ranged from 31.2% to 58.5%.

- Reading First schools had higher percentages of Hispanic students and lower percentages of White students than the All Elementary Schools category.

- Cohorts 1, 2 and 3 had significantly higher percentages of Hispanic students than Cohort 4. Additionally, African American students were significantly over-represented in Cohort 1 compared to the other cohorts and the All Elementary Schools category.

- Most of the LEAs in Cohort 1 were designated as serving large or mid-sized cities, while Cohort 2 ranged from large to mid-size fringe categories. Cohort 3 included mainly mid-size cities, urban fringe of large and mid-size cities and rural designations. Cohort 4 had the highest percentage of rural LEAs.

- Schools participating in Reading First for two or more years have steadily increased their percentage of teachers with full credentials. Cohort 4, which entered the program in the 2006-07 school year, entered the program with a high percentage of fully credentialed teachers.

- Using a weighted teacher qualification index based on 2006-07 CBEDS data, Cohort 1 Reading First schools had lower weighted teacher qualification indices than the other cohorts and the All Elementary Schools category.

- In 2007, all cohorts had more than 95% of their teachers fully credentialed.

Chapter 2: Achievement

This chapter addresses the questions: What is the impact of the Reading First program on K-3 students in participating districts and schools? What evidence is there that the Reading First program has improved the effectiveness of participating schools and districts? The chapter also addresses a *new* question: To what extent does participation in the K-3 Reading First program improve student achievement in grade 4? The key findings in this chapter are:

- After controlling for school demographic characteristics, Reading First implementation is a statistically significant predictor of achievement on all achievement metrics, especially those associated with grades 2 and 4, at the 95% confidence level. The more faithfully the program is applied, the greater the effect on achievement.

- The Reading First Achievement Index (RFAI), a composite of K-3 achievement metrics for Reading First schools, has risen an average of 3.4 points per year, equivalent to 17 points over 5 years.

- Reading First schools out-perform a statistical control group by 1.6 points per year on the RFAI, equivalent to an 8-point advantage over 5 years.

- Since 2002, Reading First schools have shown significantly more growth than either non-Reading First schools or the statistical control group.

- Reading First effects generalize to all performance levels of the Reading First student population and to the student population as a whole. On the California Standards Test (CST) metrics, the migration of students into "Proficient & Above" is matched by a comparable migration of students out of "Below and Far Below Basic." These migrations are confirmed by average student CST scale score gains on the order of 20 scale score points over a 5-year period.

- These findings are replicated in grade 4. Reading First schools grew 4.1 CST scale score points per year (20.5 scale score points over 5 years) in grade 4, versus 2.4 scale score points per year (12.0 points over 5 years) for the control group, a difference of 8.5 scale score points. Thus the program effect is sustained in grades to which the program is not administered.

Achievement results for Reading First schools are presented in terms of the Standardized Testing and Reporting (STAR) Program assessments – the California Standards Test (CST) and the California Achievement Test (CAT/6) – and the Reading First End-of-Year (EOY) curriculum-embedded assessments. As of this report, grade 4 CST results are included to assess the sustained effects of Reading First. Achievement is compared in four ways:

1. between years (gain scores)

2. between Reading First and non-Reading First schools

3. between Reading First schools and a statistical control group

4. between high implementation and low implementation Reading First schools[1]

The objective of this evaluation is to determine whether or not, and to what degree, the Reading First program is effective. What is meant by "effective"? According to the federal guidelines for Reading First, the program is effective to the degree it ensures "that every student can read at grade level or above not later than the end of Grade 3" (U.S. Department of Education, 2002). There are several ways to examine the effect of Reading First on reading in California given the limitations of a non-experimental design.

1. Measure the size of the achievement gains of the Reading First schools for grade 3 and other grades that are related to grade 3, such as grade 2 and grade 4

2. Compare Reading First schools to comparable non-Reading First schools

3. Compare Reading First schools to a "statistical control group" by using statistical methods to profile how a school that is similar to Reading First schools would perform without access to the program

4. Compare high implementation Reading First schools to low implementation Reading First schools

The first approach looks at the absolute size of the achievement gains of Reading First schools from just before they started, and implementation had not yet occurred, to the present, when the program has been in place and is presumably well implemented. A significant positive gain would suggest the Reading First program is working. However, it is difficult to rule out the possibility that such gains are the effect of other causal factors that came into play over the same time period, especially factors that may cause all schools to show an increase or decrease in scores.

The second approach, comparing Reading First schools to comparable non-Reading First schools, was discontinued in Year 4 of the evaluation for reasons that are explained in Chapter 2 of that report.[2] Given the constraints of the study, it is not possible to identify non-Reading First schools that are not to some

---

[1] A detailed discussion of Reading First program implementation as embodied in the Reading First Implementation Index (RFII), an implementation statistic computed using responses to surveys administered to teachers, coaches, and principals in every Reading First school, is deferred to Chapter 3 of this report.
[2] See Chapter 2 of The California Reading First Year 4 Evaluation Report, available at: *www.eddata.com/resources/publications/.*

degree employing the same program elements that are required of Reading First schools, making comparisons between them problematic.

The statistical control group approach employed in the Year 4 and Year 5 Reports uses multiple regression to calculate the achievement gains that would be expected of schools that are similar to Reading First schools but that do not implement the Reading First program. This approach relies on the existence of a school implementation measure, the Reading First Implementation Index (RFII) described in detail in Chapter 3. Mathematical in nature, the RFII is based on a calculated relationship between implementation and achievement, which is used to extrapolate the performance of "non-implementing" schools, even though these do not exist *per se*.

The fourth approach is statistically similar to the third, but it entails comparing a sample of Reading First "low implementing" schools with a sample of Reading First "high implementing" schools.

Based on these four approaches, Reading First will be said to show evidence of being effective to the degree that:

1. Achievement gains in Reading First schools are positive for grades 2, 3, and 4.

2. Reading First schools show higher achievement gains than all non-Reading First schools for grades 2, 3, and 4.

3. Reading First schools show higher achievement gains than what would be predicted from a statistical control group for grades 2, 3, and 4.

4. High Implementing Reading First schools show higher achievement gains than Low Implementing Reading First schools for grades 2, 3, and 4.

Measures of Achievement Gains

School progress or growth, also called achievement gains, is measured using the CSTs, the CAT/6 Mean Percentile Ranks (called here "MeanPR"), the Reading First End-of-Year (EOY) tests, and the Reading First Achievement Index (RFAI), which is a composite of the others and is used to make decisions about continued Reading First funding for LEAs. Each metric has unique characteristics described below.

The California Standards Test (CSTs). The CSTs are administered to all California students in grades 2 and above toward the end of the school year. We use the English language arts (ELA) component of the CSTs for grades 2, 3, and 4. The inclusion of grade 4 commences with the Year 5 Report because students in grade 4 can be expected to have experienced Reading First since kindergarten. Within ELA, we study the percentage of students per school that fall within each of the two following performance categories, which are a simplification of the five CST performance categories (Advanced, Proficient,

Basic, Below Basic, Far Below Basic). We also study the average CST scale score of the students in those grades.

1. "Proficient and Above" means the percentage of students in a school that are in the Proficient and Advanced performance categories. This is the primary metric for measuring growth that is used for accountability purposes under NCLB.

2. "Below and Far Below Basic" means the percentage of students in a school that score in the bottom two performance categories. It is just as important to measure growth out of the bottom categories, as it is to measure growth into the top categories, making it possible to assess whether Reading First is effective for low-scoring students.[3] A negative change in the percent of students testing "Below and Far Below Basic" means that students are exiting that performance level and moving to higher performance levels. Thus, a negative "gain" in this context means that growth is occurring.

3. "Mean Scale Score" refers to the average CST score of the students in the grade. A scale score is a number ranging from approximately 200 to 500, which describes a student's performance on a test in a way that facilitates valid comparisons. Using scale scores (with equal intervals) to measure growth reduces anomalies due to statistical artifacts caused by unequal intervals between values. Mean scale scores have not been provided in previous evaluation reports due to the NCLB emphasis on percentage of students in a performance category and to a desire to use reporting metrics with which the public and legislators are likely to be familiar. We introduce them in the Year 5 Report to address possible misinterpretation that growth is limited only to those students who move into the "Proficient and Above" category from below, or out of the "Below and Far Below Basic" category. This confusion may have led to the perception that the rest of the students who do not change categories somehow have not grown, and that Reading First has not affected them. The mean scale score metric makes it clear that growth caused by Reading First is pervasive across the Reading First student population.

The CST gain score reported in the tables of this chapter is the 2007 percentage of students in a specified category minus the corresponding percentage in the year immediately *preceding* the first year of Reading First funding. The change in scale scores is calculated using the same time frame. The gain scores are

---

[3] The "Basic" category has been discontinued in the Year 5 Evaluation Report because change in the percentage of students scoring in this category is not interpretable. For instance, if a large migration of students into "Proficient and Above" is exactly matched by an exodus of students out of "Below and Far Below Basic," the net change in the "Basic" category would be zero, a phenomenon that has in fact been observed in previous reports. This could lead to the erroneous conclusion that Reading First has no effect on students in the "Basic" category, when in fact it has a large effect. Change in this category can also yield a false finding of Reading First effectiveness.

averaged across a specified population of schools to produce the tabular statistics presented in this chapter.

CAT/6 MeanPR. As of the spring 2005 administration of the California STAR assessment, the CAT/6 component was discontinued in all elementary grades except for grade 3, so only grade 3 CAT/6 Reading, Language Arts, and Spelling data are used in this study. The "MeanPR" of a school is the average of the National Percentile Rank (NPR) scores of each of its students. The National Percentile Rank tells what percentage of students nationwide is expected to score below the student with a given NPR. An NPR of 45 would mean that the student is likely to score better than 45% of the national student population who take the tests. The MeanPR gain score for each school is its MeanPR in 2007 minus its MeanPR in the year immediately *preceding* its first year of Reading First implementation. The CAT/6 gain scores reported in the tables of this chapter are an average of these MeanPR gain scores across a specified sample of schools. Note that they are interpreted as a change in national percentile ranking, not as a change in the percentage of students meeting some benchmark or performance standard.

End-of-Year (EOY) Test. As the name suggests, the EOY is a curriculum-based test administered by all Reading First schools to students in grades K-3 at the end of the academic year. The kindergarten EOY test consists of eight subtests: Consonants, Lower Case Letters, Phonics, Rhyming, Syllables, Upper Case Letters, Vowels, and Consonant-Vowel-Consonant. The EOY tests for grades 1, 2 and 3 consist of a timed oral reading in which fluency is measured in terms of words correct per minute. The EOY is unique and valuable for this study because it is the only test that can be used to measure achievement in kindergarten and grade 1. It is also the only test used in this evaluation that is administered in Spanish to students in "waivered" Reading First classrooms (in which instruction is conducted in Spanish). The EOY score for each grade within a school consists of the percentage of students that meet the benchmark established for that grade based on national norms recommended by Hasbrouck & Tindal (2005). The gain score for that grade is its 2007 EOY score minus its EOY score at the end of the *first* year of Reading First funding (not the year previous), which for schools in the program 4 or 5 years is 2004. For schools in the program 3 years, it is 2005.

Reading First Achievement Index (RFAI). The RFAI is a weighted combination of school-level percentages of students meeting various performance levels and benchmarks drawn from the CSTs, the CAT/6 Mean PR, and the EOY, with the heaviest weights placed on the CSTs. Refer to Appendix E for a detailed explanation of how the RFAI is computed. The RFAI was first computed in 2004. As of this study YIP 5 has four years of RFAI data (2004, 2005, 2006, 2007), as compared to six years of data for the CSTs. That is because the RFAI was not available in 2002 or 2003. Like the CST, each school RFAI can be interpreted as a percentage of students meeting a set of combined benchmarks and performance

levels. Because the RFAI is not based on a single benchmark or performance level, it is not interpretable as a single percentage. The RFAI gain score for each school is its 2007 RFAI minus its RFAI at the end of its *first* year of Reading First implementation.

Grouping of Schools by "Years in Program" (YIP)

Starting with the Year 4 report, for analyses of achievement schools have been grouped by Years in Program (YIP) rather than LEA funding cohort. As explained in the Year 4 Report, there are cases where LEAs that received funding starting in one year added schools to Reading First in a later year. For purpose of measuring program effects, it was deemed necessary to group schools according to the actual year in which they started implementing the program rather than by the funding cohort of their LEA.

It is often found in educational research that intervention program effects often vary over time and across cohorts. There are also changes in the behavior of tests over the years, which would influence the YIPs differentially. In the case of Reading First, both the YIPs and the achievement metrics have different characteristics depending on starting year. YIP 5 is notably more urban than YIP 4 and has had different rates of implementation. The grade 3 achievement metric experienced a statewide dip in 2004 which yields qualitatively different trend-lines for YIPs that started before the dip compared to those that started after.

In 2007, we focus on just those Reading First schools that have been in the program for 5 years (the longest), 4 years, and 3 years (YIPs 5, 4, and 3). We have omitted schools in YIPs 1 and 2. They have relatively few schools (92 combined) and it has been established in previous reports that implementation tends to be relatively weak for many schools until the second or third year of the program.

Because the various achievement metrics did not all become available at the same time, the baselines for the achievement metrics vary. The CST metrics take 2002 (the year previous to implementation) as their baseline, whereas the EOY and RFAI take 2004 as their baseline. Each achievement gain takes the earliest year for which that achievement metric was available for that YIP. There is an additional complication relating to the baseline year for the Spanish version of the EOY test, which only became available in 2005. For more details about the relationship between the Reading First YIPs and the various achievement metrics, see Chapter 4 of the Year 4 Report.

Comparison of Reading First to Non-Reading First Schools

Prior to the Year 4 Report, efforts were made to identify a sample of non-Reading First schools that would be comparable to the Reading First population and yet not contain Reading First-style program elements. These efforts were abandoned in Year 4 as it became increasingly clear that there was no way to control for the increasing similarity between the two groups of schools as regards their use of state-

adopted reading programs, common professional development resources, and use of reading coaches. In place of a sample of comparable non-Reading First schools, we instituted the concept of the "statistical control group," described in detail below. Nonetheless, we continue to report on the gains of the non-Reading First elementary school population in California in order to provide an overview of the rest of the state and show how it has been trending since 2002. This provides an essential context for studying the Reading First gains, for we see that the Reading First upward trend is mirrored in the rest of the state. However, it is emphasized that the non-Reading First group is demographically dissimilar to the Reading First group and that caution should be exercised when comparing them.[4]

Comparison of High Implementation and Low Implementation Reading First Schools

One defining characteristic of this evaluation is that Reading First is studied not only in terms of student achievement but also in terms of program implementation at the school level. Chapter 3 and Appendices A, B, and C describe the teacher, coach, and principal surveys that were administered in all Reading First schools and used to compute a Reading First Implementation Index (RFII) statistic for each school with sufficient respondents. The RFII is intended to measure the degree to which the teachers, coaches, and principals are implementing the Reading First program in their school. RFII measures have been computed for 2004, 2005, 2006, and 2007 based on a survey administration in the spring of each year.

The RFII was used to divide Reading First schools into two groups labeled High Implementation Schools and Low Implementation Schools. For the Year 4 Report and those preceding, a high implementation school was defined as a school whose average RFII since entering the program is greater than or equal to 36.0, the average RFII in 2004. A low implementation school had an average yearly RFII less than 36.0. Based on advice from the Evaluation Advisory Group (EAG), the definitions were changed for the Year 5 Report. Now we define a high implementation school as one whose average yearly RFII is greater than 1 standard deviation above the original 36.0 cut-point, approximately 41.4. A low implementation school continues to be one whose average yearly RFII is less than 36.0.[5] This change has the effect of introducing a more stringent definition of high implementation, and also of leaving out the schools between 36.0 and 41.4 from the high and low groups. (They continue to be represented in the "All Reading First schools" category.) Therefore, the number of high implementation schools in 2007 is not comparable to that in 2006 or earlier.

---

[4] In the trend-line charts presented later in this chapter, the All Non-Reading First Elementary Schools group (which has a much higher starting point than the Reading First schools) is adjusted to have the same starting point as the Reading First schools so that their trend-lines can more conveniently be compared.

[5] An EAG recommendation to define "low implementing" schools as those with an RFII more than one standard deviation below the mean was not implemented because it was found that this yielded a very small number of low implementing schools, not sufficient for statistical comparisons.

It may be wondered why we used the "average yearly RFII" in defining the high and low groups. In trying to explain total gain in achievement since a school started in the program, since each year's implementation contributes to the total achievement gain score for a school, we need to take into account each year's implementation (RFII). Therefore, we sum the RFIIs across all the years the school has been in the program and divide by the number of years to come up with an average yearly RFII.

Nonetheless, acting on advice of the EAG, when reporting an individual RFII for each school we average its preliminary RFII (computed from the 2007 surveys) and its 2006 RFII on the theory that a rolling 2-year average is more stable and reliable than the RFII computed from a single year's worth of data.

Calculating Achievement for the Statistical Control Group

As discussed in chapters 2, 3 and 4 of the Year 4 Report, the statistical control group is defined using regression models to calculate the 2007 achievement score that a school which is similar to the Reading First schools (the same demographic and starting characteristics as the Reading First YIP under consideration) *would* obtain if it were *not* implementing the program. For reasons described in Chapter 3 of the Year 4 Report, we chose an RFII of 25 to signify a school that is not implementing the program. Thus, 25 is entered into the regression equation to calculate an expected 2007 achievement score and gain score for the statistical control group. As stated previously, the statistical control group is not a literal group of schools but an extrapolation based on a relationship between achievement and implementation derived statistically from the Reading First schools. (Non-Reading First schools could not be used to compute this relationship since they do not take the surveys and do not receive an RFII.) The detailed procedure for computing the statistical control group achievement statistics is described in Chapter 4 of the Year 4 Report.

Achievement Results

The following pages present a series of tables and trend-line charts showing starting scores, ending (2007) scores, and gains on each of 12 achievement metrics. They are the heart of the Year 5 Report and the basis of our finding that Reading First is an effective program. Table 2.1 summarizes the gains of *all* Reading First schools taken as a whole, not broken out by Years in Program (YIP). Presenting gains of schools that have been in the program differing lengths of time, this table compares them using an "average yearly achievement gain" metric. This metric differs from the metric in the YIP-specific tables, which report *total* achievement gain since the starting year.

Table 2.2 reports total RFAI gains broken out for YIPs 3, 4, and 5.

Tables 2.3, 2.4, and 2.5, with accompanying trend-line charts, show total gains on the CST and CAT/6 metrics for YIP 5, Grades 2, 3, and 4. Similar tables and charts for YIPs 3 and 4 are available in Appendix D.

Before presenting the achievement results, we touch on two points that may prove useful in interpreting the data in the tables:

1. Interpreting Significance Tests. The statistics in the achievement tables provided in this chapter are sometimes accompanied by superscripts "a", "b", and "c." These refer to tests for statistical significance. Significance tests answer the question, "How likely is it that the observed difference would have occurred by chance?" As noted below each table, the superscript "a" means that the group in question (the one with the superscript) has a gain score that is "significantly" higher than that of the Statistical Control Group at the 95% confidence level, which means that the probability of the difference occurring by chance is less than 0.05 (i.e., $p < 0.05$). The "b" means the group is significantly higher than the "All Elementary Schools" group. The "c" means the new group average is significantly higher than where it started from, i.e., that the change is significantly larger than zero. Three pieces of information go into a significance test: the difference *between* groups, the amount of variation *within* each group, and the *number* of schools within each group. A large difference between groups with little variation within each group and a large number of schools within each group will be more likely to yield a "statistically significant" difference.

2. Rounding Errors. Sometimes we report a gain score that does not appear to equal the difference between the starting score and the ending score for a given metric that may be off a decimal value. The explanation is that the reported starting and ending scores have been rounded to one decimal place, whereas the reported difference or gain was computed at more than 8 decimal places. Thus the reported gain is (slightly) more accurate than the difference between the reported starting and ending scores.

Summary Gains (Table 2.1)

Table 2.1, reports average yearly gains for all Reading First and non-Reading First schools across all the YIPs (Years in Program) for each achievement metric. As such, it summarizes all the primary findings of the Year 5 report and answers the question, "What has been the effect of Reading First on all schools currently in the program?" Because it combines all five YIPs in one set of statistics, it does not report starting scores and ending scores since these naturally differ for each YIP. For the same reason, it is not accompanied by a trend-line chart.

Table 2.1: Summary Gains, All YIPs Combined, All Grades, Mean Yearly Gain

All YIPs Combined All Grades Mean Yearly Gain (Average Change Per Year)	Reading First Schools				All Non-Reading First Elementary Schools
	All Reading First Schools	High Implementation Schools (Avg. RFII > 41.4)	Low Implementation Schools (Avg. RFII < 36.0)	Statistical Control Group (RFII = 25.0)	
Grade 2, CSTs	(N=831)	(N=137)	(N=295)	(N=N/A)	(N=4053)
% Proficient and Above	3.8[abc]	4.3[abc]	3.5[abc]	2.8	2.9
% Below and Far Below Basic	-3.3[abc]	-3.9[abc]	-3.0[abc]	-2.2	-1.5
Scale Score Metric	4.5[abc]	5.1[abc]	4.1[abc]	3.1	3.5
Grade 3, CSTs	(N=832)	(N=138)	(N=296)	(N=N/A)	(N=4048)
% Proficient and Above	1.6[abc]	1.8[bc]	1.4[bc]	1.4	0.2
% Below and Far Below Basic	-2.8[abc]	-2.9[abc]	-2.7[bc]	-2.3	-0.8
Scale Score Metric	2.9[bc]	3.1[bc]	2.7[bc]	2.6	0.5
Grade 3, CAT/6, Mean Percentile Rank	(N=832)	(N=138)	(N=296)	(N=N/A)	(N=4045)
Reading, Mean PR Metric	1.0[abc]	1.2[abc]	1.0[abc]	0.7	0.1
Language, Mean PR Metric	1.2[abc]	1.3[bc]	1.1[bc]	1.0	0.5
Spelling, Mean PR Metric	2.4[abc]	2.8[abc]	2.2[abc]	1.7	1.0
Grade 4, CSTs	(N=255)[1]	(N=26)	(N=101)	(N=N/A)	(N=3992)
% Proficient and Above	3.2[abc]	4.1[abc]	2.8[bc]	2.0	2.7
% Below and Far Below Basic	-3.1[bc]	-4.0[abc]	-2.7[bc]	-2.3	-1.3
Scale Score Metric	4.1[abc]	5.2[abc]	3.6[bc]	2.4	3.7
Reading First Achievement Index	(N=826)	(N=135)	(N=293)	(N=N/A)	-
RFAI Metric	3.4[abc]	3.6[abc]	3.4[abc]	1.8	-

[a] Significantly different (p < 0.05) relative to the "Statistical Control Group."

[b] Significantly different (p < 0.05) relative to "All Non-Reading First Elementary Schools."

[c] Significantly different (p < 0.05) relative to the starting year, i.e., significantly different from a gain of zero.

[1] The grade 4 sample includes only YIP 5 schools, hence the much smaller N.

These statistics report the average difference between a school's starting score, in the year previous to entry into Reading First (except for the RFAI, which started in 2004 and is relative to the first implementation year), and its ending year in 2007, divided by the number of years it has been in the program. Thus it is the average growth per year on a variety of metrics. Because these statistics reflect *average* yearly gains rather than *total* gains, they are smaller than the statistics reported in Tables 2.2 – 2.5. Multiply by 5 to get a 5-year expected gain.

The story is consistent. Growth is substantial in grades 2 and 4, more modest in grade 3. Reading First schools grow faster than the statistical control group and the other elementary schools in the state. High implementing schools grow faster than low implementing schools.

Note that the number of schools in each grade is not necessarily the same. This reflects the fact that not all schools teach the same grades or have complete data. The grade 4 number of schools reflects the fact

that grade 4 data were collected only for YIP 5 schools. The N's of the high and low implementing schools do not necessarily add up to the N of all implementing schools because many schools have RFII statistics higher than 36.0 and less than 41.4 and don't fall in either the "low" or "high" category.

RFAI Gains (Table 2.2)

Table 2.2 reports starting points, ending points, and total RFAI gains for YIP 3, 4, and 5 schools, starting with 2004 (the first year the RFAI was computed) or from the first year of Reading First implementation. Because the RFAI is only administered to Reading First schools, there are no comparable statistics for non-Reading First schools.

Table 2.2: RFAI Gains, YIPs 3, 4 and 5

	Reading First Schools			
	All Reading First Schools	High Implementation Schools	Low Implementation Schools	Statistical Control Group
Year in Program: 5				
Number of Schools	261	28	102	N/A
2004	36.4	38.1	35.4	36.4
2007	45.4	48.5	44.4	44.9
RFAI Gain	9.0[c]	10.4[c]	8.9[c]	8.5
Year in Program: 4				
Number of Schools	371	75	119	N/A
2004	34.5	35.7	33.2	34.4
2007	44.9	46.9	42.7	42.3
RFAI Gain	10.4[ac]	11.2[ac]	9.5[ac]	8.0
Year in Program: 3				
Number of Schools	151	26	57	N/A
2005	34.4	37.0	31.0	34.4
2007	42.8	44.6	40.2	40.7
RFAI Gain	8.4[ac]	7.5[c]	9.3[ac]	6.4

[a] Significantly different (p < 0.05) relative to the "Statistical Control Group."
[c] Significantly different (p < 0.05) relative to the starting year, i.e., significantly different from a gain of zero.

The RFAI gains in general support the hypothesis that Reading First schools are growing, that they grow more quickly than the statistical control group, and that high implementing schools grow faster than low implementing schools (though not all differences are statistically significant). The schools in YIP 3 offer an exception to the pattern, however. High implementing schools show a smaller gain than low implementing schools. In considering this exception and the differences that are not statistically significant, it is worth bearing in mind that the high implementing schools had a substantially higher starting RFAI (37.0) than the low implementing schools, which might have depressed their growth. It is

also worth bearing in mind that the RFAI statistic is 45% composed of data from grade 3 which, as is discussed later in this chapter, has a more complex relationship with implementation and years in program than grade 2 does.

CST Results for Grade 2 (Table 2.3 and Figures 2.3a – 2.3c)

Table 2.3 reports the starting and ending grade 2 CST scores of students in schools that have been in the program five years.

Table 2.3: CST Metric, YIP = 5, Grade = 2

Years in Program: 5 Grade: 2	Reading First Schools				All Non-Reading First Elementary Schools
	All Reading First Schools	High Implementation Schools (Avg. RFII > 41.4)	Low Implementation Schools (Avg. RFII < 36.0)	Statistical Control Group (RFII = 25.0)	
Number of Schools	259	28	101	N/A	4,053
% Proficient and Above					
2002	15.4	14.8	14.8	15.4	37.8
2007	34.2	36.7	33.0	30.4	52.3
Change Since Starting Year	**18.9**[abc]	**22.0**[abc]	**18.2**[abc]	**15.0**	**14.5**
% Below and Far Below Basic					
2002	54.3	53.8	55.6	54.3	30.5
2007	36.7	33.6	38.7	41.1	23.0
Change Since Starting Year	**-17.6**[abc]	**-20.2**[abc]	**-16.9**[abc]	**-13.2**	**-7.6**
Mean Scale Score					
2002	299.8	299.5	298.5	299.8	333.4
2007	324.7	328.6	322.3	318.8	350.9
Change Since Starting Year	**25.0**[abc]	**29.0**[abc]	**23.8**[abc]	**19.0**	**17.5**

[a] Significantly different ($p < 0.05$) relative to the "Statistical Control Group."

[b] Significantly different ($p < 0.05$) relative to "All Non-Reading First Elementary Schools."

[c] Significantly different ($p < 0.05$) relative to the starting year, i.e., significantly different from a gain of zero.

The gains in "% Proficient and Above" have risen from 15.7 percentage points in the Year 4 Report to 18.9 percentage points in Year 5, continuing a strong growth trend, although the gain is somewhat less from 2006 to 2007 than it was in previous years, or five per year. The growth rate is equivalent to a gain of 25 scale score points on the grade 2 CST over five years. Consider that the scale score difference between "Basic" (which starts at 300) and "Proficient" (which starts at 350) is 50 scale score points and that the CSTs range from approximately 200 to 500. If Reading First schools continue their current growth trajectory, they will have moved one whole performance level in 10 years, from 300 to 350, from the average student scoring "Basic" to the average student scoring "Proficient." This trajectory is more remarkable when one remembers that this gain is at the school level, with new students entering

kindergarten each year. Since each student cohort can be assumed to start at roughly the same average level of ability in kindergarten, one could interpret this rate of growth to mean that Reading First schools are now entering each new cohort of students 25 scale score points further up the scale in three years than they were, with similar cohorts, five years ago. Relative to the ordinarily slow pace of school improvement, and in light of the fact that this average comprises more than 30,000 students in YIP 5 alone, the pace of change is considerable.

We see that the rest of the state's elementary schools have also shown significant growth, but they lag behind Reading First schools by 7.5 scale score points. We also see that lower performing students are moving out of the bottom performance levels at the same rate that mid-range students are moving into the top two performance levels, a pattern not reproduced in non-Reading First schools. Students in non-Reading First schools exit the lower categories at almost half the rate that students enter the top categories. This is a key and important difference between Reading First and non-Reading First schools, one that holds up even in light of the fact that the two groups of schools are not ordinarily comparable.

We also see that gains for high implementing schools are 4 scale score points higher than for Reading First schools as a whole, more than 5 scale score points higher than for low implementing schools. This demonstrates that fidelity of implementation makes a measurable improvement in achievement.

Figures 2.3a, 2.3b, and 2.3c make the same points graphically by showing how school groups with differing degrees of participation in the program (i.e., implementation) experience different rates of growth.

Note that the "non-Reading First Schools" trend-line has been adjusted downward to have the same starting point as "All Reading First Schools" to make it easier to compare their trend-lines.

Figure 2.3a: CST % Proficient & Above, YIP = 5, Grade = 2
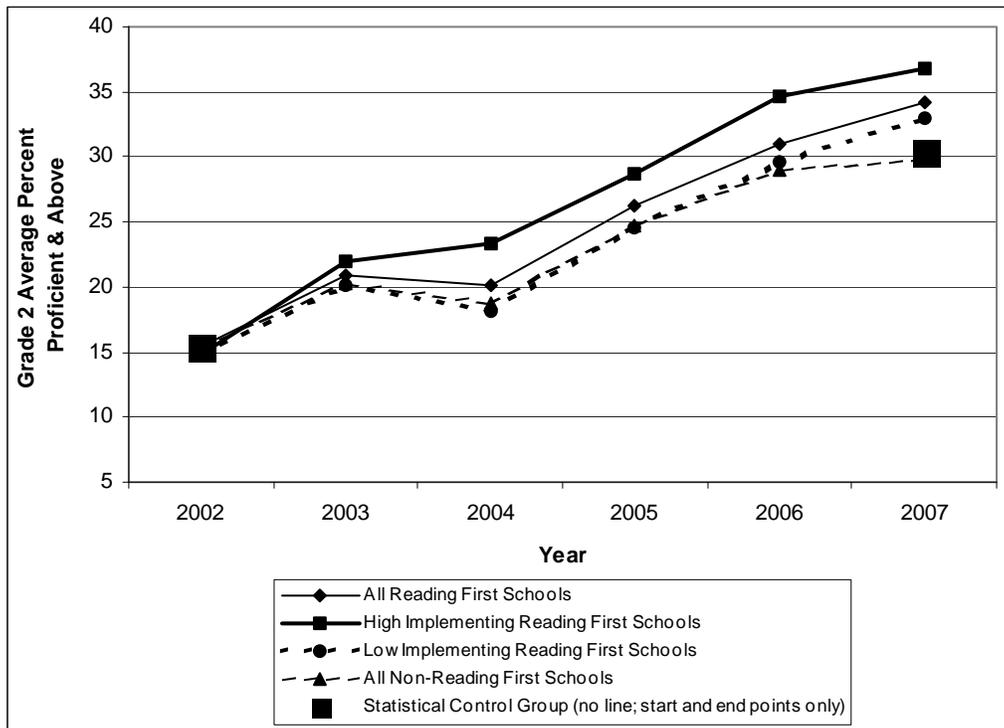
Figure 2.3b: CST % Below and Far Below Basic, YIP = 5, Grade = 2
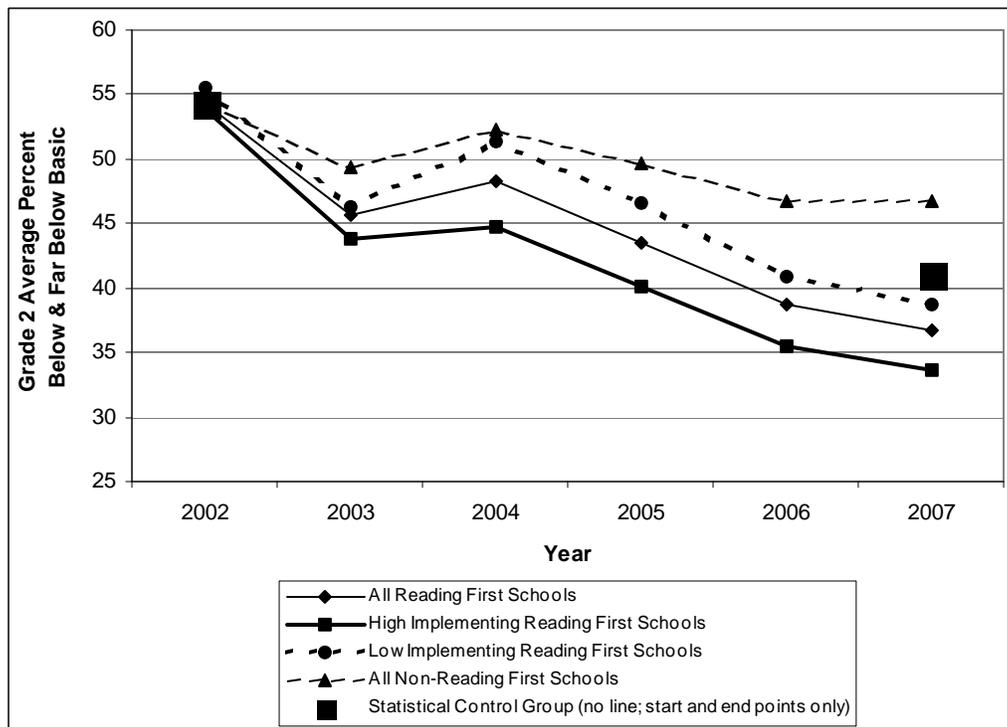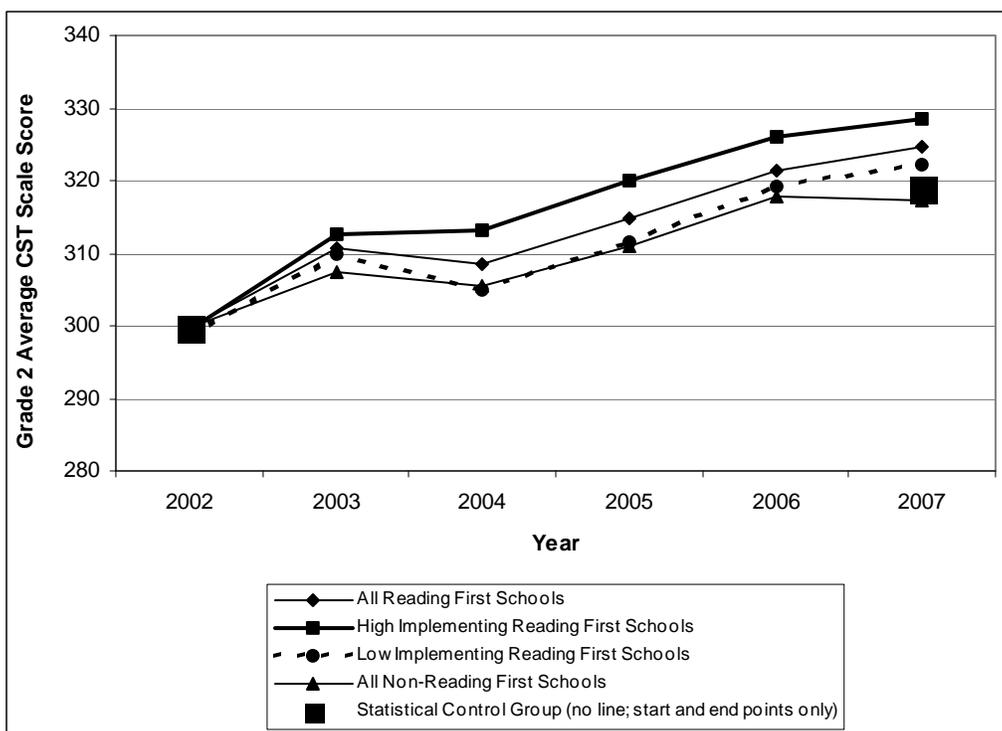
Figure 2.3c: CST Mean Scale Score, YIP = 5, Grade = 2

In addition to the patterns discussed above, we see that growth on the grade 2 metric has, with the exception of 2004, been fairly steady. We see that the high and low implementation schools started at approximately the same location on the scale and fanned out according to their level of implementation. This "fan" pattern strongly supports the finding of program efficacy and rules out the hypothesis that different growth rates are an artifact of different starting points. We also see that the statistical control group and the non-Reading First population have similar growth rates, supporting our contention that the statistical control group is a reasonable proxy for comparable non-Reading First schools.

Notice that the growth from 2006 to 2007 is flatter than for previous years. If this flattening continues in 2008 it may be evidence of a "plateau" effect, about which we have hypothesized in previous reports.

CST and CAT/6 Results for Grade 3 (Table 2.4 and Figures 2.4a – 2.4f)

Table 2.4 reports gains, starting scores, and ending scores for grade 3. In addition to CST scores, grade 3 offers CAT/6 scores for three subject areas: Reading, Language Arts, and Spelling. Grade 3 is unique in this regard, and the extra information proves critical in interpreting the grade 3 results.

Table 2.4: CST and CAT/6 Metrics, YIP = 5, Grade = 3

Years in Program: 5 Grade: 3	Reading First Schools				All Non-Reading First Elementary Schools
	All Reading First Schools	High Implementation Schools (Avg. RFII > 41.4)	Low Implementation Schools (Avg. RFII < 36.0)	Statistical Control Group (RFII = 25.0)	
Number of Schools	259	28	101	N/A	4,048
% Proficient and Above					
2002	14.8	13.2	14.7	14.8	40.1
2007	20.8	25.8	20.0	20.0	41.0
Change Since Starting Year	**6.0[bc]**	**12.6[abc]**	**5.3[bc]**	**5.2**	**1.0**
% Below and Far Below Basic					
2002	57.9	58.4	57.9	57.9	31.2
2007	45.0	40.3	46.6	46.3	27.4
Change Since Starting Year	**-12.9[bc]**	**-18.1[abc]**	**-11.3[bc]**	**-11.6**	**-3.8**
Mean Scale Score					
2002	294.5	293.7	294.3	294.5	333.9
2007	307.4	313.8	305.8	306.7	336.1
Change Since Starting Year	**12.9[bc]**	**20.2[abc]**	**11.4[bc]**	**12.2**	**2.3**
CAT/6 Mean Percentile Rank Reading					
2002	22.5	22.8	22.1	22.5	45.8
2007	27.4	30.5	26.4	26.7	46.3
Change Since Starting Year	**4.9[bc]**	**7.6[abc]**	**4.3[bc]**	**4.1**	**0.5**
CAT/6 Mean Percentile Rank Language					
2002	24.8	24.4	24.9	24.8	44.6
2007	30.4	33.4	29.6	29.9	47.2
Change Since Starting Year	**5.6[bc]**	**9.0[abc]**	**4.6[bc]**	**5.1**	**2.6**
CAT/6 Mean Percentile Rank Spelling					
2002	36.5	35.3	36.3	36.5	52.2
2007	49.6	52.4	47.6	46.6	57.4
Change Since Starting Year	**13.1[abc]**	**17.1[abc]**	**11.2[bc]**	**10.1**	**5.2**

[a] Significantly different ($p < 0.05$) relative to the "Statistical Control Group."

[b] Significantly different ($p < 0.05$) relative to "All Non-Reading First Elementary Schools."

[c] Significantly different ($p < 0.05$) relative to the starting year, i.e., significantly different from a gain of zero.

Gains in grade 3 "% Proficient and Above" are less impressive than those for grade 2, one-third as much. Movement out of the bottom categories is substantial and lags grade 2 by only 5 percentage points. As with grade 2, the Reading First schools strongly out-perform non-Reading First schools in moving students out of the lower performance levels. The mean scale score gain is half that of grade 2. Gains relative to the Statistical Control Group are small, in most cases not significant. On the other hand, the differences between high- and low-implementing schools are much larger than for grade 2. Figures 2.4a – 2.4f reveal that grade 3 has a complexity not shared by the other grades.
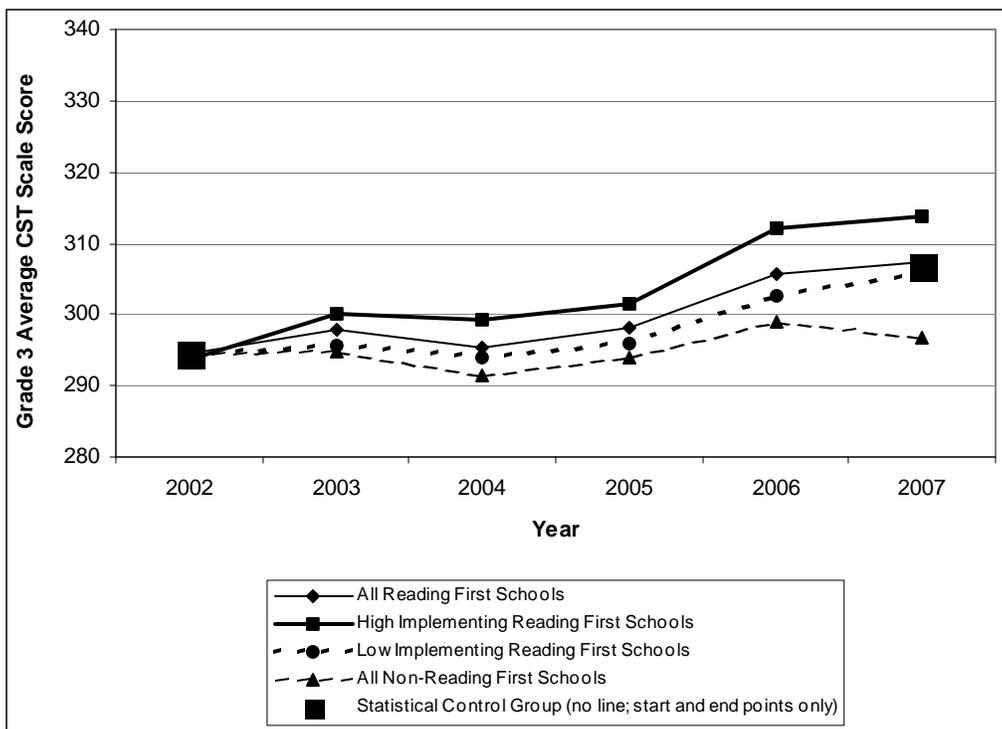
Figure 2.4a: CST % Proficient & Above, YIP = 5, Grade = 3
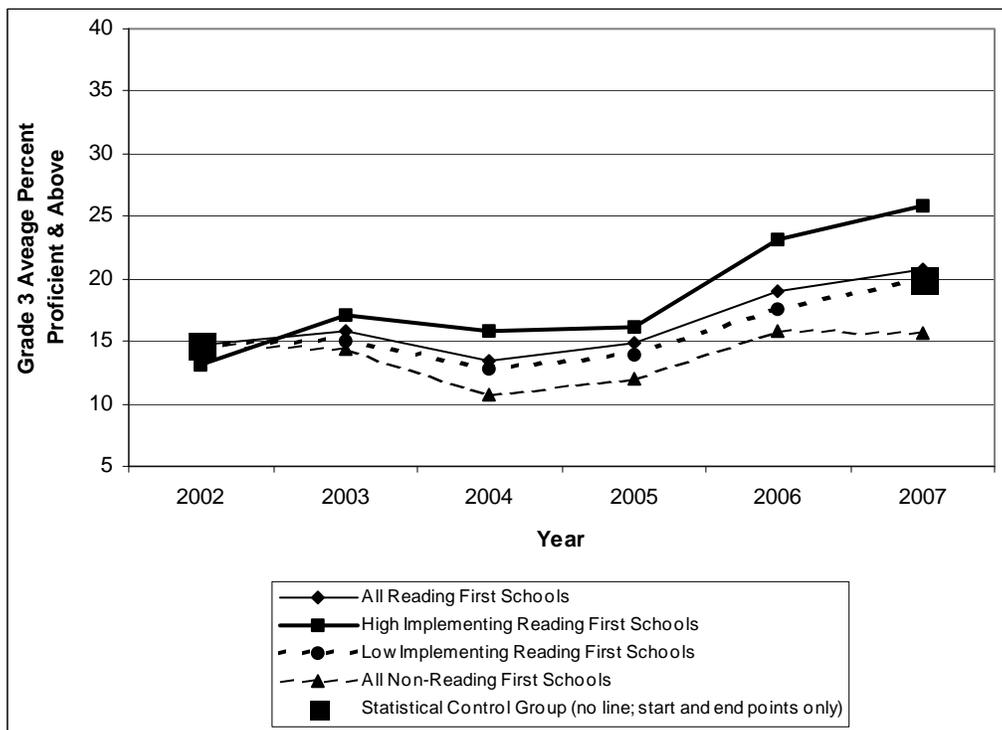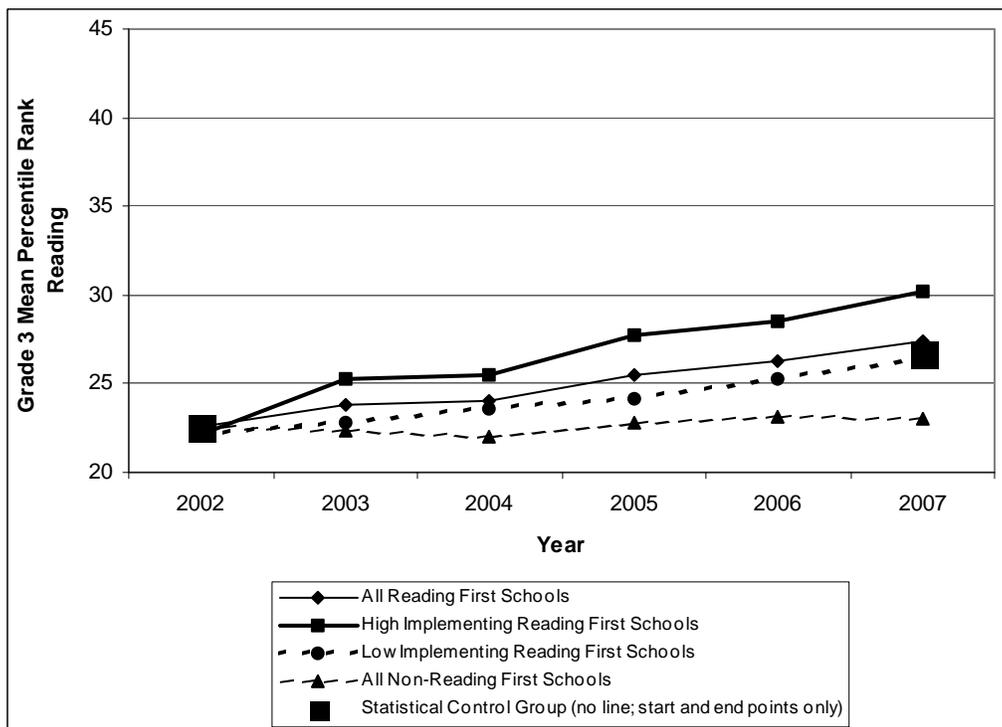

Figure 2.4b: CST % Below and Far Below Basic, YIP = 5, Grade = 3
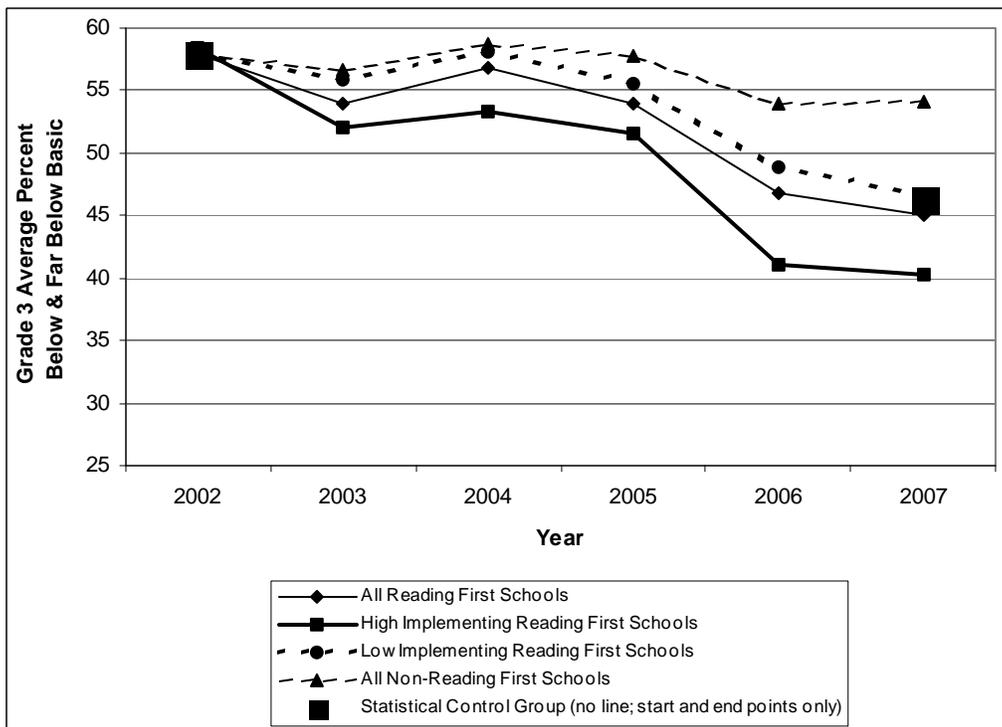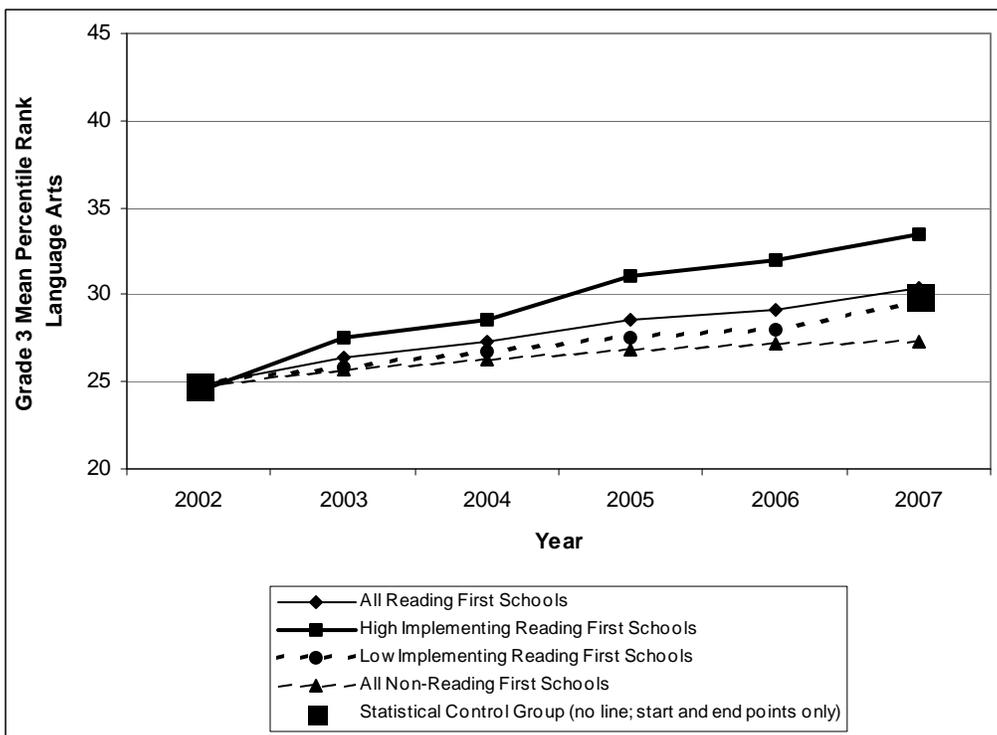
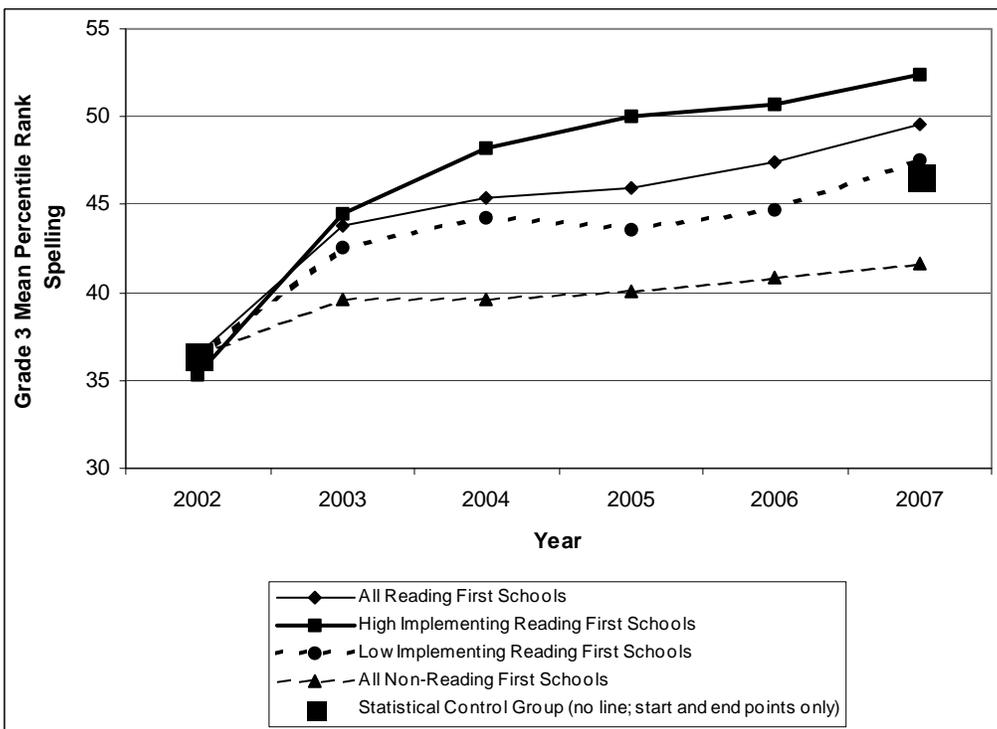**Figure 2.4c: CST Mean Scale Score Per Student, YIP = 5, Grade = 3**

Figure 2.4d: CAT/6 Reading, Mean Percentile Rank, YIP = 5, Grade = 3

Figure 2.4e: CAT/6 Language, Mean Percentile Rank, YIP = 5, Grade = 3

Figure 2.4f: CAT/6 Spelling, Mean Percentile Rank, YIP = 5, Grade = 3

Figures 2.4a – 2.4f reveal a number of important patterns that are not readily apparent in the statistics of Table 2.4. The most obvious, noted in preceding reports, is that the grade 3 CST scores dip substantially in 2004, creating a "U" shape. We see that even though the CST trends for Reading First schools is somewhat flat relative to grade 2 (Figures 2.4a – 2.4c), they are substantially more positive than those for the non-Reading First schools. After 2004, the trends are steadily positive, with a bit of flattening from 2006 to 2007.

A second pattern is that the CAT/6 trend-lines are qualitatively different than those for the CSTs. There is no "U" shape, just a steady positive trend ranging from slight in the cases of Reading and Language to large in the case of Spelling. In combination with the grade 2 and grade 4 results (below), this cautions us not to place too much weight on the shape and relative direction of the grade 3 CST trend-lines.

As regards the CAT/6 trend-lines, Spelling has a substantially higher starting point than Reading and Language and its trend lines range from 40 to 50 on the Mean Percentile Rank metric. This puts its trend-lines around the lower inflection point of the nationally normed CAT/6 population, where a given amount of ability growth is likely to show the largest changes in the percentile metric. The Reading and Language trend-lines are lower in the distribution where the same amount of ability growth will cause a smaller change in percentiles. This warns us that the absolute size of the trends in the CAT/6 metric may be in part an artifact of their position on the distribution.

A fourth pattern is that the statistical control group tends to show much higher gains, positive or negative, than the non-Reading First schools. In other words, for grade 3 the control group does not seem to behave like "comparable non-Reading First schools." It is behaving more like the low-implementing Reading First schools and is only marginally lower than the trend-line for all Reading First schools.

This highlights a fifth pattern. While the relative proximity of the low-implementing schools, all Reading First schools, and the statistical control group would seem to indicate a weak statistical relationship between implementation and achievement on the grade 3 metric, we see that the high implementing schools show *dramatically* higher gains than all the other schools. Thus, there does seem to be a strong implementation effect for grade 3, but only above a certain threshold of implementation, presumably around an RFII of 41. Schools below this threshold tend to show much more modest growth.

The sixth pattern is that while the "All Reading First" trend-lines may be modest relative to the high implementing schools, the trend-lines for non-Reading First schools show little or no growth on all the grade 3 achievement metrics. They do not seem to have improved much at all over the same period.

Thus, despite considerable statistical complexity, we find that Reading First efficacy is supported by the grade 3 achievement trend-lines.

CST Results for Grade 4 (Table 2.5 and Figures 2.5a – 2.5c)

Table 2.5 reports the CST results for grade 4 which have been collected only for YIP 5 schools. (The grade 2 and grade 3 results for YIPs 3 and 4 are reported in Appendix D.) Table 2.5 and its accompanying trend-lines demonstrate that Reading First is having a sustained effect that supports the student population as they move into the upper grades. This may prove to be the most telling of the Reading First effects since it supports the hypothesis that students in Reading First classrooms are learning skills that generalize beyond the course content and test material of the first few grades in elementary school.

Table 2.5: CSTs, YIP = 5, Grade = 4

Years in Program: 5 **Grade: 4**	Reading First Schools				All Non-Reading First Elementary Schools
	All Reading First Schools	High Implementation Schools (Avg. RFII > 41.4)	Low Implementation Schools (Avg. RFII < 36.0)	Statistical Control Group (RFII = 25.0)	
Number of Schools	255	26	101	N/A	3,992
% Proficient and Above					
2002	15.2	14.2	15.9	15.2	42.1
2007	31.3	34.8	30.0	27.5	55.7
Change Since Starting Year	**16.1**[abc]	**20.6**[abc]	**14.1**[ac]	**12.3**	**13.6**
% Below and Far Below Basic					
2002	47.8	48.6	46.7	47.8	23.2
2007	32.2	28.4	33.4	34.9	16.9
Change Since Starting Year	**-15.6**[abc]	**-20.2**[abc]	**-13.3**[bc]	**-12.9**	**-6.3**
Mean Scale Score					
2002	306.8	305.1	307.7	306.7	340.9
2007	327.3	331.3	325.8	322.1	359.5
Change Since Starting Year	**20.5**[abc]	**26.2**[abc]	**18.1**[ac]	**15.4**	**18.6**

[a] Significantly different (p < 0.05) relative to the "Statistical Control Group."

[b] Significantly different (p < 0.05) relative to "All Non-Reading First Elementary Schools."

[c] Significantly different (p < 0.05) relative to the starting year, i.e., significantly different from a gain of zero.

Grade 4 sharply reinforces the growth picture presented by the grade 2 trend-lines. Reading First schools grow significantly faster than the control group. They grow marginally faster than the non-Reading First schools on the "% Proficient and Above" and "Mean Scale Score" metrics, but dramatically faster on the "% Below and Far Below Basic" metric. Movement out of the bottom two categories matches movement into the top categories, unlike non-Reading First schools. The average scale score growth is 20 points over five years, not far shy of the 25 points seen in grade 2. What makes this table remarkable is that Reading First is only administered in grades K-3. There is no grade 4 Reading First program. Yet the

CST scores are almost what one would expect if Reading First extended to grade 4. This demonstrates that Reading First students have been able to carry with them the skills and habits that they developed in the earlier grades, and that rigorous instruction in the lower grades lays the groundwork for large gains in the higher grades.

Non-Reading First schools also show substantial gains over this period, but the gains are slightly smaller and much less uniform across the population as can be seen in Figures 2.5a – 2.5c.
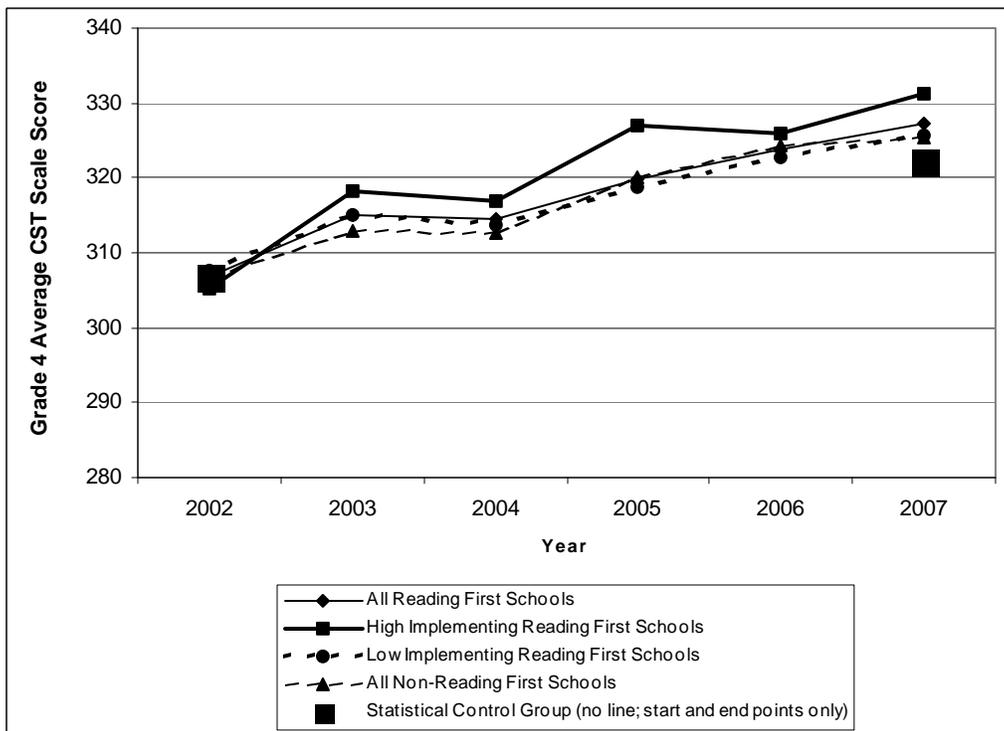
Figure 2.5a: CST % Proficient & Above, YIP = 5, Grade = 4
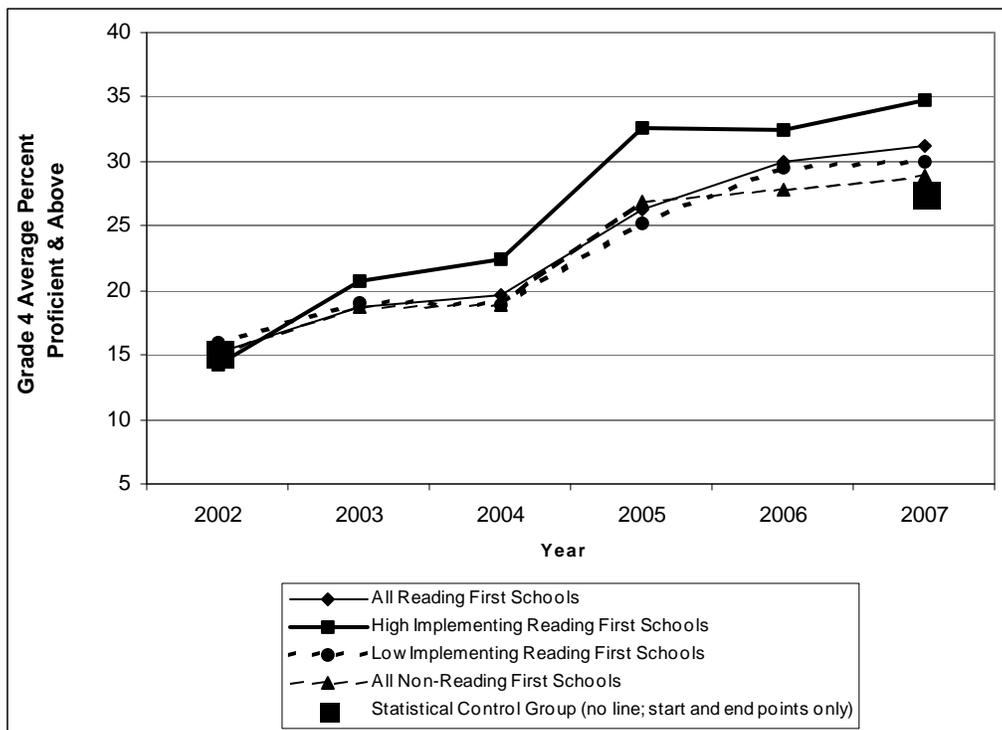
Figure 2.5b: CST % Below and Far Below Basic, YIP = 5, Grade = 4
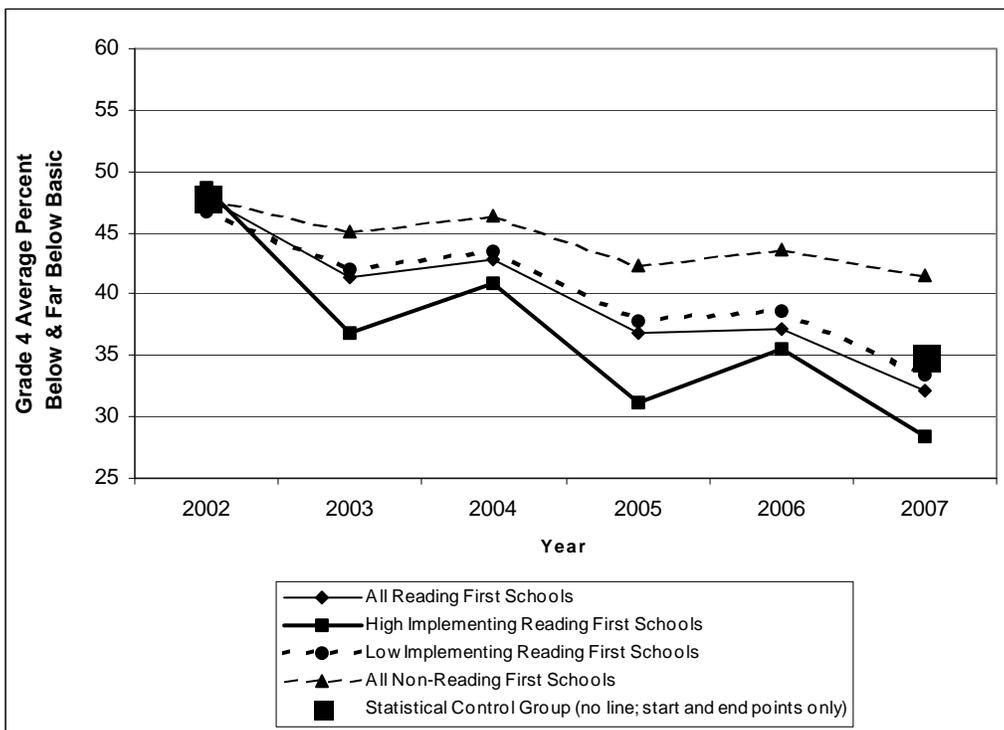
Figure 2.5c: CST Mean Scale Score, YIP = 5, Grade = 4

Figures 2.5a (% Proficient and Above) and 2.5c (Mean Scale Score) suggest that the All Reading First, Low Implementing, and non-Reading First trend-lines, as well as the statistical control group, do not grow in ways that are particularly different. They all show robust growth. Unlike for grade 3, the significance statistics for grade 4 (Table 2.5) best capture the differential effect of Reading First.

However Figure 2.5b, the effect of Reading First on movement out of the lower categories, reveals that Reading First schools far outpace non-Reading First schools in the lower performance levels. Low-performing students in non-Reading First schools run a real risk of becoming mired in the Below Basic and Far Below Basic performance levels, unable to get beyond elementary reading tasks even as their peers surge ahead. By comparison, low-performing students in Reading First schools appear to enjoy a decisive advantage and have the tools to keep up with their peers.

In addition, we see in grade 4 a repetition of the pattern that was so evident in grade 3, that high implementing schools "break out" from the rest of the schools and produce distinct and impressive trend-lines. While the grade 4 high implementing trend-lines tend to be somewhat jagged, that is probably an artifact of the relatively small number of schools (26 out of 255) in this group. The smaller the sample, the less stable the trend-line. The "break out" effect suggests, again, that there is some threshold of implementation above which schools experience a qualitatively higher level of achievement and sustainability.

The grade 4 effect strongly supports the strategy of focusing on the early grades by providing funds, professional development, coaching, and curricular coherence. This is consistent with extensive research that documents the importance of a strong foundation of early reading development, a concept that is central to the national Reading First initiative (e.g., Foorman & Torgesen, 2001; National Reading Panel, 2000; Snow, Burns & Griffin, 1998).

Regression Effect Sizes

We have mentioned that in order to calculate gains for the statistical control group we perform a regression analysis for every achievement variable. To summarize the results of these regressions, we present regression tables that show the effect of Reading First implementation on the RFAI and grade 4 CST "% Proficient and Above" metrics. In both cases, the dependent variable to be predicted is the 2007 score for the relevant achievement variable, i.e., the 2007 RFAI and the 2007 grade 4 percent Proficient and above. We select the RFAI as an outcome variable because it embodies data from the grade 2 and 3 CSTs, the CAT/6, and K-3 EOY data. We select the grade 4 percent Proficient and above as an outcome variable because it measures the degree to which participation in the K-3 Reading First program influences performance in subsequent grades. Also, it is the only achievement variable not contained in the RFAI.

Regression analysis involves identifying a number of "predictor" variables that contribute information regarding the "dependent" variable. In this case, we found that, in addition to "Average Yearly RFII" and "Years in Program," the percentage of Socio-Economically Disadvantaged (SED) students in a school was a significant predictor of achievement. The percentage of English learners proved not, in general, to be a significant predictor of the RFAI and grade 4 % Proficient and above dependent variables and was not included.

In a separate regression we multiply Average Yearly RFII by Years in Program to create a composite variable that reflects total Reading First implementation over the years in the program. This effect is presented with the others in Tables 2.6 and 2.7 and is highlighted with bold type. The footnotes refer to both tables below.

Table 2.6: Effect Size of Variables Predicting Percent of Students Proficient & Above on Grade 4 CSTs in 2007 ($R^2 = 0.22$)[1]

Predictor Variable (Predicting 2007 Grade 4 % Proficient & Above)	Standardized Beta Coefficient Effect (standard deviation units)[2]	t-test (t > 1.96 implies significance with 95% confidence)[3]	Probability the Effect is by Chance[4]
Starting CST Gr. 4 % Proficient & Above	0.40	11.0	0.0000
Number of Years in Program	0.18	5.4	0.0000
Average Yearly RFII	0.14	4.3	0.0000
Yearly RFII * Years in Program[5]	**0.21**	6.6	0.0000
Percent of SEDs in School	-0.12	-3.5	0.0006

Table 2.7: Effect Size of Variables Predicting the 2007 RFAI ($R^2 = 0.51$)[1]

Predictor Variable (Predicting the 2007 RFAI)	Standardized Beta Coefficient Effect (standard deviation units)[2]	t-test (t > 1.96 implies significance with 95% confidence)[3]	Probability the Effect is by Chance[4]
Starting RFAI	0.68	27.4	0.0000
Number of Years in Program	0.21	8.7	0.0000
Average Yearly RFII	0.07	2.9	0.0045
Yearly RFII * Years in Program[5]	**0.22**	9.2	0.0000
Percent of SEDs in School	-0.06	-2.5	0.0126

[1]The R^2 statistic reports the percentage of variance that is explained by the model.
[2]The "Standardized Beta Coefficient" shows how many standard deviations the CST "% Proficient & Above" increases for every one standard deviation increase of that predictor variable.
[3]The "t-test" shows how many times larger the effect is than what would be predicted by chance.
[4]The "Probability" column uses the t-statistic to compute the probability that the observed effect occurred by chance.
[5]The "Yearly RFII * Years in Program" predictor variable is the product of a school's "Average Yearly RFII" and its "Number of Years in Program" (equal to the sum of its RFII statistics over time). To avoid collinearity, its effect size was computed in a separate regression run in which "Average Yearly RFII" and "Years in Program" were removed.

How to Interpret the Regression Tables

The predictor variables we are interested in are "Number of Years in Program" and "Average Yearly RFII" and the variable that is obtained by multiplying them together, "Yearly RFII*Years in Program." The latter can be thought of as a school's total amount of Reading First implementation over time. The "Percent of SEDs in School" is not a variable of primary interest, but its role here is to remove confounding influences that socio-economic status might have on the implementation effect. The role of the "Starting" variable is to remove the effect of the school's achievement starting point so that we can treat all schools in the sample as if they started at the same performance level.

The two right columns – the t-test and the probability -- answer the question: How likely is it that we would have encountered the observed effect size in this row by chance? We see that the t-test statistics are all above 1.96 and the probabilities (which are calculated from the t-statistics) are all well below 0.05. That means it is very unlikely that we would have obtained these effect sizes by chance.

That Reading First implementation is a significant predictor of achievement gain supports the claim of efficacy. We have demonstrated that Reading First implementation matters and that it is not an artifact of SED or Starting Point or percent of English Learners (not shown here because it is not a significant predictor). Though we are naturally drawn to examine the effect size, called here the "standardized beta coefficient," it is very difficult to interpret and there are no accepted industry standards on how to decide whether a given standardized beta coefficient is "good" or not. To examine "effect size," it is better to examine the relative gains reported in the tables earlier in this chapter, especially relative to the statistical control group gains that were computed, in fact, from regression tables just like these.

That said, the accepted way to interpret a standardized beta coefficient is as the change in standard deviation units that is expected in the dependent variable given a change of one standard deviation in the predictor variable. We see that the "Yearly RFII*Years in Program" beta coefficient equals 0.21 and 0.22 for the Year 4 Proficient & Above and RFAI dependent variables. That means for every one standard deviation increase on the total implementation scale we can expect achievement to increase about one fifth of a standard deviation on the achievement scale. Calculating the standard deviations of the variables and looking at typical total implementation levels, we see that this means that total implementation will generally increase both the RFAI and grade 4 scores around 6 to 8 points in a 5-year period for an average implementing school. And that is what we see in the trend-lines – generally a 6 to 8 point difference between the statistical control group and the average Reading First school after 5 years, depending on the achievement metric.

Measurement Error Lowers Effect Size

While a 0.21 or 0.22 effect size is reasonably large for a study of this kind, it strongly *understates* the "true" effect of Reading First implementation on achievement, both in the regression tables and in the trend-line charts and tables.

The standardized beta coefficient and the gain score differences between high and low implementing schools and between all Reading First schools and the statistical control group – all of these differences assume that our RFII implementation measures and achievement measures are *perfectly* precise. Obviously they are not. Achievement tests have a wide margin of error when measuring student ability (reduced when aggregated to the school level). Most important, the RFII has a very large margin of error, which arises from a variety of sources:

- Ambiguity in the survey questions

- Biases caused by teachers overstating or understating their school's level of implementation

- Uncertainty caused by teachers not understanding the questions or encountering questions that do not apply to them

- Schools that show high achievement gains but report low Reading First implementation because of the use of effective non-Reading First programs, initiatives, and other causal factors

- Schools that report high Reading First implementation but show low achievement gains due to circumstances out of their control or perhaps to biased reporting.

As measurement error increases, the measured or *observed* effect becomes smaller in accordance with the statistical law known as "regression to the mean." This will happen even in cases where there is a nearly perfect causal effect. For instance, even the observed relationship between physical exercise and muscle tone can approach zero if the instruments used to measure physical exercise and muscle tone have a high degree of measurement error.

Therefore, it is important to remember that the effect sizes reported in the Reading First evaluation are on the conservative side, as they are for most evaluation studies. There is no obvious or widely accepted way to correct for measurement error with these kinds of variables.

Is the implementation effect as strong in the Year 5 Report as it was in the Year 4 Report?

The answer is yes, for the most part. The Year 4 "Yearly RFII*Years in Program" effect was 0.17. The corresponding Year 5 statistic is 0.22. However, the Average Yearly RFII effect, taken in isolation,

dropped from 0.09 to 0.07. As regards CST % Proficient & Above, even though the Year 4 report used the grade 2 results and the Year 5 report uses grade 4 results, we find that the effect sizes for all the implementation variables are virtually identical. Thus, grade 2 and grade 4 substantially confirm each other. The grade 3 CSTs have an anomalous relationship with all the remaining achievement variables, including the grade 3 CAT/6.

Why is the Average Yearly RFII coefficient different between the RFAI and Grade 4?

The Average Yearly RFII effect is 0.14 for grade 4 Proficient and Above but 0.07 for the RFAI – half as much. This is a consequence of the 45% weight of grade 3 achievement metrics in the calculation of the RFAI. Whatever anomalies exist in grade 3 are to some extent inherited by the RFAI. However, because the non-grade 3 achievement metrics "straighten out" grade 3's U-shaped CST trend-line, the Years in Program effect is larger on the RFAI than for grade 4, and the two implementation variables together are sufficient to yield an overall 0.22 "Mean RFII * Years in Program" effect for the RFAI.

Should "Years in Program" be included in the implementation effect if all schools are trending upward?

When Average Yearly RFII is multiplied by Years in Program, we see that the total implementation effect on the RFAI is 0.22 (consistent with, or higher than, effects reported in the Year 3 and Year 4 reports), whereas the effect of Average Yearly RFII alone is smaller (0.07). We argue that the 0.22 effect size is probably the more valid estimate of the total Reading First effect, but this requires us to assume that the Years in Program effect is sensitive primarily to the school's implementation of Reading First and that it is independent of non-Reading First effects on achievement.

This is a strong assumption but not unreasonable. Reading First schools agree, as a condition of funding, not to implement competing programs or initiatives that are not aligned with Reading First. This has an important theoretical implication. Because we see strong achievement gains in schools with high RFIIs – in fact gains that are stronger than for lower implementing schools – and because higher RFIIs imply that such schools are implementing Reading First more *exclusively*, we can conclude that most of the Years in Program effect that we observe in high implementing Reading First schools is caused by Reading First and *not* by non-Reading First reading programs or non-Reading First pedagogical practices that would be precluded by the program.

This observation leads to a vitally important question. If the achievement gains experienced by Reading First schools over a five-year period are primarily a Reading First effect, as we suggest, why do non-Reading First schools also show substantial gains (though not as large) over the same period of time?

Assuming the gains are real, and not an artifact of the tests, we see two possibilities:

Non-Reading First schools have, over the same period, begun implementing non-Reading First educational strategies that happen to be effective; or

Non-Reading First schools have been implementing some or all of the same program elements that make Reading First effective.

A review of state educational initiatives supports the second possibility. The state, in January of 2002, adopted two reading curricula for K-8 schools to use. These are the same Houghton-Mifflin and Open Court reading programs required in Reading First. Schools that adopt these programs have access to SB 472 teacher professional development, AB 75 principal professional development, and the 6-8 week skills assessments. In addition, many LEAs and schools have opted to hire reading coaches at their own expense. Such non-Reading First schools become virtually indistinguishable from Reading First schools in terms of educational practices in the classroom. The main difference is that the non-Reading First schools must use other funding sources to hire reading coaches and provide professional development.

Therefore, we believe that the statewide trend is fundamentally an effect of the same educational practices and program elements that are required by Reading First. To prove this hypothesis, a separate non-Reading First implementation study is necessary. This would mean administering a version of the Reading First implementation survey to a representative sample of non-Reading First schools.

Conclusions

The conclusions in the Year 5 Report reinforce and extend those of the Year 4 Report. We began the chapter by stating that Reading First would be said to show evidence of being effective to the degree that:

1. Achievement gains in Reading First schools are positive for grades 2, 3, and 4.

2. Reading First schools show higher achievement gains than non-Reading First schools for grades 2, 3, and 4.

3. Reading First schools show higher achievement gains than what would be predicted from a statistical control group for grades 2, 3, and 4.

4. High implementing Reading First schools show higher achievement gains than low implementing Reading First schools for grades 2, 3, and 4.

The Year 5 Report finds that the answer continues to be "yes" to all four questions with a small number of exceptions in particular instances. It extends these findings especially to grade 4. This confirms the findings of the previous grades and supports the hypothesis that students who have progressed through Reading First programs in grades K-3 are better prepared for higher grades than students who have not.

We conclude this chapter by restating from the Year 4 Report an important idea discussed in this chapter. Reading First implementation, and thus Reading First exclusivity at the school site, is a significant predictor of positive cross-year gains. This fact supports the hypothesis that the upward trend in reading scores in such schools since 2002 is the result of Reading First. Because the rest of the state K-3 schools have shown similar, though less dramatic, upward trends over the same time period, it is possible that the statewide trend in non-Reading First schools is being driven by the same program elements that are driving the Reading First gains. Anecdotal evidence suggests this to be the case. If subsequent research should bear out this hypothesis, it would validate efforts to make such program elements available to all California schools, not just those in Reading First.

References

Foorman, B. R., & Torgesen, J. (2001). Critical elements of classroom and small-group instruction promote reading success in all children, Learning Disabilities Research and Practice, 16(4), 203–212.

Hasbrouck, J., & Tindal, G. (2005). *Oral Reading Fluency: 90 Years of Measurement (Tech. Rep. No.33).* Eugene, Oregon: University of Oregon, College of Education, Behavioral Research and Teaching.

National Reading Panel. (2000). Report of the National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Washington, DC: National Institute of Child Health and Human Development.

Parker, J. (2003). *Do Interim Assessments for Reading First Predict Success on the English/Language Arts California Standards Test?* Paramount Unified School District. Paper presented to the California Educational Research Association (CERA), November 14, 2003.

Snow, C. S., Burns, S. M., & Griffin, P. (1998). Preventing reading difficulties in young children. Washington, DC: National Academy Press.

U.S. Department of Education (2002). The No Child Left Behind (NCLB) Act, Title I, Part B, Subpart 1, Section 1201. Signed into federal law January 2002.

Yovanoff, P., & Tindal, G. (2003). *Scaling Early Reading Performance Tasks (Tech. Rep. No. 21).* Eugene, Oregon: University of Oregon, College of Education, Behavioral Research and Teaching.

Chapter 3: Implementation of Reading First

This chapter presents data gathered from surveys of Reading First participants used to address the question: How well has the Reading First program been implemented in each participating school and district? Principal, reading coach, and teacher surveys provide a global perspective on implementation in Reading First schools as well as information about specific dimensions of program implementation such as professional development, material and instructional resources, understanding of Reading First Assurances and curricular materials, and perceptions of the Reading First program.

To evaluate the implementation of Reading First in California, Educational Data Systems (EDS) developed three surveys– one each for Reading First teachers, coaches, and principals – and administered them annually from 2004 to 2007. Because participation in the evaluation process is part of the commitment that local education agencies (LEAs) make when they apply for funding, the response rate on the surveys has been high. In 2007, a total of 17,261 usable surveys were received from teachers, 1,028 from reading coaches, and 1,073 from principals, totaling 19,362 and yielding a response rate of 91%.[1] Results of the surveys can be found in Appendices A – C of this report.

This chapter primarily discusses the analysis of the survey data to compute a Reading First Implementation Index (RFII) for each school. This index is used to evaluate the overall implementation at the school level.

Key points in this chapter are:

- Measuring implementation is an essential element in assessing program effectiveness (i.e., the potential of a program to produce achievement gains given a sufficient level of implementation).

- Most schools in the Reading First program are implementing the program adequately.

- The average level of implementation has risen throughout the duration of the Reading First program. The average (RFII) across all schools was 39 in 2006 and 2007, compared to 36 in 2004 and 2005.

- Schools that have been in Reading First for two or more years have higher average implementation than newcomers.

- The RFII can be interpreted as a (theoretical) percentage of times that teachers rate their schools "more than adequate" on relevant survey questions. Using the distribution of school RFII measures, it is possible to state how many schools in the state meet the "more than adequate" standard from the point of view of teachers on selected dimensions.

---

[1] For response rates and specific information from previous years, the reader is referred to past reports available at: www.eddata.com/resources/publications/.

Measuring Reading First Program Implementation

To fully evaluate the effectiveness of an educational program, it is not enough to look at student achievement gains alone. Rather, it is necessary to examine achievement gains in relation to the degree of implementation of the program elements, or implementation fidelity (Dane & Schneider, 1998; Ruiz-Primo, 2006). If it is found that duration and intensity of program implementation are significant predictors of achievement, then we can say that evidence exists that the program has an impact on achievement, the ultimate desired program outcome. If achievement gains bear no relation to the degree of program implementation, no evidence of program efficacy can be claimed (Schiller, 2001).

Fidelity of implementation is defined as "the degree to which an intervention [or program] is implemented as planned" (Gresham, Gansle & Noell, 1993). Studies of implementation have found significant correlations between degree of implementation of an educational program and student outcomes (Dane & Schneider, 1998; Leinhardt, Zigmond & Cooley, 1981). Therefore, the monitoring of implementation fidelity provides evidence regarding the extent to which the program elements are being applied according to design so that those responsible for program oversight can determine whether adjustments are needed to improve effectiveness (Power, Blom-Hoffman, Clarke, Riley-Tillman, Kelleher, & Manz, 2005).

In this chapter, we use survey data to quantify the degree of implementation occurring within each Reading First school. For each school, multiple respondents completed the survey, providing the perspectives of the site principal, the reading coach, and participating teachers. A school that may report a low level of use of curricular materials, neglects professional development, or skimps on instructional time, for example, would not be considered to be implementing the program. When "implementation" is defined in this more tangible way, assuming it can be measured with reasonable accuracy, it becomes feasible to decide whether the program has the *potential* of working if it is well implemented.

Rationale for Using a Survey

To directly measure the presence, absence, or degree of implementation of Reading First in *all* participating schools and districts is a daunting task. There is no statewide database that would definitively reflect Reading First implementation, and it is impossible within the scope of this evaluation to conduct observations at all sites. In 2007 there were 886 Reading First schools in California. To measure implementation in each school, the external evaluator would ideally send trained auditors to observe each Reading First classroom over an extended period of time. While this would not be practical for the complete population of schools, it could in theory be done with a representative sample of schools (absent legal restrictions). However, the State has specifically requested in its Request for Proposals an implementation measure for *all* Reading First schools. To obtain information about implementation from

all Reading First schools and districts, teachers, principals, and reading coaches in all Reading First schools were asked to complete a comprehensive survey constructed to gather information about the presence, absence, and degree of utilization of the critical elements that define the implementation of the Reading First program.

The advantage of using a survey is that it is feasible to administer and analyze results from all schools, and the respondents (teachers, coaches, principals) are the most knowledgeable regarding what is happening inside their schools and classrooms throughout the school year. Nonetheless, there are unavoidable limitations and sources of bias:

1. The respondents are, to a certain extent, reporting on themselves. This could lead to upward bias in estimations of school implementation since respondents may feel a desire to respond "appropriately," or they may be unclear regarding what "full" implementation looks like.

2. Similarly, if school officials believe that survey results could be used to reduce or deny funding, there would be a strong incentive for some school personnel to encourage respondents to respond in a way that would raise the school's implementation score, also leading to an upward bias.

3. While an upward bias would probably apply to all schools to some degree, it might be more pronounced in some schools than others. This would introduce an extra source of error in the *relative* measures of schools.

4. In order for a survey to be specific enough to be useful, it needs to have questions tailored to particular types of respondents. For instance, there need to be questions tailored specifically to teachers, coaches, and principals, and to users of Open Court and Houghton Mifflin in the Spanish and English versions. This impairs our ability to compare schools when they have different proportions of each respondent type.

5. To the degree the survey instrument is changed from year to year, results could lose their cross-year comparability.

6. Each question, taken on its own, inevitably carries ambiguities and imprecision. It is often difficult to be clear exactly what dimensional construct is being measured by a question, and whether it is indeed "implementation."

These issues have been discussed at length in previous reports and accepted survey analysis models have been used to ameliorate these potential limitations throughout the four years of the survey use.[2] To summarize, the above issues are addressed as follows:

---

[2] The reader is referred to previous annual reports at www.eddata.com/resources/publications/ for details about the development of the survey and analysis procedures.

6. Schools are measured relative to each other rather than against an absolute standard.

7. Teachers complete the survey anonymously, enhancing their ability to report truthfully about the program. Because in most schools there is only one principal and one reading coach, their responses are not entirely anonymous, though school code numbers and not school names are used in the analysis process. A school's implementation measure pools together the teacher, principal, and coach responses.

8. Questions are worded so that their "correct" answers are not immediately obvious, increasing the chance that respondents select truthful answers.

9. There are numerous opportunities for cross-verification of findings across respondents within a school. Respondents not only report their own use of program elements but also rate other respondent types (coaches rate teachers, teachers rate coaches, etc.).

10. The implementation survey provides data that are used for making program adjustments and no "high-stakes" funding decisions rest on results. The "significant progress" regulations[3] approved in fall 2007 are based entirely on achievement data.

11. Equating methods are used to equate responses across respondent groups and across program years.

12. The potential ambiguity at the question level is addressed by using statistical methods to group items' coherent dimensions that seem to cluster together statistically and are validated by experts in the California Technical Assistance Center (C-TAC) and the Evaluation Advisory Group (EAG).

The reliability (Cronbach-alpha) of the Reading First Implementation Index has been well established in previous reports and has ranged from .90 to .92 (a reliability of 0.85 is widely considered sufficient). Additionally, the validity of using the RFII as a measure of school-level implementation has been previously established. Given the high content validity of the Reading First survey and its level of detail, the use of methodological tools that correct for common sources of bias, and the statistical and psychometric characteristics of the RFII, we consider the RFII to be sufficiently valid and reliable as a means for measuring implementation at the school level.

---

[3] Information on "significant progress" as available at: http://www.cde.ca.gov/pd/ca/rl/rdfst06achievedef.asp.

Changes to the Survey

From year to year, it has been necessary to make minor changes to the survey to reflect programmatic changes or to clarify ambiguous items. In each round of changes, equating procedures have been employed to allow for cross-year comparisons. The changes over time are summarized in this section.

Individual questions throughout the survey underwent editorial modifications, often to clarify routing from section to section on the web survey. In 2005, based on a change in the Reading First program to include Spanish curricular materials for waiver classrooms (instruction in Spanish), the teacher survey was expanded to include additional questions involving the receipt and use of the Spanish versions of curricula. In 2006, further revisions were made to clarify which curricular materials were referenced in specific questions. In 2007, very minor wording changes clarified some items thought to be potentially confusing or no longer relevant in a program that has been in place for several years. In each round of revisions, efforts were made to retain enough "old questions" to link the different survey administrations together.

Anecdotal information received from teachers and coaches indicates that it took 20 to 30 minutes to complete the survey.

Calculating the Reading First Implementation Index (RFII)

Previous reports have described in detail the steps by which the RFII was constructed and how it is calculated. In short, the procedure is as follows:

Using an Item Response Theory program called Facets, subsets of questions across the three surveys are used to generate measures on 17-19 dimensions.[4]

Three of these dimensions are used to calculate each school's RFII. They are: School Implementation Overall (SIO), Overall Reading First Understanding (OUND), and Teacher/Coach Professional Development (TCPD).

---

[4] There are a number of methods for analyzing survey data. The method used here, the Many-Facet Rasch Model or Facets, is well-suited to judging and equating designs in which there are large amounts of missing data and the data consist of "subjective judgments" (Linacre, 1994). Facets is a generalization of the Rasch Model, which is one of a number of psychometric models organized under the rubric of "Item Response Theory." These are the models behind many large-scale student assessments and licensure examinations, chosen especially for their ability to equate test forms so that students who are exposed to different test forms can nonetheless be measured accurately on a common scale.

The measures on these dimensions are weighted and combined to calculate the school's RFII. The weights are:

School Implementation Overall (SIO)   = 70%

Overall Reading First Understanding (OUND) = 20%

Teacher/Coach Professional Development (TCPD) = 10%

The resulting RFII statistic is scaled to be between 0 and 100 and to have a distribution similar to that of the Reading First Achievement Index (RFAI).

Based on advice from the EAG, starting in 2007, the RFII of a school in a given year is averaged with its RFII from the preceding year. It is hoped that this will make the RFII more robust to changes in the sample of teachers in each school who take the survey each year while allowing it to be reflective of the school's recent implementation history. For this report, when we refer to the 2007 RFII, it is actually the average of the 2006 and 2007 RFIIs for each school.

Implementation Results

<u>Distribution and Interpretation of the RFII</u>

Figure 3.1 shows how the RFII was distributed across all Reading First schools in 2007. The mean 2007 RFII was 39; the standard deviation around the mean was 5. This can be practically interpreted as follows: Reading First *teachers* on average found their schools to be "more than adequate" 39% of the time (i.e., on 39% of the relevant items). Interpreting the RFII as a percentage of items is not strictly correct. The RFII is actually based on a statistical probability that teachers in a school will rate their school "more than adequate" across the test. It is a theoretical statistical parameter used to explain the data, not a literal count of responses. Interpreting it as a percentage of items scored "more than adequate" makes it easier to understand, however.

Note the emphasis on teachers; the RFII was intentionally calibrated relative to teacher perceptions of "more than adequate implementation." Teachers tended to give lower scores to their schools than coaches and principals. While most of the dimension measures in Table 3.1 in the next section are calibrated relative to teachers, some of the dimensions are calibrated relative to coaches and principals as indicated in the footnotes to the table.
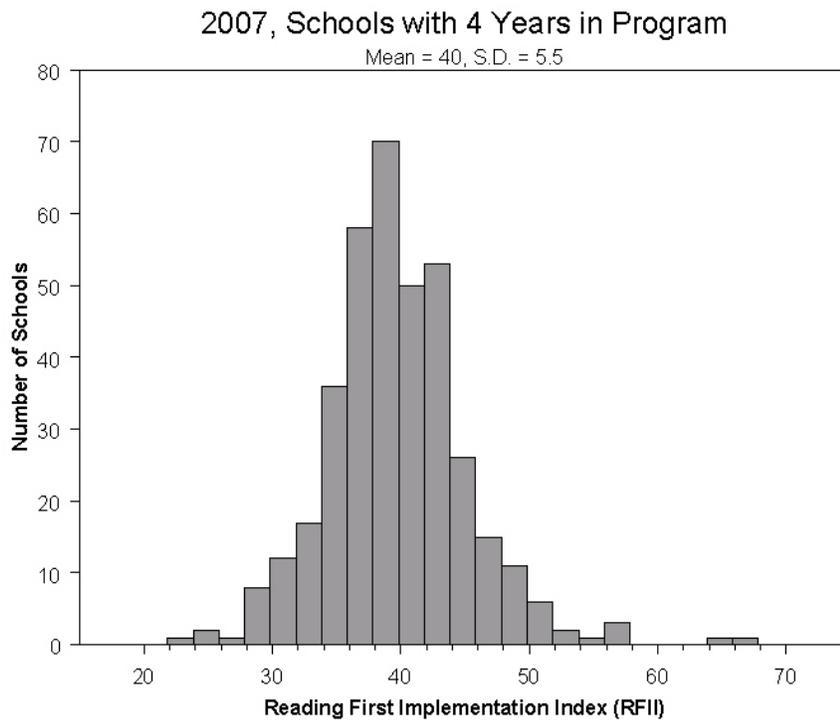
Figure 3.1: All Schools – 2007 Reading First Implementation Index (RFII), Distribution of Schools
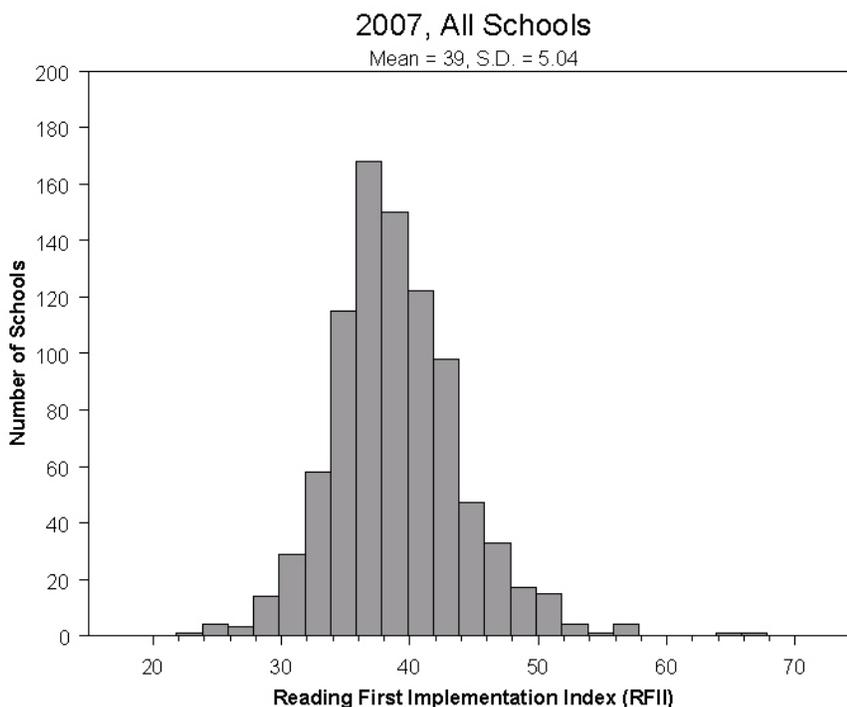
Degree of implementation is likely to develop over time. Schools with more years in the program (Years In Program, or YIP) may have different implementation profiles than schools newer to the program. The histograms in Figures 3.2 through 3.6 show the distribution of the RFII for schools participating in Reading First for different lengths of time. Measures on the far right tail of the distributions of the figures (above 55) should be viewed with caution; such schools tend to show unusually high inter-respondent agreement, possibly suggesting coached answers or collaborative group completion. In the histograms, some patterns are evident. First, there is a consistent pattern of the distribution across YIPs and all are similar to the All Schools graph in Figure 3.1. Schools in the program only one or two years (YIP 1 and YIP 2) have modes slightly lower than schools with more years in the program. YIP 2 schools show a bimodal pattern, which repeats a bimodal distribution observed in the Year 3 Report for schools that were at that time (2005) in YIP 2. (In 2006, the distribution of those schools, then in YIP 3, coalesced to a single-mode distribution, presumably because the lower mode "caught up" with the rest of the cohort.) It appears that bimodal distributions may be a recurring characteristic of schools in the second year of implementation; they divide into "fast adopters" and "slow adopters." YIP 1 schools, a smaller number of schools altogether, seem to require additional time to build their level of implementation. YIPs 3, 4, and 5 seem most closely to resemble the All Schools distribution.
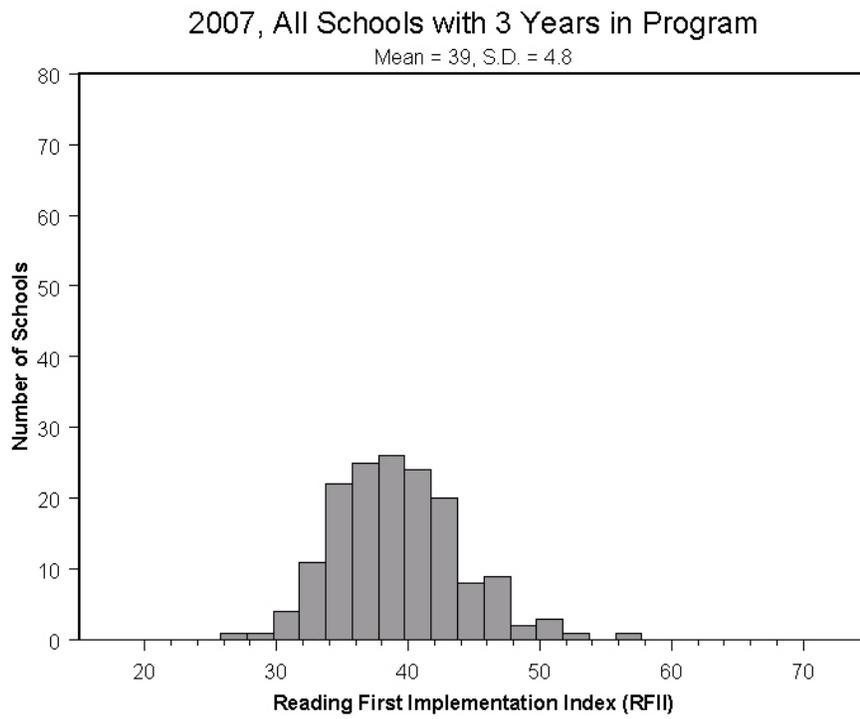
Figure 3.2: YIP = 5 –2007 Reading First Implementation Index (RFII), Distribution of Schools
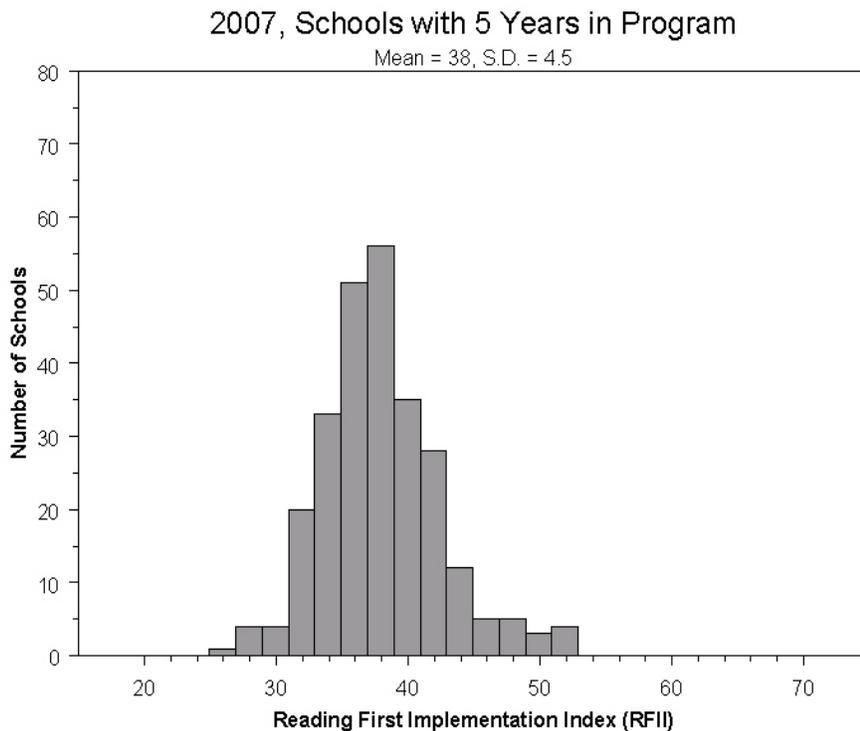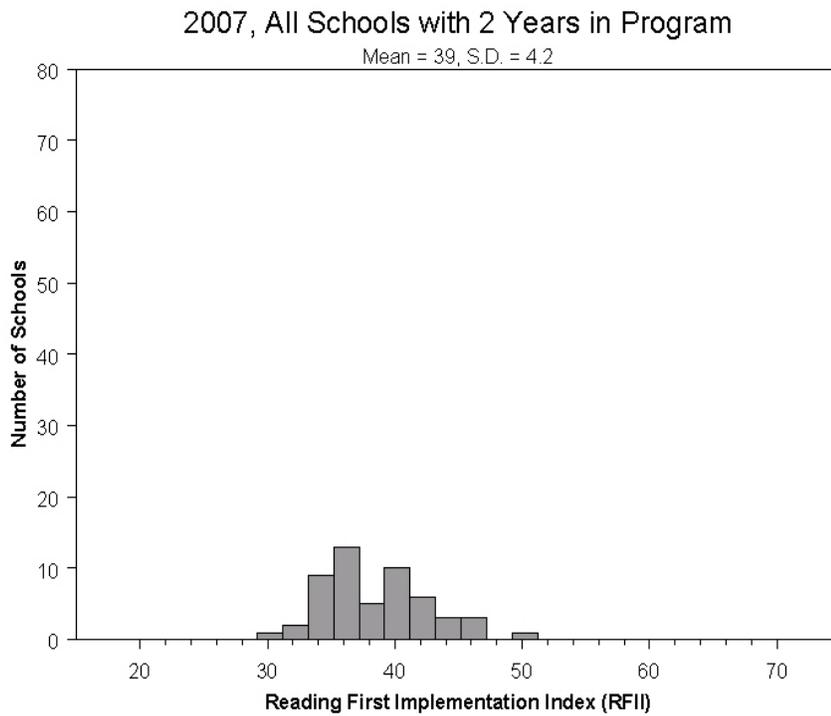

Figure 3.3: YIP = 4 –2007 Reading First Implementation Index (RFII), Distribution of Schools

Figure 3.4: YIP = 3 –2007 Reading First Implementation Index (RFII), Distribution of Schools
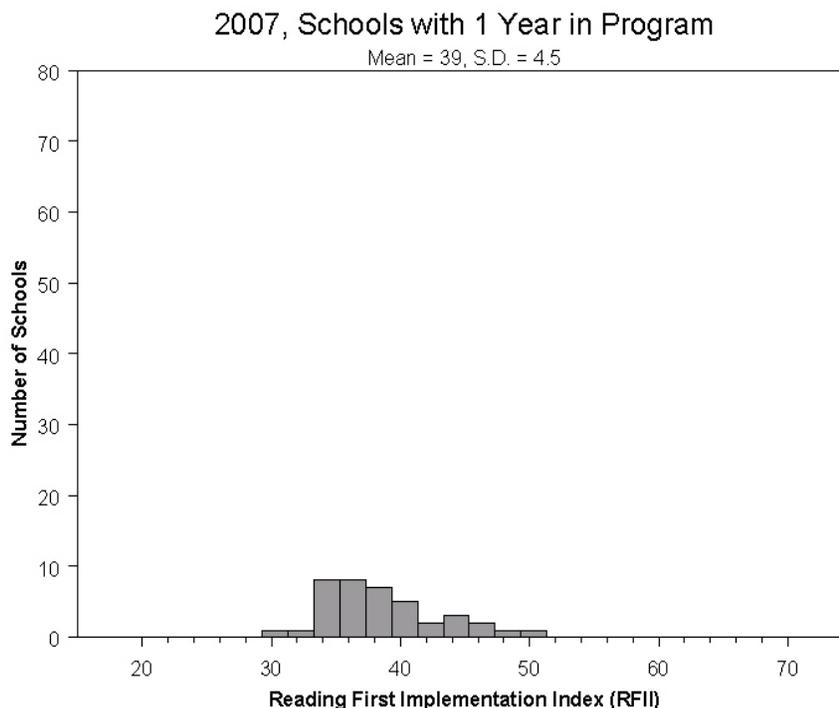

2007, All Schools with 3 Years in Program
Mean = 39, S.D. = 4.8

Figure 3.5: YIP = 2 –2007 Reading First Implementation Index (RFII), Distribution of Schools

2007, All Schools with 2 Years in Program
Mean = 39, S.D. = 4.2

Figure 3.6: YIP = 1 –2007 Reading First Implementation Index (RFII), Distribution of Schools[5]

Dimensions of Implementation

Table 3.1 shows the dimensions derived in the RFII calculation process, their means and standard deviations for each year from 2004 to 2007, and the mean and standard deviation of the RFII for each of those years. The 2007 RFII had a mean of 39 and a standard deviation of 5. In the bottom row, we see the mean school RFIIs for 2004, 2005, 2006 and 2007 for all schools in the Reading First population, with standard deviations (labeled "Plus or Minus"). A list of 18 dimensions and the number of items that comprised each dimension are also reported along with their means and standard deviations for each year. Three of these dimensions, set in bold type, were used to calculate the RFII. The means in the columns by year may be interpreted as the average percent of times (items) that teachers rated their school "more than adequate" on that dimension, averaged across schools. This is the same standard used for the RFII.

For this report, we introduce a new column in Table 3.1 that shows the degree to which each dimension correlates with the 2007 RFII. (Correlations range from -1.00 to +1.00, where 0.00 means there is no relationship at all.) Thus, if a school wishes to increase its RFAI most expeditiously, it should focus on those dimensions which: a) are low relative to the state average; and b) have a high correlation with the

---

[5] Note that YIP 1 schools only have one year in the program and, therefore, averaging the 2006 and 2007 RFII was not possible. Only the 2007 RFII was calculated.

RFII. In interpreting the dimensions, note that some are contained within others. For instance, "School Implementation Overall" is composed of items from all the implementation dimensions.

Table 3.1: All Schools, N (2007) = 885, Mean and "Plus or Minus" for Each Dimension, 2004-2007[1, 2, 3, 4]

	Dimension	# Items, 2007	% of the time teachers rated their school "More than Adequate"								Correlation with 2007 RFII
			2004		2005		2006		2007		
			Mean	Plus or Minus	Mean	Plus or Minus	Mean	Plus or Minus	Mean	Plus or Minus	
1	Teacher Professional Development	9	38	14	36	10	35	10	34	9	0.48
2	Coach Professional Development	7	58	22	56	21	48	22	33	21	0.17
3	Principal Professional Development	3	48	30	46	30	57	30	56	30	0.16
4	**Teacher Coach Professional Development**	**11**	**40**	*16*	**37**	*11*	**35**	*10*	**34**	*10*	**0.45**
5	Evaluation of Professional Development	5	11	6	14	7	15	9	15	8	0.48
6	Implementation, Assurances	11	44	18	48	16	46	16	45	17	0.55
7	School Implementation, Materials	175	36	10	37	9	41	11	41	10	0.53
8	School Implementation, Instruction	28	34	6	36	6	40	6	40	6	0.79
9	**School Implementation Overall**	**210**	**39**	*7*	**40**	*6*	**43**	*7*	**43**	*7*	**0.96**
10	Coaching Implementation	32	46	16	48	14	50	12	49	12	0.68
11	Teacher Implementation	33	48	5	50	5	54	5	54	5	0.62
12	Teacher RF Understanding	17	27	6	29	5	30	6	30	5	0.34
13	Coach RF Understanding	17	36	15	39	14	38	8	39	8	0.42
14	Principal RF Understanding	17	17	9	19	10	20	6	20	6	0.40
15	**Overall RF Understanding**	**17**	**23**	*5*	**25**	*5*	**26**	*5*	**26**	*5*	**0.34**
16	Teacher RF Evaluation	4	14	7	14	7	16	9	15	8	0.58
17	Coach RF Evaluation	6	20	18	19	18	24	21	23	20	0.32
18	Principal RF Evaluation	6	23	24	24	24	23	20	20	19	0.24
19	**RF Implementation Index (RFII)**	**238**	**36**	*6*	**36**	*5*	**39**	*6*	**39**	*5*	**1.00**

[1] Dimensions 4, 9, and 15 are in bold because they are weighted contributors to Dimension 19, the RFII. The 2007 statistics are across 885 schools from the point of view of teachers for dimensions 1, 4, 5, 7, 8, 9, 10, 11, 12, 15, 16, and 19. Dimensions 2, 13, and 17 are from the point of view of coaches. Dimensions 3, 14, and 18 are from the point of view of principals. Dimension 6 is from the point of view of coaches and principals together. The 2004 statistics are across 628 schools; the 2005 statistics across 808 schools; and the 2006 statistics are across 856 schools.

[2] The statistics in the right column report the dimension's correlation with the RFII. The closer to 1.00, the more it captures what is meant by "implementation" as embodied by the RFII.

[3] The N-count of schools in this table does not exactly match those from all Reading First schools as reported in Chapters 1 and 2. The schools reported here are those whose teachers, coaches, and principals returned surveys.

[4] The phrase "Plus or Minus" refers to the average distance from the mean of all the measures that went into the mean. This is also known as the Standard Deviation.

Conclusions

Are Schools Implementing "Adequately"?   To interpret the implementation data, we rely on the procedures developed in prior reports that validate the RFII as a satisfactory measure of implementation. The RFII serves as a comparative benchmark for examining implementation by every school in the Reading First program. The RFII of an individual school can be viewed relative to some standard reference point that characterizes the population of schools as a whole.   In the first year of implementation, the average RFII was 36. This became the cut-point – somewhat arbitrary – between "High Implementation" schools and "Low Implementation" schools.   This distinction was used in conjunction with school achievement measures in other chapters to track the different achievement trend-lines for high implementing and low implementing Reading First schools (see Chapter 4 of the Year 4 Report and Chapter 2 of the Year 5 Report). To preserve comparability over time, the 36 as a cut-point continues to be used to define the upper boundary of the lower implementing schools. However, based on advice in 2007 from the EAG, the "High Implementation" schools have been redefined to be at least one standard deviation above 36 – a new cut-point of 41.4. This has the benefit of sharpening the distinction between high and low implementing schools, but at the cost of leaving out schools that are in the mid-range between 36 and 41.4.

Because the cut-point of 36 has over the course of the evaluation been used to distinguish high from low implementing schools, it serves as a reasonable definition of the lower bound of "Adequate."[6] By that criterion, the histograms and Table 3.1 above reveal that schools are on average doing an "adequate" job of implementing the Reading First program, since the mean 2007 RFII of 39 is greater than 36 by half a standard deviation.

Examining the mean RFII over time, it appears that the index has risen. In 2004 and 2005, the mean RFII was 36 while in 2006 and 2007, it was 39. It stands to reason that program integrity would increase over time and the rise in the RFII statistic supports that conclusion.

---

[6] Note, however, that this usage of the term "adequate" differs fundamentally from that used in previous reports. In the Year 4 Report and earlier, "adequate" was defined in a manner parallel to "more than adequate" – i.e., as a teacher's propensity to score a school in or above the "adequate" rating scale category for each item. While psychometrically defensible, this definition has proven needlessly confusing and is here replaced with a simpler "cut-point based" definition that is in harmony with how implementation is conceptualized in the achievement section of the evaluation.

References

Dane, A. V. & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? Clinical Psychology Review, 18(1), 23-45.

Gresham, F., Gansle, K. A., & Noell, H. (1993, January 1). Treatment Integrity in Applied Behavior Analysis with Children. *Journal of Applied Behavior Analysis, 26*(2), 257.

Leinhardt, G., Zigmond, N., & Cooley, W. (1981). Reading instruction and its effects. *American Educational Research Journal, 18*, 343-361.

Linacre, J. M. (1994), Many-Facet Rasch Measurement, Chicago: MESA Press. Also visit www.rasch.org/rmt/rmt162h.htm.

Power, T., Blom-Hoffman, J., Clarke, A., Riley-Tillman, T., Kelleher, C., & Manz, P. (2005). Reconceptualizing Intervention Integrity: A Partnership-Based Framework for Linking Research with Practice. *Psychology in the Schools*, *42*(5), 495-508.

Ruiz-Primo, M. (2006, February 1). A Multi-Method and Multi-Source Approach for Studying Fidelity of Implementation. CSE Report 677. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*, (ERIC Document Reproduction Service No. ED492864) Retrieved October 30, 2007, from ERIC database.

Schiller, E. (2001). Policy Brief I: Using Implementation Data to Study State, District, and School Impacts, Bethesda, MD: Abt Associates Inc.