The Challenge of Scaling Up Educational Reform

Findings and Lessons from First Things First

Final Report

Janet Quint
Howard S. Bloom
Alison Rebeck Black
LaFleur Stephens

with

Theresa M. Akey

July 2005

Principal funding for First Things First comes from the Institute of Education Sciences, U.S. Department of Education. Additional support to supplement the core project comes from the Ford Foundation, the Bill and Melinda Gates Foundation, the William T. Grant Foundation, and the Ewing Marion Kauffman Foundation. A grant from the Pew Charitable Trusts for MDRC's research methodology initiatives was an important source of funding for the Classroom Observation Study.

Dissemination of MDRC publications is also supported by the following foundations that help finance MDRC's public policy outreach and expanding efforts to communicate the results and implications of our work to policymakers, practitioners, and others: The Atlantic Philanthropies; the Alcoa, Ambrose Monell, Bristol-Myers Squibb, Ford, Grable, and Starr Foundations; and the Open Society Institute.

The findings and conclusions in this report do not necessarily represent the official positions or policies of the funders.

For information about MDRC and copies of our publications, see our Web site: www.mdrc.org.

Overview

First Things First (FTF) is a major comprehensive school reform that includes three central components: *small learning communities* of up to 350 students and their key teachers who remain together for several years; a *family advocate system*, in which each student is paired with a staff member who meets regularly with the student, monitors his or her progress, and works with the student's parents to promote success; and *instructional improvement efforts* aimed at making lessons more engaging and rigorous, as well as better aligned with state and local standards.

FTF was initially launched in Kansas City, Kansas, and subsequently tested in 12 middle schools and high schools in four additional districts (Houston, Texas; the Riverview Gardens School District in suburban St. Louis County, Missouri; and Greenville and Shaw, Mississippi) through the Scaling Up First Things First Demonstration, a five-year research and demonstration project supported by the Institute of Education Sciences in the U.S. Department of Education. The scaling-up project was a collaboration of two organizations: the Institute for Research and Reform in Education (IRRE), which developed the program model and provided support and technical assistance to partner schools and districts, and MDRC, which evaluated the initiative. This report describes the implementation and effects of the program model in these five districts, all of which serve high proportions of minority and economically disadvantaged students.

With respect to implementation, the researchers found that FTF has evolved continuously not only at the sites but also in the minds of its developers, as IRRE personnel have learned from both successes and challenges. Implementation progressed further in settings where district and school leaders provided consistent support for the initiative and IRRE staff offered intensive technical assistance; predictably, changes in structure (for example, the creation of small learning communities) took hold more easily than changes in instruction.

The impacts of FTF were measured using a comparative interrupted time-series design. In summary, the key findings are:

- Middle and high school students in Kansas City, Kansas, registered large gains on a wide range of academic outcomes that were sustained over several years and were pervasive across the district's schools; similar gains were not present in the most comparable schools in the state. The improvements occurred over the course of eight years of substantial effort by the school district and IRRE to implement FTF as the district's central educational reform.

- It is not yet clear whether the expansion sites, which had operated FTF for two or three years at the time of the research follow-up, will replicate the impressive findings for Kansas City.

Contents

Overview iii
List of Tables and Figures vii
Preface xiii
Acknowledgments xv

Executive Summary ES-1

1 Introduction 1
The Initiative's Program Model and Theory of Change 4
The Study Districts and Schools 8
Data Sources 11
The Scope and Contents of This Report 11

2 Implementing First Things First 13
Introduction and Key Findings 13
Change and Stability of the FTF Model and Implementation Process 14
Implementing FTF in Kansas City, Kansas 21
Implementing FTF at the Expansion Sites 22
Variation Among Expansion-Site Districts and Schools: Factors Influencing Implementation 28
Implementing the Key Components 34
Small Learning Communities 36
The Family Advocate System 41
Instructional Improvement 43

3 Changes in Support and Engagement Among Students and Teachers 49
Support and Engagement Among Teachers 54
Support and Engagement Among Students 56
Interpreting the Findings 62

4 The Impacts of First Things First on Student Outcomes 65
Estimating Impacts 65
FTF in Kansas City, Kansas 73
FTF in Houston, Texas 91
FTF in Riverview Gardens, Missouri 110
FTF in the Delta Region of Mississippi 128
Conclusions 134

5 Reflections and Lessons 135
District Support 136
Extended Follow-Up 137
Balancing Personalization and Instructional Improvement 138
Intensive Technical Assistance 140
School E 141

Appendixes

A Measuring the Implementation of First Things First 143
B Support and Engagement Among Teachers and Students: Results for Individual Schools 151
C Estimating the Impacts of First Things First 157
D Supplementary Tables for Chapter 4 175

References 209

Previous Publications on Scaling Up First Things First 211

List of Tables and Figures

Table		Page
ES.1	Estimated Impact of First Things First on Student Test Scores: Kansas City, Kansas	ES-5
ES.2	Estimated Impact of First Things First on Student Test Scores: Replication Sites	ES-7
1.1	School Districts and Secondary Schools Implementing First Things First	3
1.2	The Seven Critical Features of First Things First	7
1.3	Selected Characteristics of First Things First Schools in the Planning Year	9
2.1	Evolution of the Key Elements of First Things First	15
2.2	Extent of Implementation of First Things First: Across Schools	26
2.3	Extent of Implementation of First Things First: Individual Schools and Districts	29
2.4	Leadership and Technical Assistance in the First Things First Sites	30
2.5	Extent of Implementation of Key Dimensions of First Things First	35
3.1	Survey Items Measuring Support and Engagement Among Students	51
3.2	Survey Items Measuring Support and Engagement Among Teachers	52
3.3	Teachers' Average Scale Scores: Support and Engagement	55
3.4	Percentage of Teachers in High and Low Categories of Support and Engagement: All Schools	57
3.5	Percentage of Teachers in High and Low Categories of Support and Engagement: 2001 Cohort Schools	58
3.6	Students' Average Scale Scores: Support from Teachers and Engagement	59
3.7	Percentage of Students in High and Low Categories of Support from Teachers and Engagement: All Schools	60
3.8	Percentage of Students in High and Low Categories of Support from Teachers and Engagement: 2001 Cohort Schools	61
4.1	High School State Assessment Test Scores for First Things First Schools, Comparison Schools, and the State: Kansas City, Kansas	75
4.2	Middle School State Assessment Test Scores for First Things First Schools, Comparison Schools, and the State: Kansas City, Kansas	85
4.3	High School Test Results for First Things First Schools, Comparison Schools, the District, and the State: Houston, Texas	93

Table		Page
4.4	Middle School Test Results for First Things First Schools, Comparison Schools, the District, and the State: Houston, Texas	109
4.5	High School MAP Test Scores for First Things First School, Comparison Schools, and the State: Riverview Gardens, Missouri	122
4.6	Middle School MAP Test Scores for First Things First Schools, Comparison Schools, and the State: Riverview Gardens, Missouri	124
4.7	High School SATP Test Scores for First Things First Schools, Comparison Schools, and the State: Delta Region of Mississippi	130
A.1	Structural and Functional Dimensions of Implementation to Be Measured	145
B.1	Teachers' Average Scale Scores for Individual Schools: Support	153
B.2	Teachers' Average Scale Scores for Individual Schools: Engagement	154
B.3	Students' Average Scale Scores for Individual Schools: Support from Teachers	155
B.4	Students' Average Scale Scores for Individual Schools: Engagement	156
C.1	Outcome Measures for the First Things First Impact Analysis	160
C.2	Selection of Comparison Schools	165
D.1	Estimated Impact of First Things First on 11th-Grade State Reading Tests: Kansas City, Kansas	176
D.2	Estimated Impact of First Things First on 10th-Grade State Math Tests: Kansas City, Kansas	178
D.3	Estimated Impact of First Things First on High School Attendance, Dropout, and Graduation Rates: Kansas City, Kansas	180
D.4	Estimated Impact of First Things First on 8th-Grade State Reading Tests: Kansas City, Kansas	182
D.5	Estimated Impact of First Things First on 7th-Grade State Math Tests: Kansas City, Kansas	184
D.6	Estimated Impact of First Things First on Middle School Attendance Rates: Kansas City, Kansas	186
D.7	Estimated Impact of First Things First on the Percentage of 10th-Grade Students Passing the TAAS/TAKS in Reading and Math: Houston, Texas	187
D.8	Estimated Impact of First Things First on the Percentage of 10th-Graders Scoring At/Above the 50th Percentile and At/Below the 25th Percentile on the SAT-9 in Reading: Houston, Texas	189

Table		Page
D.9	Estimated Impact of First Things First on the Percentage of 10th-Graders Scoring At/Above the 50th Percentile and At/Below the 25th Percentile on the SAT-9 in Math: Houston, Texas	191
D.10	Estimated Impact of First Things First on High School Attendance Rates and 9th-Grade Persistence Rates: Houston, Texas	193
D.11	Estimated Impact of First Things First on the Percentage of 8th-Grade Students Passing the TAAS/TAKS in Reading and Math: Houston, Texas	195
D.12	Estimated Impact of First Things First on the Percentage of 8th-Graders Scoring At/Above the 50th Percentile and At/Below the 25th Percentile on the SAT-9 in Reading: Houston, Texas	197
D.13	Estimated Impact of First Things First on the Percentage of 8th-Graders Scoring At/Above the 50th Percentile and At/Below the 25th Percentile on the SAT-9 in Math: Houston, Texas	199
D.14	Estimated Impact of First Things First on Middle School Attendance Rates: Houston, Texas	201
D.15	Estimated Impact of First Things First on High School State Test Scores, Attendance Rates, Dropout Rates, and Graduation Rates: Riverview Gardens, Missouri	203
D.16	Estimated Impact of First Things First on Middle School State Test Scores and Attendance Rates: Riverview Gardens, Missouri	205
D.17	Estimated Impact of First Things First on 10th-Grade State Test Scores in English II: Delta Region of Mississippi	206
D.18	Estimated Impact of First Things First on 9th-Grade State Test Scores in Algebra: Delta Region of Mississippi	207

Figure		
1.1	The Initiative's Theory of Change	5
2.1	Teachers' Responses to the Critical Features of First Things First in the Planning Year	24
4.1	Design Diagrams for the Impact Analysis	68
4.2	Changes from Quasi-Baseline Levels in the Percentage of 11th-Graders Scoring Proficient or Unsatisfactory on the State Reading Test: Kansas City, Kansas	77
4.3	Changes from Quasi-Baseline Levels in the Percentage of 10th-Graders Scoring Proficient or Unsatisfactory on the State Math Test: Kansas City, Kansas	80

Figure		Page
4.4	Changes from Quasi-Baseline Levels in High School Attendance, Dropout, and Graduation Rates: Kansas City, Kansas	82
4.5	Changes from Quasi-Baseline Levels in the Percentage of 8th-Graders Scoring Proficient or Unsatisfactory on the State Reading Test: Kansas City, Kansas	86
4.6	Changes from Quasi-Baseline Levels in the Percentage of 7th-Graders Scoring Proficient or Unsatisfactory on the State Math Test: Kansas City, Kansas	88
4.7	Changes from Quasi-Baseline Levels in Middle School Attendance Rates: Kansas City, Kansas	90
4.8	Changes from Baseline Levels in the Percentage of 10th-Graders Passing the TAAS/TAKS in Reading and Math for the 2001 Cohort High School (School E): Houston, Texas	96
4.9	Changes from Baseline Levels in the Percentage of 10th-Graders Scoring At/Above the 50th Percentile on the SAT-9 in Reading and Math for the 2001 Cohort High School (School E): Houston, Texas	98
4.10	Changes from Baseline Levels in High School Attendance Rates and 9th-Grade Persistence Rates for the 2001 Cohort High School (School E): Houston, Texas	100
4.11	Changes from Baseline Levels in the Percentage of 10th-Graders Passing the TAAS/TAKS in Reading and Math for the 2002 Cohort High Schools (Schools F and G): Houston, Texas	103
4.12	Changes from Baseline Levels in the Percentage of 10th-Graders Scoring At/Above the 50th Percentile on the SAT-9 in Reading and Math for the 2002 Cohort High Schools (Schools F and G): Houston, Texas	105
4.13	Changes from Baseline Levels in High School Attendance Rates and 9th-Grade Persistence Rates for the 2002 Cohort High Schools (Schools F and G): Houston, Texas	107
4.14	Changes from Baseline Levels in the Percentage of 8th-Graders Passing the TAAS/TAKS in Reading and Math for the 2001 Cohort Middle School (School S): Houston, Texas	112
4.15	Changes from Baseline Levels in the Percentage of 8th-Graders Scoring At/Above the 50th Percentile on the SAT-9 in Reading and Math for the 2001 Cohort Middle School (School S): Houston, Texas	114
4.16	Changes from Baseline Levels in Middle School Attendance Rates for the 2001 Cohort Middle School (School S): Houston, Texas	116
4.17	Changes from Baseline Levels in the Percentage of 8th-Graders Passing the TAAS/TAKS in Reading and Math for the 2002 Cohort Middle Schools (Schools U, V, and T): Houston, Texas	117

Figure		Page
4.18	Changes from Baseline Levels in the Percentage of 8th-Graders Scoring At/Above the 50th Percentile on the SAT-9 in Reading and Math for the 2002 Cohort Middle Schools (Schools U, V, and T): Houston, Texas	119
4.19	Changes from Baseline Levels in Middle School Attendance Rates for the 2002 Cohort Middle Schools (School U, V, and T): Houston, Texas	121
4.20	Changes from Baseline Levels in the Percentage of High School Students Scoring in the Bottom Two Categories of the MAP State Test: Riverview Gardens, Missouri	123
4.21	Changes from Baseline Levels in High School Attendance, Dropout, and Graduation Rates: Riverview Gardens, Missouri	125
4.22	Changes from Baseline Levels in the Percentage of Middle School Students Scoring in the Bottom Two Categories of the MAP State Test: Riverview Gardens, Missouri	127
4.23	Changes from Baseline Levels in Middle School Attendance Rates: Riverview Gardens, Missouri	129
4.24	Changes from Quasi-Baseline Levels in the Percentage of 10th-Graders Passing English II: Delta Region of Mississippi	132
4.25	Changes from Quasi-Baseline Levels in the Percentage of 9th-Graders Passing Algebra: Delta Region of Mississippi	133
C.1	Design Diagrams for the Impact Analysis	159

Preface

Nearly four years after the passage of the federal No Child Left Behind Act, students, schools, and school districts continue to struggle with meeting high expectations for performance in times of fiscal uncertainty. In the past couple of years, in particular, greater political and public focus has centered on low-performing high schools, where many students fail to graduate and where those who do are often unprepared for the challenges of postsecondary education and work.

This report offers important findings on the scaling up of First Things First, a comprehensive school reform that seeks major changes in school structure, instruction, and accountability and governance policies. The study examines the implementation and effects of First Things First in both middle schools and high schools.

The authors of this study tell two stories — one of success and the other of the challenges of replicating success. First, this study corroborates earlier research showing that high school and middle school academic outcomes improved substantially in Kansas City, Kansas, the first site where First Things First was implemented. This success came after years of focused support from school district leaders and intense technical assistance from the Institute for Research and Reform in Education, the developer of the reform initiative.

The second story — of the replication of First Things First — confirms what the Kansas City experience shows: Expecting success in the short term is sure to disappoint, especially when trying to boost student achievement. In the brief, one- to three-year follow-up period for the expansion sites (in Houston, suburban St. Louis, and the Mississippi Delta), the pattern of results was mixed but included some glimmer of hope. The findings indicate that improving instruction is a particularly difficult aspect of education reform but may be central to the goal of increasing student achievement.

This study demonstrates that developing a successful school reform model is possible. However, once accomplished, repeating success requires intense effort, consistent leadership, and the patience to allow the interventions time to work.

Gordon L. Berlin
President

Acknowledgments

As this five-year endeavor draws to a close, the authors wish to thank a number of people whose insights and assistance were central to the completion both of the project as a whole and of this report in particular. First and foremost, the evaluation would not have been possible without the cooperation of administrators, teachers, and students at the districts and schools participating in the Scaling Up First Things First Demonstration. They were willing to share their experiences and reflections in both interviews and surveys and to subject their work to scrutiny. While these individuals are far too numerous to acknowledge individually, we are especially grateful to Harry Selig of the Houston Independent School District and to Dan Wright of the Kansas City, Kansas, Public Schools for their assistance in providing data central to the impact analysis in their respective sites.

On-site field researchers Thelma Collins, Hines Cronin, Belita Leal, and Marianne Wilson displayed the special combination of inquisitiveness and sensitivity that is essential to collecting high-quality information. Linda Kuhn of Survey Research Management administered teacher and student surveys at the expansion-site schools and monitored the preparation of the resulting data files. The surveys themselves were designed by Carolyn Eldred.

Robert Granger, now President of the William T. Grant Foundation, was instrumental to the project's early development and has provided ongoing support. Earlier work by Michelle Gambone helped the researchers conceptualize this evaluation. At the Institute for Research and Reform in Education, James Connell and Laurie Levin supplied important insights from the developers' perspective and offered detailed comments on two drafts of the report. The assistance of Susan Bloom, Julie Broom, Linda Gerson, Freida Inmon, and Adena Klem is also gratefully acknowledged. Phyllis Blumenfeld at the University of Michigan played a central role in helping MDRC researchers design the observational study of classroom instruction, and Teresa McMahon assumed responsibility for analysis of the observational data.

Numerous MDRC staff members, present and past, were important to our work on this project. We are especially grateful to Fred Doolittle for his guidance and ongoing support of the effort over many years. Corinne Herlihy and William Corrin modeled collegiality in weekly project meetings and had many helpful suggestions. Marla Sherman managed the numerous activities involved in readying staff and student surveys for administration, assisted by Shirley Campbell and by Shirley James and her capable staff. D. Crystal Byndloss played a key role in conducting field research, and Angela Estacion guided the classroom observation study, and the work of Amy Karwan and Veronica Fellerath helped to inform the evaluation design. Julian Brash, Rasika Kulkarni, Bernice Melamud, Judith Scott, Nickisha Stephenson, and Laura Sztejnberg played important roles in programming and otherwise preparing the data for analy-

sis. Rebecca Kleinman prepared many of the tables and figures. James Kemple and Jason Snipes offered searching and careful reviews of earlier drafts of the report, and Gordon Berlin's comments were also extremely helpful to shaping the overall message. Glee Holton provided support throughout the process. Vivian Mateo contributed her skills in creating figures and tables. John Hutchins and Amy Rosenberg offered many useful editorial suggestions, and Stephanie Cowell prepared the report for publication. First Vannett Davy and then Mario Flecha helped the authors remain organized throughout.

The Authors

Executive Summary

Introduction

This report on First Things First — a major comprehensive school reform — arrives at an opportune moment, when President George W. Bush, the nation's governors, and business and foundation leaders have announced a renewed commitment to reforming American high schools. Now operating in more than 70 schools in nine districts across the country, First Things First (FTF) seeks to improve low-performing schools by strengthening relationships between teachers and students and by making classes more engaging and rigorous. FTF was initially launched in Kansas City, Kansas, and subsequently tested in 12 middle schools and high schools in four additional districts through the Scaling Up First Things First Demonstration, a five-year research and demonstration project supported by the Institute of Education Sciences in the U.S. Department of Education. The project was a collaboration of two organizations: the Institute for Research and Reform in Education (IRRE), which developed the program model and provided support and technical assistance to partner schools and districts, and MDRC, which evaluated the initiative.

This report, the last of four produced by MDRC, describes the implementation and effects of the program model in these five districts, all serving high proportions of minority and economically disadvantaged students. It complements and updates a report on the FTF program in Kansas City, Kansas.[1] In summary, the key findings of this study are:

- Middle and high school students in Kansas City, Kansas, registered large gains that were sustained over several years and were pervasive across the district's schools; similar gains were not present in the most comparable schools in the state. The improvements occurred over the course of eight years of substantial effort by the school district and IRRE to implement FTF as the district's central educational reform.

- It is not yet clear whether the expansion sites, which had operated FTF for two or three years at the time of the research follow-up, will replicate the impressive findings for Kansas City.

[1]Michelle Gambone, Adena Klein, Jean Summers, Theresa Akey, and Cynthia Sipe, *Turning the Tide: The Achievements of the First Things First Education Reform in the Kansas City, Kansas, Public School District* (Philadelphia: Youth Development Strategies, Inc., 2004).

What Is First Things First?

FTF entails major changes in school structure, instruction, and accountability and governance policies. The model includes three components:

- **Small learning communities.** In this initiative, small learning communities (SLCs) contain groups of up to 350 students and their core-subject and other key teachers who remain together for several years. They are organized around broad themes (for example, "Science and Technology" and "Performing Arts") that are meant to inform instruction and provide the SLCs with unique identities.

- **Family Advocate System.** Each student is paired with a staff member — generally a teacher in the student's SLC — who is expected to meet regularly with the student and monitor his or her academic, social, and emotional progress. The advocate is responsible for assisting the student, creating a more positive relationship between the school and the student's family, and working with parents to promote their child's academic success.

- **Instructional improvement efforts.** Teachers work with their colleagues to align curricula with state and local standards, and they participate in professional development activities designed to help them learn, practice, and regularly use strategies that make classroom instruction rigorous and engaging.

How Was FTF Evaluated?

To assess program implementation, the report draws on a combination of quantitative data from teacher and student surveys and qualitative findings from classroom observations and interviews with administrators, teachers, students, and others. While a random assignment design — considered the "gold standard" for evaluating program impacts — was not feasible for this study, MDRC used a rigorous research method, called a "comparative interrupted time-series analysis," to estimate the effect of FTF. In principle, the impact of FTF on a student outcome equals the *difference* between what that outcome was after the school reform was under way and what it would have been without the reform (the "counterfactual"). In practice, one can estimate this difference by comparing the change over time in a student outcome for schools that adopted FTF with the corresponding change for similar schools that did not adopt the reform, and variants of this approach were used for each of the five sites in the evaluation. Ideally, the evaluation design that is used to produce impact estimates should comprise data on consistently measured student outcomes for multiple pre-intervention baseline years, multiple post-intervention follow-up years, multiple FTF schools, and multiple comparison schools that are closely matched with the FTF schools. Unfortunately, this set of ideal conditions did not exist in any of the study sites; in-

stead, the evaluators had to make the best of the data that were available, while remaining mindful of the limitations of the resulting analyses.

How Was FTF Implemented?

FTF began operations in Kansas City, Kansas, in 1998-1999 in one of the district's four comprehensive high schools, along with that high school's feeder middle and elementary schools. The three remaining comprehensive high schools and their feeder schools were added to the program over the next two years. The "expansion," or "scaling-up," sites discussed in this report were phased in over a two-year period, beginning in 2001-2002. They include three high schools and four middle schools in Houston, Texas; a high school and its two feeder middle schools in the Riverview Gardens School District in suburban St. Louis County, Missouri; and two high schools in Greenville and Shaw, Mississippi.

FTF is a complex reform whose implementation requires change at every level. It demands much both of personnel in the schools and districts mounting the reform and of the staff of IRRE, who are responsible for guiding and assisting local efforts. The following key findings emerge from MDRC's analysis of the initiative's implementation in these various sites.

- **FTF has evolved continuously not only at the implementation sites but in the minds of its developers, as IRRE personnel have learned from both successes and challenges.**

For example, the Family Advocate System was not added to the mix of program elements until 2000-2001. IRRE's role in providing technical assistance in the area of instructional improvement increased considerably, and that assistance became more comprehensive and systematic over time.

- **Predictably, changes in structure took hold more quickly and more easily than changes in instruction; the instructional improvement efforts associated with FTF were implemented most fully in Kansas City, Kansas.**

While the creation of small learning communities was relatively easy and popular among teachers and students alike, changing teachers' instructional practices proved challenging. Central office support for instructional improvement in Kansas City helped ensure progress in this area. Over time, teachers at the expansion sites followed their Kansas City counterparts, moving forward in aligning curricula and assessments with state standards and in making greater use of active engagement strategies in their lessons.

- **District and school leadership and outside technical assistance were the key determinants of implementation success at the expansion sites.**

Implementation progressed further in settings where district and school leaders provided consistent support for the initiative, where the principal and School Improvement Facilitator (a school district employee working at each school to guide implementation of the reform) had a cooperative and mutually respectful relationship, and where IRRE staff offered intensive technical assistance.

Did FTF Make a Difference for Student Outcomes?

MDRC looked at the impact of FTF both in its original site — Kansas City, Kansas — and in the later, scaling-up districts. The key findings follow.

- **In Kansas City, Kansas, high school and middle school academic outcomes improved substantially as FTF was implemented, while similar trends were not observed in comparison schools, pointing to the initiative's central role in improving academic performance.**

These academic outcomes included increased rates of student attendance and graduation, reduced student dropout rates, and improved student performance on the Kansas state tests of reading and mathematics. As Table ES.1 shows, the estimated effects on student test scores reflect double-digit increases in the percentage of students who scored at levels deemed "proficient" or above by the state and double-digit reductions in the percentage of students who scored at levels deemed "unsatisfactory."

For example, on the most recent state reading test (for spring 2004), FTF high schools experienced an 11.1 point relative gain in the percentage of student scores that were proficient or above. In other words, the increase in the percentage from its initial level three years earlier was 11.1 points greater for FTF high schools than for their comparison schools. Even larger relative improvements were observed for the percentage of student scores that were unsatisfactory. In spring 2004, this percentage had dropped by 15.5 points more for FTF high schools than for comparison schools. Findings for FTF middle schools indicate a relative increase of 13.7 points in the percentage of student scores that were proficient or above and a relative decline of 13.6 points in the percentage of scores that were unsatisfactory. Thus, overall, there was a pronounced and consistent pattern of relative improvements in the reading performance of FTF high school students and middle school students.

Findings for math in Table ES.1 reveal correspondingly large and consistent improvements for middle schools. By spring 2004, FTF middle schools had experienced a 9.6 point increase (relative to their comparison schools) in the percentage of math scores that were proficient or above and a 9.0 point relative decrease in the percentage of scores that were unsatisfactory. Math scores for FTF high schools also showed signs of improvement (with relative declines

The First Things First Evaluation

Table ES.1

Estimated Impact of First Things First on Student Test Scores: Kansas City, Kansas

	Spring 2002[a]	Spring 2003	Spring 2004
	Impact on Percentage Proficient[b]		
High schools			
11th-grade reading test	6.9	10.2 **	11.1 **
10th-grade math test	1.2	3.4	-4.4 *
Middle schools			
8th-grade reading test	3.0	23.1 ***	13.7 ***
7th-grade math test	5.0	11.0 ***	9.6 **
	Impact on Percentage Unsatisfactory[b]		
High schools			
11th-grade reading test	-5.4	-11.1 **	-15.5 ***
10th-grade math test	-10.8 ***	-6.7 **	-5.2
Middle schools			
8th-grade reading test	-5.4	-22.3 ***	-13.6 ***
7th-grade math test	-7.3 *	-13.1 ***	-9.0 **

NOTES: Sample includes students from four FTF high schools and eight FTF middle schools.

The "impact" was calculated as the difference between the change from the quasi-baseline level for FTF schools in spring 2001 and the corresponding change over time for comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]Spring 2002 is the fourth year of implementation for three schools, the third year of implementation for three schools, and the second year of implementation for six schools.

[b]"Proficient" is defined as the sum of the top three (of five) performance categories on the Kansas state test. "Unsatisfactory" is defined as the bottom performance category on the Kansas state test. Improved student performance is represented by a relative increase in the percentage proficient and a relative decrease in the percentage unsatisfactory.

in the percentage of scores that were unsatisfactory). However, there was no clear pattern over time in changes in the percentage of scores that were proficient or above.

These positive effects — plus those for other high school and middle school outcomes — reflect improvements that, in many cases, were sustained over time and occurred at numerous schools. Therefore, despite the inability to select comparison schools that closely matched Kansas City's exceptionally low-performing schools (an issue that is discussed further in the

report), the multiplicity and magnitude of the improvements that were observed for FTF schools in this district support the conclusion that the reform model was critical to causing them.

- **It is not yet clear whether the expansion sites will replicate the robust findings for Kansas City.**

Findings for the FTF replication districts are less pronounced or less consistent than those for the reform model's original district. As Table ES.2 illustrates, estimates of impacts on state test scores are almost never as large as those for Kansas City; they vary markedly across districts; and they are seldom statistically significant. The lack of statistical significance reflects, in part, the small number of schools in the replication study from each district and, thus, the limited statistical precision of the study to detect impacts at these schools.

One of the largest urban high schools where implementation was most complete has registered positive effects on student achievement. There are some suggestive signs of success at other schools as well; however, the overall pattern of findings leaves considerable uncertainty about how much improvement in student performance was produced by the reform.

It is important to note several limitations of the analyses presented in this report, which make them a conservative test of the program's effects. First, as noted, the statistical precision for measuring impacts at a single school or a few schools (as is the case in most of the expansion sites) is often too limited to identify with confidence impacts other than those that are exceptionally large. Second, in the current educational environment, there are strong pressures on all schools to improve, so that the outcomes of FTF are measured against those of comparison schools that may also be trying to change. Third, because expansion sites began their efforts recently, there is only a brief window of time through which to view their success. Finally, because at some sites the benchmark used to gauge improvement is a "quasi-baseline" year after implementation had already begun, any impacts produced before or during this year are "netted out" of the analysis and thus are not attributed to the initiative.

What Are the Policy Implications of This Study?

This report tells a complex story about a complicated initiative. The implementation findings indicate that mounting the intervention is hard; doing it well requires commitment, persistence, and effort. The positive effects of FTF in its home district of Kansas City, Kansas, were sizable, pervasive, and sustained. So far, the schools participating in the scaling-up demonstration, with one exception, have not registered similar effects.

What does this say about FTF — and about school reform efforts more generally? The experience in Kansas City, Kansas, points to four conditions that were *sufficient* to produce

The First Things First Evaluation

Table ES.2

Estimated Impact of First Things First on Student Test Scores: Replication Sites

	Follow-Up Year 1	Follow-Up Year 2	Follow-Up Year 3
	Impact on Percentage Passing[b]		
Houston, Texas[a]			
High schools			
10th-grade reading test	-1.1	6.6	8.8 *
10th-grade math test	-3.3	4.2	7.0
Middle schools			
8th-grade reading test	-1.6	-5.1 **	-1.9
8th-grade math test	2.5	1.8	6.5
Delta Region of Mississippi			
High schools			
10th-grade English test	12.9	8.4	
9th-grade algebra test (Shaw High School)	-10.0	-3.3	
9th-grade algebra test (Greenville-Weston High School)	15.8	-15.6	
	Impact on Percentage of Low Performers[c]		
Riverview Gardens, Missouri			
High school			
11th-grade communication arts test	5.6	-7.1	1.5
10th-grade math test	-2.8	-7.6	-10.0
Middle schools			
7th-grade communication arts test	3.7	6.5	0.0
8th-grade math test	-4.5	-9.5	-7.1

NOTES: Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

The "impact" was calculated as the difference between the change from the baseline or quasi-baseline level for FTF schools and the corresponding change over time for comparison schools.

[a]In follow-up Years 1 and 2, impacts presented are the average of impacts for three high schools and four middle schools. In follow-up Year 3, impacts presented are for one high school and one middle school.

[b]Improved student performance is represented by a relative increase in the percentage passing.

[c]Low-performing students are those scoring in the bottom two proficiency categories established by the State of Missouri. Improved student performance is represented by a relative decrease in the percentage of low performers.

meaningful impacts on a wide array of outcomes in secondary schools serving disadvantaged populations. Whether these conditions are also *necessary* remains an open question.

1. **A districtwide focus, with the district's staying the course for many years in its provision of pressure and supports for the reform's changes**

From the outset, the Kansas City, Kansas, school district took ownership of FTF as its major school reform initiative. In contrast, at the scaling-up sites, the school districts did not provide similarly consistent support and oversight to the schools mounting the initiative. The experience of the successful Houston high school suggests that lack of strong district support may be offset if exceptionally strong school-level leadership is in place.

2. **Schools that had operated FTF for many years when their impacts were measured**

The findings of this report are consistent with other research indicating that comprehensive school reforms are more effective when they have been in place for at least five years. At the time of the research follow-up, FTF had been in operation for a much longer period in Kansas City than in the expansion sites. Even the latest-starting Kansas City schools had been in operation for four years when the last impact data were collected. On the other hand, the follow-up data at the expansion sites reflect a maximum of only two or three years of experience operating the intervention.

3. **Balancing a need for more personalized learning environments with a comprehensive and intensive approach to improving instruction that emphasizes alignment, rigor, and student engagement**

The FTF experience suggests that striking such a balance is not easy. For instance, the expansion schools were able to implement small learning communities quickly, during the demonstration's planning year. By all accounts, they enabled teachers and students to develop closer relationships with each other. In contrast, early instructional improvement efforts in the scaling-up sites were much less systematic than in Kansas City.

4. **Intensive and responsive technical assistance from providers who are willing to make midcourse adjustments where needed**

IRRE was on the scene in Kansas City, Kansas, throughout the planning period and early implementation years of FTF in the district. But IRRE's capacity was stretched by the addition of so many new sites at once and by the subsequent expansion into two more large urban districts that are not part of this report.

The experiences of operating FTF in Kansas City, Kansas, and in the expansion sites suggest that school reform is too difficult to expect results without long-term support from high-level leaders and without sufficient technical assistance. The success of the model in Kansas City points to the critical role that districts play in providing a unified message and the pressure and support that all educators need to keep their eyes on the prize: better teacher-student relationships and improved teaching and learning in the classroom.

Chapter 1

Introduction

This report considers the implementation and impacts of First Things First, one of the major initiatives aimed at changing low-performing high schools that emerged during the last decade. The report arrives at an opportune moment, when President George W. Bush, the nation's governors, and business and foundation leaders have announced a renewed commitment to reforming American high schools, especially schools serving large numbers of low-income students and students of color. Critics have noted the large size and anonymity of many such schools, their poor working conditions, the lower qualifications and inexperience of many teachers, and the rarity of rigorous, challenging instruction. First Things First (FTF) — a comprehensive school reform initiative currently operating in over 70 schools in nine districts across the country — attempts to combat these problems by focusing on building strong relationships, improving teaching and learning, and reallocating resources to meet those first two goals. This report discusses the program's implementation and impacts in the first five districts to launch the initiative.

Designed by the Institute for Research and Reform in Education (IRRE) — headed by James P. Connell, a developmental psychologist — FTF includes changes in school structure, instructional practices, and accountability and governance that are aimed at making schools more engaging places for students and adults alike and at improving students' academic performance. Implementation of these changes is intended to require only modest and temporary increases in resources. The model is based on research conducted by Connell and others on the factors making for high engagement and high achievement among adolescents, the literature on organizational change and effective educational practices, and the experiences of schools that have succeeded with students who might otherwise be at high risk of school failure [1]

FTF was first launched in Kansas City, Kansas, a city with a largely low-income, nonwhite population of some 150,000 people that is situated across the Wyandotte and Missouri Rivers from Kansas City, Missouri. Planning for FTF in Kansas City began at the district level in 1996; officials decided to adopt the initiative districtwide and to phase it in over several years. The first of the district's four comprehensive high schools, along with that high school's feeder middle and elementary schools, began planning for FTF during the 1997-1998 school year and started implementation in the 1998-1999 school year. The 1998-1999 school year also marked the planning year for a second high school "cluster," which began implementation in 1999-2000. The two remaining clusters began planning in 1999-2000 and implementation in 2000-2001. Positive early outcomes in this pioneering district led IRRE to seek to test the initiative in other locations

[1] See, for example, Connell and Wellborn, 1991; Skinner, Zimmer-Gembeck, and Connell, 1998.

through the Scaling Up First Things First Demonstration, a research and demonstration project supported by the Institute of Education Sciences in the U.S. Department of Education.

The demonstration project represents a collaboration of two organizations. IRRE has provided support and technical assistance to the participating schools and districts through its own small core staff and a network of experienced practitioners and consultants. Working with many districts at once presented a new challenge for the organization, which had previously concentrated its efforts in a single location. IRRE has also produced reports and guides directed toward school and district administrators that discuss the practical and policy issues involved in implementing the initiative. MDRC has provided oversight for the project as a whole and has studied the program's implementation and impacts at the expansion sites that are part of the scaling-up effort. Building on an evaluation of FTF in Kansas City, Kansas, that was conducted by Youth Development Strategies, Inc. (YDSI), MDRC has further analyzed FTF's impacts at its original home site and added a year of follow-up to the YDSI study.[2]

The new schools and districts — referred to in this report as the "expansion," or "scaling-up," sites or locations — comprise secondary schools in a variety of urban, suburban, and rural settings.[3] These new schools were phased in over a two-year period, in two groups. Schools in the earlier-implementing group began planning for the initiative during the 2000-2001 academic year and started implementation during the 2001-2002 school year; consequently, throughout this report, they are referred to as the "2001 cohort" schools. They include a high school and middle school in Houston, Texas; a high school and its two feeder middle schools in the Riverview Gardens School District in suburban St. Louis County, Missouri; and two high schools in Greenville and Shaw, Mississippi, located in the Mississippi Delta.[4] The second group of schools — which started planning and implementation activities one year after the first group and is therefore referred to as the "2002 cohort" — includes two additional high schools and three middle schools in Houston.

Table 1.1 lists the five school districts and the secondary schools that are discussed in this report. The experiences of these schools reflect some common circumstances and some unique challenges.

---

[2]Two reports by an independent evaluator discuss the implementation and impacts of FTF in Kansas City, Kansas (Gambone, Klem, Moore, and Summers, 2002; Gambone et al., 2004). The present report relies on these two studies for information on the implementation of FTF in Kansas City.

[3]Formally, an impact evaluation of the last two clusters of secondary schools to implement in Kansas City, Kansas, is also supported under the Scaling-Up First Things First Demonstration. For ease of reference, however, in this report, the term "scaling-up site" is restricted to schools or districts outside Kansas City, Kansas.

[4]During the 2001-2002 academic year, the two high schools in Greenville, Mississippi — Greenville High School and Weston High School — merged to form one high school with two campuses, now known as Greenville-Weston High School. Unlike previous MDRC reports on FTF, which examined the two campuses separately, this report treats Greenville-Weston High School as one school.

The First Things First Evaluation

Table 1.1

Schools Districts and Secondary Schools Implementing First Things First

Kansas City (KS) Public Schools

Wyandotte High School
Central Middle School
Northwest Middle School

Washington High School
Arrowhead Middle School
Eisenhower Middle School

Harmon High School*
Argentine Middle School*
Rosedale Middle School *

Schlagle High School*
Coronado Middle School*
West Middle School*

Houston (TX) Independent School District

Lee High School*†
Sharpstown Middle School*†

Sam Houston High School*††
Sharpstown High School*††
Fondren Middle School*††
Fonville Middle School*††
Welch Middle School*††

Riverview Gardens (MO) School District

Riverview Gardens High School*†
Central Middle School*†
East Middle School*†

Greenville (MS) Public Schools

Greenville-Weston High School*†

Shaw (MS) Public Schools

Shaw High School*†

SOURCES: IRRE and MDRC documents.

NOTES: *Denotes an expansion site under the OERI Scaling Up First Things First contract.
†Denotes a 2001 cohort school: planning year 2000-2001; implementation year 2001-2002.
††Denotes a 2002 cohort school: planning year 2001-2002; implementation year 2002-2003.

This is the final report in a series of MDRC publications about FTF.[5] Earlier reports focused on the program's planning and early implementation and on classroom instruction at the expansion sites. This document has three main goals:

- To present new findings on program impacts in Kansas City, Kansas — findings that constitute an especially important addition to the impact story, because they show the effects of FTF at schools that have attained a considerable degree of operational maturity
- To carry forward the implementation analysis and to present early findings on program impacts at the expansion sites, bearing in mind that these impacts represent the effects of newly minted programs
- To explore the factors contributing to these findings and helping to explain different patterns of findings

The remainder of this introductory chapter considers the theory of change that underlies FTF, presents data on the schools that are included in the study, and discusses the data sources and organization of the report.

The Initiative's Program Model and Theory of Change

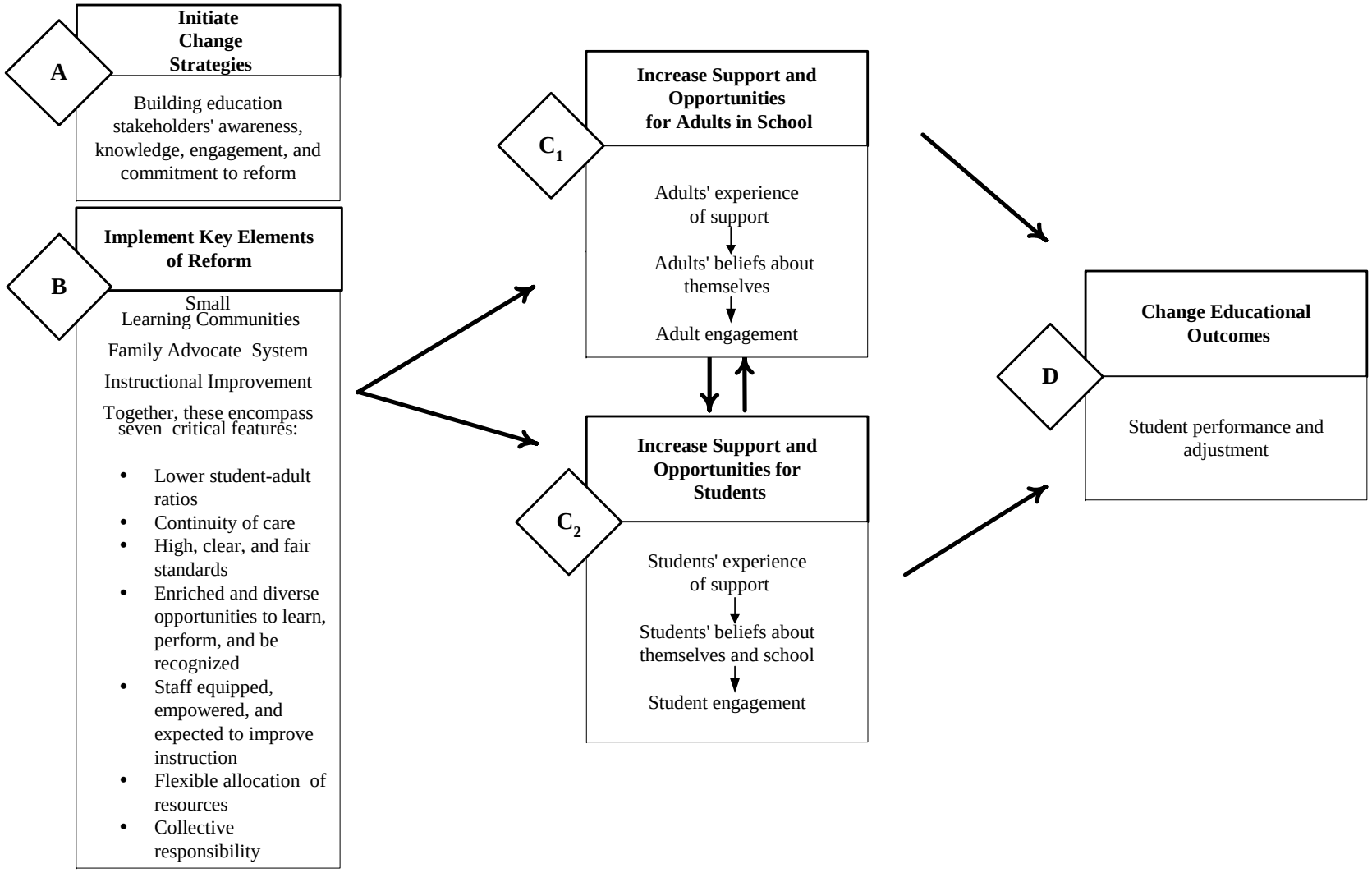
At the core of the FTF initiative is a research-based theory of change that articulates how and why the intervention is expected to increase student achievement. A key premise underlying the theory is that humans have fundamental needs to feel *competent*, to feel *autonomous*, and to feel *related*. That is, people need to feel that they can act in ways that will produce desired effects, that they can make independent choices, and that they are securely attached to important others. Two further premises are that positive development is facilitated by social contexts that meet these fundamental needs and that there are specific elements within these contexts that support or hinder such development.

Figure 1.1 illustrates the FTF theory of change. Creating a commitment to change through exposure to the change strategies (Box A in the figure) is the first step in the theory, both logically and temporally. Thus, implementing whole-school change requires that key stakeholders in the community, the school districts, and the schools themselves perceive a need to change. It also calls for a clear understanding of the change that is sought and an intense and sustained commitment on the part of administrators, teachers, and others to pursuing that change.

[5]See Quint, 2002; Quint, Byndloss, and Melamud, 2003; Estacion, McMahon, and Quint, 2004.

Figure 1.1

The Initiative's Theory of Change



The theory holds that, by implementing three key elements, impersonal, low-performing schools can transform themselves into environments that satisfy the fundamental needs of both students and teachers. These three elements, shown in Box B of Figure 1.1, are:

- **Small learning communities (SLCs).** In FTF, these are groups of up to 350 students and their core-subject and other key teachers who remain together for several years. SLCs are organized around broad themes (for example, "Science and Technology," "Performing Arts") that inform instruction and provide the SLCs with unique identities. SLC teachers meet regularly to discuss their students' progress and problems and to explore ways of making instruction more rigorous and engaging.

- **Family Advocate System.** Each student is paired with a staff member — generally a teacher in the student's SLC — who is expected to meet regularly with the student and monitor his or her academic, social, and emotional progress and advocate on his or her behalf. The advocate is also expected to maintain contact with the student's family, serving as the key liaison between the family and the school and helping to engage families in the education of their children. An important aspect of the component is the Family Advocate Period, a specific time reserved for students and staff to meet in a group setting.

- **Instructional improvement efforts.** Teachers are expected to work with their colleagues to learn, practice, and regularly use strategies that make classroom instruction rigorous and engaging.

Together, these three key elements encompass seven principles, referred to in program parlance as the "critical features" of FTF. These critical features are shown in abbreviated form in Figure 1.1. and are elaborated in Table 1.2. It is worth pointing out that the reform principles are not original or unique to FTF. They are found, singly or in combination, in many whole-school reform initiatives and thus can be taken as reflecting the best current thinking about the aspects of schools that make them most conducive to learning. What FTF brings to schools, as discussed later in this report, is not merely a set of elements and principles but also a set of strategies for putting them in place.

Putting in place these key elements, it is hypothesized, will increase feelings of support and engagement among teachers and students (shown as Boxes C1 and C2 of Figure 1.1). Small learning communities and the Family Advocate System are intended to create strong, caring teacher-student relationships as well as strong collegial relationships among SLC teachers. As teachers come to care more about their students' academic success, they will be more motivated to adopt instructional practices that help their students succeed, especially because their fellow

The First Things First Evaluation

Table 1.2

The Seven Critical Features of First Things First

Structural changes

1. Lower student-adult ratios to 15:1 during language arts and math classes for at least 10 hours per week.

2. Provide continuity of care across the school day, across the school years, and between school and home by forming small learning communities. The same core group of eight to ten professionals stays with the same group of 150-350 students for extended periods during the school day for all three years of middle school and for at least two-year periods in high school. The Family Advocate System is also aimed at ensuring continuity of care between staff of the small learning communities and students' families.

Instructional changes

3. Set high, clear, and fair academic and conduct standards that define clearly what all students will know and be able to do by the time they leave high school and at points along the way. Performance on standards-based tests is linked directly to students' advancement and grading, drives curriculum and instruction in all courses, and is discussed regularly with students and their families. Adults and students agree on conduct standards, which are reinforced by adults modeling positive behaviors and attitudes and which are sustained by clear benefits to students and adults for meeting them and consequences for violating them.

4. Provide enriched and diverse opportunities to learn, by making learning more active and connected in safe and respectful learning environments; to perform, by linking assessment strategies that use multiple modes of learning and tie performance directly to standards; and to be recognized, by creating individual and collective incentives for student achievement and by providing leadership opportunities in academic and nonacademic areas.

5. Equip, empower, and expect all staff to improve instruction by creating a shared vision and expectation of high-quality teaching and learning in all classrooms; supporting small learning communities' implementation of research-based instructional strategies to fulfill that vision; and engaging all staff in ongoing study to improve curricular and instructional approaches.

Accountability and governance changes

6. Allow for flexible allocation of available resources by teams and schools, based on instructional and interpersonal needs of students. Resources include people (students and staff); instructional facilities; time for instructional planning and professional development; and discretionary funds.

7. Assure collective responsibility by providing collective incentives and consequences for small learning communities, schools, and central office staff that are linked to change in student performance.

SOURCE: IRRE documents.

teachers will support and encourage their instructional improvement efforts. For their part, students who feel that their teachers support and care about them, and who find their classes more interesting and more challenging, will feel more autonomous and more confident about their own abilities and will engage more fully in their coursework. Such "engagement" entails a belief that doing well is personally important and includes a set of behaviors and feelings that back up that belief and put it into practice (for example, trying hard, preparing for class, paying attention, taking responsibility, and avoiding anger and blame when academic setbacks occur).

The vertical arrows connecting Boxes C1 and C2 in both directions indicate that there are reciprocal influences between increased supports and opportunities for students and for adults. Changes in one promote changes in the other, and vice versa. For example, teachers may modify their instruction in ways that promote student engagement, and such engagement will encourage teachers to strengthen and broaden their commitment to instructional improvement.

Increased student engagement, in turn, is seen as a critical antecedent to the intervention's desired long-term outcomes: higher scores on tests measuring academic achievement, better attendance, and improvements in other academic outcomes (Box D in the figure).

The Study Districts and Schools

The hope of improved student scores on high-stakes tests attracted many school districts to FTF. In selecting districts and schools to participate in the Scaling-Up First Things First Demonstration, IRRE specifically sought out sites whose demographic characteristics would make it more likely that the schools involved would benefit from the kinds of reforms that FTF offers. All the schools had to serve a substantial percentage of economically disadvantaged young people, and the schools had to be large enough to be divided into several SLCs. Site selection criteria also involved the developers' subjective judgments of local administrators' will and capacity to undertake major reforms.

Table 1.3 presents selected characteristics of the Kansas City, Kansas, and the scaling-up schools studied in this report. As the table illustrates, the number of students enrolled in the study schools differed greatly from school to school. However, all the schools served predominantly nonwhite students: mostly African-American students in Riverview Gardens and the Mississippi schools and mostly Hispanic or African-American students in the Kansas City and Houston schools. The majority of these students were poor, as evidenced by their eligibility for free or reduced-price lunches. And many students at all schools exhibited low academic achievement: For example, during the FTF planning year, 62 percent of tenth-grade students across the three Houston high schools mounting the initiative scored at or below the 25th percentile on the reading part of the SAT-9, a nationally normed standardized test of achievement,

The First Things First Evaluation

Table 1.3

Selected Characteristics of First Things First Schools in the Planning Year

District and School	FTF Planning Year	Total Enrollment	Ethnicity (%)		Eligible for Free/ Reduced-Priced Lunch (%)	Percentage of Students with Low Achievement on Standardized Tests[b]			
			Black	Hispanic		Math (State)	Reading (State)	Math (National)	Reading (National)
Kansas City, Kansas[a]									
Wyandotte High School	1997-1998	1,273	64.7	11.7	73.0	79.0	64.2	NA	NA
Washington High School	1998-1999	1,152	61.0	3.1	50.6	71.0	54.9	NA	NA
Harmon High School	1999-2000	1,190	28.6	33.8	65.2	69.5	44.7	NA	NA
Schlagle High School	1999-2000	1,067	74.7	2.3	58.2	78.8	47.4	NA	NA
Central Middle School	1997-1998	699	32.6	32.6	91.6	70.0	40.5	NA	NA
Northwest Middle School	1997-1998	443	98.0	1.1	89.2	77.1	62.2	NA	NA
Arrowhead Middle School	1998-1999	497	48.5	5.8	48.3	35.8	32.3	NA	NA
Eisenhower Middle School	1998-1999	752	68.1	2.7	58.8	50.2	36.5	NA	NA
West Middle School	1999-2000	429	69.2	1.6	76.5	73.2	36.5	NA	NA
Coronado Middle School	1999-2000	431	68.7	5.3	73.3	65.8	31.3	NA	NA
Argentine Middle School	1999-2000	555	27.4	40.2	76.8	60.0	37.4	NA	NA
Rosedale Middle School	1999-2000	500	33.2	30.2	80.2	54.6	31.4	NA	NA
Houston, Texas									
Lee High School	2000-2001	2,564	12.3	73.5	62.7	27.8	37.3	49.7	69.1
Sam Houston High School	2001-2002	3,139	6.3	87.8	68.8	9.7	11.8	40.4	64.7
Sharpstown High School	2001-2002	2,188	33.5	48.8	57.4	15.7	13.8	42.1	48.7
Sharpstown Middle School	2000-2001	1,656	24.6	61.6	78.5	11.5	8.9	39.7	36.2
Fonville Middle School	2001-2002	1,231	7.5	85.8	91.4	12.7	10.9	56.3	51.9
Fondren Middle School	2001-2002	1,453	58.9	35.5	77.6	18.9	13.5	47.3	37.5
Welch Middle School	2001-2002	1,724	63.0	29.2	59.6	5.9	3.5	24.2	30.4
Riverview Gardens, Missouri									
Riverview Gardens High School	2000-2001	1,715	89.6	0.1	50.9	90.9	56.4	NA	NA
Central Middle School	2000-2001	820	86.3	0.1	61.0	84.3	47.9	NA	NA
East Middle School	2000-2001	356	99.4	0.0	87.2	84.6	71.3	NA	NA

Table 1.3 (continued)

District and School	FTF Planning Year	Total Enrollment	Ethnicity (%) Black	Ethnicity (%) Hispanic	Eligible for Free/ Reduced-Priced Lunch (%)	Percentage of Students with Low Achievement on Standardized Tests[b] Math (State)	Reading (State)	Math (National)	Reading (National)
Greenville, Mississippi									
Greenville-Weston High School	2000-2001	1,667	99.3	0.1	87.8	38.9	43.5	NA	NA
Shaw, Mississippi									
Shaw High School	2000-2001	314	99.4	0.0	95.2	6.0	52.6	NA	NA

SOURCES: In Kansas City, test score data were obtained from individual student records from a Kansas state data file. For other outcomes in Kansas City, school-level records were obtained from Kansas City, Kansas, Public Schools. In Houston, individual student records were obtained from the Houston Independent School District. In Missouri, school-level records were obtained from the Missouri Department of Secondary and Elementary Education. In Mississippi, school-level records were obtained from the Mississippi Department of Education.

NOTES: [a]While the planning year is as shown, the test score data presented are from the 2000-2001 school year, the baseline year for MDRC's study of program impacts.

[b]The data presented for the percentage of students with low achievement represent, in Kansas City, the percentage of 8th- and 11th-graders failing the reading section and the percentage of 7th- and 10th-graders failing the math section of the Kansas State Assessment. In Houston, the data represent the percentage of 8th- and 10th-graders failing the reading and math sections of the Texas Assessment of Academic Skills (TAAS). For national tests, the data shown represent the percentages of students scoring at the 25th percentile or less. In Riverview Gardens, the data represent the percentage of 7th- and 11th-graders scoring in the bottom two proficiency categories of the reading section and the percentage of 8th- and 10th-graders scoring in the bottom two proficiency categories of the math section of the Missouri Achievement Program (MAP). In Mississippi, the data represent the percentage of 10th-graders failing the reading section and the percentage of 9th-graders failing the math section of the Subject Area Testing Program (SAT-9).

and 44 percent scored at or below the 25th percentile on the math part of the test. Similarly, during the planning year, 56 percent of eleventh-grade students at Riverview Gardens High School scored in the bottom two categories of the state's test of communication arts while 91 percent of tenth-graders scored in the bottom two categories of the state math test. Clearly, IRRE selected schools where improved academic achievement was badly needed.[6]

Data Sources

This report draws on a combination of quantitative and qualitative data. The quantitative data on student outcomes — test scores, attendance, persistence, and graduation — used in the impact analysis come both from individual student records data maintained by the districts and from aggregate school data obtained from states' department of education Web sites. Wherever possible, student outcome data collection began several years prior to the introduction of FTF; in all cases, it continued through the 2003-2004 school year. Quantitative data on support and engagement among teachers and students come from surveys administered each spring at all the scaling-up schools outside Kansas City, beginning during the school's planning year and continuing through 2004.

The qualitative data largely reflect the efforts of the field researchers who worked for MDRC at the expansion sites between the fall of 2000 and the end of the 2002-2003 school year (in the Mississippi sites) or the middle of the 2003-2004 academic year (in Houston and Riverview Gardens). Over the course of the initiative, the field researchers conducted both formal and informal interviews with district officials, school administrators, teachers, and students and observed whole-school and SLC meetings, professional development sessions, and classroom lessons. In addition, MDRC staff members visited the expansion sites to interview district and school leaders. Published data on the school districts and schools rounded out the interview and field notes. In addition, conversations with IRRE staff members and IRRE documents provided the developers' perspective on progress and challenges at the demonstration.

The Scope and Contents of This Report

The organization of this report follows the theory of change presented in Figure 1.1. The report contains five chapters. After this introductory chapter:

[6]As a condition of the U.S. Department of Education funding, IRRE selected rural as well as urban sites. Some of the sites chosen had only one high school in the district. This fact has implications for the ability of the analysis to detect statistically significant impacts on student achievement and other outcomes, as discussed in Chapter 4.

- Chapter 2 examines the extent to which the key elements of the initiative had been implemented by the 2003-2004 school year and explores reasons for variation in implementation among schools and of different program elements.
- Chapter 3 focuses on changes in support and engagement among teachers and students. As noted above, the theory of change posits that increased feelings of support and engagement are important antecedents of improved academic performance.
- Chapter 4 examines the program's impacts on test scores, attendance, and other academic outcomes.
- Chapter 5 concludes the report and reflects on lessons learned and their policy implications.

Chapter 2

Implementing First Things First

Introduction and Key Findings

First Things First (FTF) is a complex reform whose implementation demands much both of personnel in the schools and districts mounting the reform and of staff of the Institute for Research and Reform in Education (IRRE), who are responsible for guiding and assisting local efforts. Issues associated with implementation can be approached from many perspectives. While the chapter discusses IRRE's work with school districts and individual schools, IRRE has written many reports addressing its role in detail.[1] The principal focus of this chapter, instead, is on the experience of the schools in mounting the components of the initiative.

The initiative itself has also changed over time, as IRRE staff have recognized the need for clearer guidelines and enriched supports in some areas. Because this fact is important for understanding the reform's trajectory at the study schools, the next section of this chapter discusses the evolution, both in theory and in practice, of the program's key elements: small learning communities (SLCs), the Family Advocate System, and instructional improvement efforts. Attention then turns to the program's implementation. While FTF's history and development in the Kansas City, Kansas, flagship site were not a focus of the MDRC evaluation, the Kansas City story is important for understanding the impacts achieved by schools in that district and is summarized briefly. The discussion then addresses the extent to which the program as a whole and its key elements had been implemented at the expansion sites by the 2003-2004 school year. (By focusing on this academic year, the implementation analysis parallels the analysis of program impacts, for which 2003-2004 represents the last year of available follow-up.) The extent of variation in implementation among the expansion-site schools and districts participating in the study is then explored. Issues associated with the implementation of specific program elements generally were not restricted to individual schools or districts, and these are discussed in the chapter's final section.

Several key findings emerge from the analysis:

- The FTF program model has evolved considerably over time, with major changes including the introduction of a new component (the Family Advocate System) and a growing role for IRRE in instructional improvement efforts.

[1]See Connell, 2003; Klem, Levin, Bloom, and Connell, 2003; Connell and Broom, 2004; Institute for Research and Reform in Education, 2004.

- FTF was implemented most fully in Kansas City, Kansas, the pioneer site, where it was adopted throughout the school district and where the central office provided considerable support to the initiative.
- Expansion-site schools varied widely in their implementation of the structural and functional dimensions of FTF, although all implemented the program's key elements to some degree.
- Predictably, changes in structure took hold more quickly and more easily than changes in the behavior of administrators, teachers, and others within those structures.
- Expansion-site middle schools, on average, got further in implementing FTF than high schools did, in part because the FTF structural changes were easier to put in place in middle schools.
- Changing teachers' instructional practices proved especially challenging, although, over time, teachers made progress in aligning curriculum and assessments with state standards and made greater use of cooperative learning strategies in their lessons.
- By the end of the study, schools that began program operations in the 2001-2002 school year had not necessarily progressed further in implementing the initiative than those that started implementation a year later.
- Leadership and outside technical assistance were the key determinants of the extent of implementation at the expansion sites. Implementation progressed further in settings where district and school leaders provided consistent support for the initiative, where the principal and School Improvement Facilitator (a school district employee working at each FTF school to guide implementation of the reform) had a cooperative and mutually respectful relationship, and where IRRE staff offered intensive technical assistance.

Change and Stability in the FTF Model and Implementation Process

FTF has evolved continuously not only at the implementation sites but in the minds of its developers, as IRRE personnel have learned and grown from both successes and challenges. Table 2.1 summarizes the evolution of the program model, showing which of the key elements were in place for successive implementation cohorts. It shows that while small learning communities and instructional improvement efforts have been consistent elements throughout the

Table 2.1

Evolution of the Key Elements of First Things First

Implementation Year	Small Learning Communities	Family Advocate System	Instructional Improvement Efforts
1998-1999 (Startup for Kansas City, Kansas, cluster 1)	No prescribed form, but commitment to stay with students over multiple years was expected. Schools could select among several models of SLC structures.	No component existed.	Delivered by KCK school district. District literacy focus announced, with professional development centering on effective strategies for literacy instruction; school literacy coaches designated; weekly early release time and common planning time available for instructional improvement.
1999-2000 (Startup for Kansas City, Kansas, cluster 2)	No prescribed form, but commitment to stay with students over multiple years was expected. Schools could select among several models of SLC structures.	No component existed.	New standards-based curriculum introduced in KCK schools. Guide defining high-quality teaching and learning by district staff and School Improvement Facilitators (SIFs).
2000-2001 (Startup for Kansas City, Kansas, clusters 3 and 4)	No prescribed form, but commitment to stay with students over multiple years was expected. Four-year thematic SLCs and two-year SLCs based on grade level were models presented to schools.	Implemented in several SLCs in KCK middle and high schools.	District professional development focused on engagement strategies. Instructional coaches replaced literacy coaches, but retained literacy focus. In Scaling-Up schools' planning year, emphasis on "read-alouds."
2001-2002 (Startup for seven Scaling-Up schools)	Expansion-site schools were given specific options from which to choose: thematic or nonthematic, two- or four-year communities. KCK leadership prescribed four-year, thematic SLCs for all high schools.	Implemented in additional KCK secondary SLCs and all SLCs in Scaling-Up schools.	In KCK, new secondary math curricula piloted. In Scaling-UP schools, emphasis on engagement strategies. Common planning time, along with early release time in some sites, available for instructional improvement.

(continued)

Table 2.1 (continued)

Implementation Year	Small Learning Communities	Family Advocate System	Instructional Improvement Efforts
2002-2003 (Startup for five Scaling-Up schools)	Three-year (for middle schools) or four-year (for high schools) mixed-grade, thematic SLCs prescribed.	Implemented in additional KCK secondary SLCs and all SLCs in Scaling-Up schools.	In KCK, wider use of new math curricula in middle schools. In Scaling-Up schools, continued emphasis on active engagement strategies. Common planning time along with early release time in some sites, available for instructional improvement.
2003-2004	Three-year (for middle schools) or four-year (for high schools) mixed-grade, thematic SLCs prescribed.	Implemented in almost all secondary SLCs in KCK and all SLCs in Scaling-Up schools. IRRE released Family Advocate Period Activities Guide.	In KCK, wider use of new math curricula in middle schools and high schools; tools for measuring engagement, alignment, and rigor in classrooms piloted with instructional leaders. In Scaling-Up schools, introduction of more comprehensive approach: instructional goals broaened to include engagement, alignment, and rigor; use of common planning time and late start or early release time for peer observation, study of student work, and planning of common assessments within courses of study.

SOURCES: IRRE documents; Gambone, Klem, Moore, and Summers, 2002; Gambone et al., 2004.

history of the initiative, at different times, they have taken different forms. The Family Advocate System, for its part, is a relatively new addition to the mix of program components. Major developments with respect to these three program elements are briefly described below.

Small Learning Communities

When the first cluster of Kansas City, Kansas, schools began planning for FTF in the 1997-1998 school year, school personnel were charged with deciding on a reorganization plan that would allow groups of students — whether in the same grade or mixed grades — to have contact with a small number of adults for longer periods each day and for more than one year. Although there was no specific requirement to implement SLCs, almost all schools in this cluster adopted that option. In 1998-1999, such a requirement was introduced for the remaining Kansas City school clusters of schools planning to mount FTF, but individual schools could decide how their SLCs were to be structured from among several models that IRRE presented to them, although the expectation that students and teachers would stay together over several years remained in force. Considerable variation in SLC structure resulted at the secondary school level: In one school, students could choose from among four-year, theme-based houses; in two schools, SLCs were organized by grade level; and in three schools, students were assigned to SLCs that were balanced by gender and ethnicity.

Based on promising early results from the Kansas City, Kansas, school that had implemented four-year thematic SLCs and on the scheduling and other challenges faced by the schools that had not, IRRE staff members strongly believed that all schools should adopt this structure. But, as in Kansas City, IRRE allowed the 2001 cohort expansion-site schools to decide whether their SLCs should be two- or four-year and whether they should be thematic or not. Teachers at these schools perceived the information that IRRE supplied to them about the benefits of thematic and four-year communities as one-sided, however, and many teachers were left feeling manipulated and ill-used. Much dissatisfaction arose, and distrust of IRRE lingered for months afterward.[2]

With the 2002 cohort expansion-site schools, IRRE took a clear and nonnegotiable position. It simply announced to these schools that SLCs would be theme-based and would extend over all four years of high school and all years of middle school, and school staff members accepted this dictum without protest. Because schools no longer had to spend time deciding about school structure, planning for other aspects of the intervention took place on an accelerated schedule, and in general the year proceeded smoothly. One lesson emerging from the experience is that there is clear value to deciding in advance what is nonnegotiable. If teachers are

[2]See the discussion in Quint (2002, pp. 58ff.), which is briefly summarized here. See also Connell, 2003.

given a say, it should be in decisions where their input is truly sought and where decisions that run contrary to the advice of the developer can be tolerated.

The Family Advocate System

Family advocacy presents another example of the way in which FTF has evolved over time. As Table 2.1 shows, this component was not part of the original program model introduced into the Kansas City, Kansas, school system. Subsequently, however, IRRE staff became aware of a prototype model of family advocacy and saw its relevance to FTF. A small number of Kansas City SLCs began to experiment with a family advocacy component in the 2000-2001 school year, and that year, too, IRRE described family advocacy as an integral part of the program model to schools interested in participating in the Scaling-Up First Things First Demonstration. Before schools initially implemented the Family Advocate System, IRRE staff and consultants provided training on the component, as well as ongoing professional development assistance thereafter. Family advocacy has now been implemented in almost all the SLCs in Kansas City as well.

Instructional Improvement Efforts

Although the FTF model and the thinking behind it have evolved in many ways over the course of the demonstration, that evolution has been most striking in the area of instructional improvement. Some features associated with such improvement — reduced student-teacher ratios; increased instructional time in mathematics and language arts classes; and enriched opportunities for students to learn, perform, and be recognized — have remained constant since the inception of the initiative in the first cohort of Kansas City, Kansas, schools.[3] IRRE's role in providing technical assistance in this area has shifted and increased considerably, however.

When FTF was introduced in the Kansas City schools, IRRE's technical assistance related to instruction largely focused on helping schools arrange students' schedules so that they could spend more time in language arts and math classes. Content and pedagogy were left to the school district, which launched several initiatives to improve instruction. In the 1998-1999 academic year (when the first cluster of schools began operating FTF), the district announced a new literacy initiative, which involved, among other things, the assignment of a literacy coach to each school. The next year, the district adopted a new standards-based curriculum and, at IRRE's suggestion, developed a guide aimed at creating a unified vision by identifying the characteristics of high-quality teaching and learning. The guide, created by district staff and by

[3]IRRE now describes reduced student-teacher ratios as a desirable goal and one to be worked toward but not as an intrinsic part of the program model. At the time the initiative was introduced to teachers at the expansion sites, the prospect of reduced ratios was one of its most appealing aspects.

representatives of the National Education Association (NEA), emphasized the importance of instruction that was challenging and standards-based, that connected with students' experiences, and that engaged students in active learning experiences, including cooperative learning. Much of the professional development offered by the district in the ensuing years centered on active engagement strategies; the district also maintained its focus on literacy.

IRRE also worked to help central office staff develop the capacity to produce different kinds of data on students and to report these data at the SLC level. The data identified students whose attendance and test scores placed them at risk of dropping out, so that teachers could focus additional attention on these students. The new information thus helped SLC members develop a sense of collective responsibility; it also helped SLCs, schools, and the district as a whole set goals for improvement.

Unlike some school reform initiatives, FTF does not include specific high-quality curricula, and although IRRE staff recognized a need to improve instruction at the expansion sites, they did not have a set of coordinated strategies for achieving such improvement at the outset of the scaling-up demonstration.[4] Instead, IRRE turned to consultants who had delivered professional development to the Kansas City, Kansas, schools and with whose work IRRE was therefore familiar. During the 2001 cohort's planning year, IRRE contracted with one group of consultants to conduct a series of workshops on the use of two related instructional strategies: the "read-aloud" and the "think-aloud."[5] In general, the training proved disappointing. Many teachers complained that the consultants were belaboring points that they had already grasped or with which they were already familiar, and math teachers especially said that they could not see how the techniques were relevant to their discipline.

IRRE staff members were aware of the lackluster response to the training and believed that a wider repertory of instructional strategies was needed. During the next year (the first year of implementation for the 2001 cohort schools and the planning year for the 2002 cohort schools), IRRE changed the focus of its technical assistance. It contracted with Kagan Cooperative Learning, Inc. — which had also worked in the Kansas City, Kansas, schools — to provide the expansion schools with training in a set of cooperative learning strategies developed by

[4]FTF curricula for struggling readers and for students making the transition to high school algebra are currently in development. IRRE has also worked to create an on-line resource of high-quality, standards-based curriculum and assessment materials to which teachers at all schools participating in FTF will have access.

[5]In a *read-aloud*, the teacher models fluent reading of fiction or nonfiction passages as a way of engaging students with text, exposing students to the rhythms of the English language, and demonstrating enjoyment or learning from the act of reading. In a *think-aloud*, the teacher models the process of gathering meaning from text — for example, determining the main idea and the author's purpose, using prior knowledge to create new knowledge, and recognizing that reading creates new questions for the reader to answer.

Spencer Kagan and designed to ensure that all students participate actively in learning.[6] By design, these strategies emphasized learning structures rather than specific subject matter content, so that teachers in all disciplines could integrate them into their lessons. Over the next years, staff at the participating schools received several days of initial and follow-up training in the use of these methods; during the summers, some staff members from each school also attended a week of intensive Kagan training in Florida or shorter local and regional training sessions.

While the cooperative learning strategies developed by Kagan required that students take a more active role in the learning process, they did not ensure that lessons would be intellectually challenging or aligned in content with state and local curriculum standards. IRRE's hiring of a new Director of Instructional Supports at the beginning of the 2003-2004 school year signaled the organization's shift to a more comprehensive and more coherent approach to instructional improvement embodied in the acronym "EAR" (engagement, alignment, and rigor). Indeed, the new IRRE staff member was a prime mover of this change; her prior experience as an instructional coach with the Houston Independent School District, as a School Improvement Facilitator (SIF) in an FTF school in Houston, and as an FTF site director made her an ideal candidate for the new position. Under her guidance, and using IRRE-developed protocols, teachers at the Houston schools began to undertake a number of structured activities that together constituted a more systematic approach to instructional improvement. First, teachers observed classes (often taught by teachers outside their SLCs, to make it less stressful for all parties) to look for evidence of engagement, alignment, and rigor and afterwards discussed what they had seen, noting both successful practices and practices needing improvement. Second, teachers in their SLCs met to examine student work presented by a colleague and to discuss how to make that work, and the lesson it reflected, more engaging, aligned, and rigorous. Third, IRRE helped schools develop a structure for staff activities aimed at ensuring greater alignment of course content with state and district standards, along with common assessments and grading standards. Finally, IRRE designed a set of instruments for measuring engagement, alignment, and rigor in classrooms; the instruments were pilot-tested by instructional leaders in Kansas City. Thus, while continuing to contract with others to develop instructional improvement materials and activities, IRRE had itself assumed an instructional leadership role at the expansion sites.[7]

[6]Typically, students are arranged in pairs or small groups to ask each other questions, share opinions, or otherwise reflect on the content of what they are learning.

[7]IRRE has contracted to develop literacy and math curricula specifically designed for students whose academic achievement lags several years behind that of their peers. These curricula are expected to be ready in the fall of 2005 and the spring of 2006, respectively.

Implementing FTF in Kansas City, Kansas

MDRC did not study the implementation of FTF in Kansas City, Kansas. That research was conducted by Youth Development Strategies, Inc. (YDSI), and the discussion below briefly summarizes two richly detailed YDSI reports addressing this subject.[8]

The foremost fact to remember about the implementation of First Things First in Kansas City, Kansas, is that, from the outset, it was adopted and planned as a districtwide reform. FTF arrived in a district whose leaders were looking for a way to reverse negative trends in student achievement. The superintendent had recently completed a series of efforts to create a systematic, data-driven planning and evaluation process that included new standardized tests and revision and promotion of graduation requirements to support higher expectations. After learning about FTF, district leadership saw the reform as a vehicle for synthesizing and promoting the district's efforts to improve.

The school board approved FTF as the cornerstone of its District-Wide Improvement Plan in the fall of 1996, and, for almost a decade now, FTF has continued to receive sustained support from the board, the superintendent, and other central office personnel. Indeed, soon after the district decided to mount the reform, the superintendent who had urged its adoption retired, and an interim superintendent was appointed until a replacement was hired a year later. Ultimately, the school board selected as superintendent a district administrator who had thrown his hat into the ring precisely because he was so committed to the reform. In this way, the board signaled its own ongoing support for the initiative. The long-term commitment of the school board, superintendent, and top-level central office personnel to FTF is itself unusual in a world where the hiring of a new superintendent often signals the advent of a new reform effort.[9]

District leadership established a number of policies and practices aimed at furthering FTF's goals: It instituted weekly early release time for professional development, reassigned curriculum specialists to serve as SIFs, and reconfigured the central office to create clearer lines of oversight and accountability. The three-year schedule for phasing FTF into all district schools was adopted, to maximize the attention and assistance that individual schools in each implementation cohort would receive.

In this process, IRRE came to be viewed as more than an "outside consultant." It offered regular advice to the superintendent and other administrators and became a sounding board for local decisions tied to FTF. IRRE staff visited the district on an almost monthly basis

[8]See Gambone, Klem, Moore, and Summers, 2002; Gambone et al., 2004.

[9]A major Kansas City, Missouri, foundation played an important role in ensuring the initiative's continuity. It not only provided financial support to the initiative but also was involved in ongoing planning and provided assistance from its own research, training, and communications departments.

to monitor the initiative as it unfolded and to provide support to personnel in both the central office and the individual schools implementing or planning for the reform. Among other efforts, IRRE provided the impetus for district officials, SIFs, and others to promulgate standards for high-quality teaching and learning.

Kansas City in many respects served as a testing ground for the expansion sites. IRRE learned a great deal from both the positive implementation lessons that the Kansas City schools offered and the issues that they presented. Thus, as noted above, IRRE's decision to mandate thematic four-year SLCs stemmed in large part from the fact that a Kansas City high school that had selected this model had had positive early outcomes, while schools that had chosen alternative SLC structures had experienced a number of problems. The Family Advocate System was also pretested in some Kansas City SLCs before being incorporated as an integral part of the program model at the scaling-up schools. And to lead professional development sessions for teachers at the expansion sites, IRRE drew on consultants who had led similar workshops in Kansas City as part of the district's broader instructional improvement agenda.

The YDSI analysis of FTF's implementation in Kansas City includes the following major findings:

- Secondary schools in the district were able to achieve reduced student-teacher ratios.
- FTF implementation was associated with a notable increase in the use of small-group teaching strategies in classrooms.
- Students in high schools were more likely to report having opportunities to work in teams on assignments and to work on interdisciplinary projects and projects connected to their futures and their lives outside school.
- Students were more likely to report that their teachers held high expectations for them, that they knew what it took to succeed academically, and that teachers provided them with models of good work.

Implementing FTF at the Expansion Sites

The remainder of the chapter concerns the course of FTF at the expansion sites, which were the focus of MDRC's implementation research. Discussion centers on the first two stages of the program's theory of change: initiating the change strategies and implementing the key components of the reform (see Boxes A and B of Figure 1.1 in Chapter 1).

Creating Initial Commitment to FTF

As Box A of Figure 1.1 suggests, the theory of change underlying FTF holds that, for program implementation to be successful, people must have an understanding of the reforms that will be put in place; they must believe that change is necessary; and they must feel committed to that change. The planning year for the initiative consists of activities aimed at building understanding and commitment on the part of all school personnel.

While the program model evolved considerably over time, IRRE's process for introducing FTF and launching program operations in the scaling-up schools remained quite similar to the one used in the Kansas City, Kansas, schools. That process is described in detail in earlier MDRC reports. In brief, a sequence of meetings known as "Roundtables" — the first being held for district and school administrators, and the second for all school personnel — was the vehicle through which administrators and teachers came to learn about FTF. At these meetings, district and school personnel heard about FTF not only from IRRE representatives but also from teachers and students from schools that had already begun to implement the initiative. Teachers subsequently joined committees in which, working under the guidance of the SIF, they planned for various aspects of program implementation. Finally, teachers and students were assigned to SLCs partly on the basis of their stated interest in the SLC's theme.[10]

As noted above, the planning year for the 2002 cohort expansion schools differed in one major way from that of its predecessor: Teachers no longer had to contend with the potentially divisive decision about school structure. Given this difference, it is worth asking whether early responses to the intervention among teachers at the two groups of schools also varied. Figure 2.1 suggests that there were differences but that, for the most part, these were neither large nor systematic. Teachers in the later-implementing group reported being more knowledgeable about the intervention at the conclusion of the planning year than their earlier-starting peers as well as being better prepared to implement the reforms. They were not, however, more likely to believe that making the changes would be essential for improving students' performance. Teachers in the 2001 cohort schools were more likely than those in the 2002 cohort schools to report that they felt "enthusiastic" or "positive" about implementing the initiative and that their colleagues were also supportive.

General Findings on Program Implementation

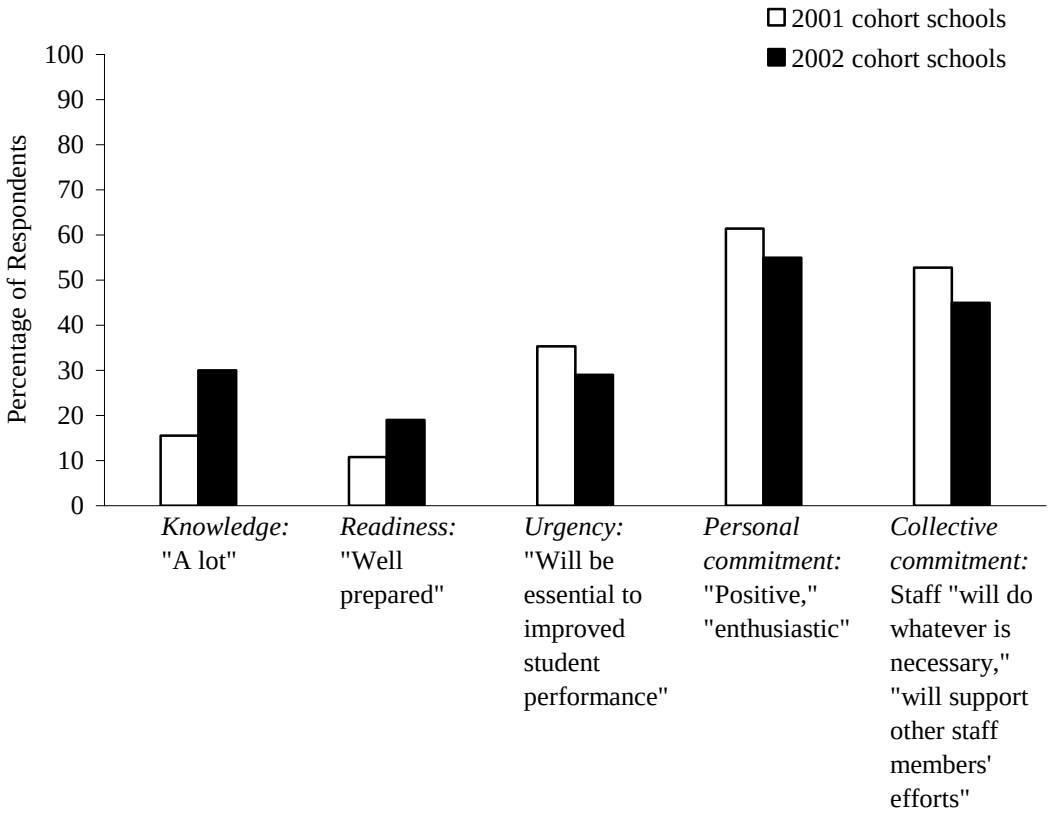
Putting the initiative's key elements of reform into place is the next stage of the FTF theory of change (shown as Box B of Figure 1.1). This section presents overall findings on the extent

[10]The process of assigning teachers to SLCs, which was handled by IRRE, also took into account teachers' certifications, areas and years of teaching experience, gender, and ethnicity — in order to ensure that SLCs would be balanced in these respects.

The First Things First Evaluation

Figure 2.1

Teachers' Responses to the Critical Features of First Things First in the Planning Year



SOURCES: 2001 and 2002 First Things First staff surveys.

NOTE: Respondents include only staff with classroom responsibilities.

of program implementation at the expansion sites, using quantitative ratings based on information contained in field research reports. To develop these ratings, MDRC staff first identified a number of dimensions associated with the successful operation of SLCs, the Family Advocate System, and instructional improvement. The dimensions, which are discussed in detail later in this chapter, are both structural and functional in nature; that is, they relate both to changes in formal configurations (for example, scheduling, student assignment patterns) and to the way that teachers and students behave within these configurations. MDRC staff then rated the implementation of each dimension at each school on a scale of 1 (the lowest rating) to 4 (the highest).[11] An individual school's overall score is the average of its ratings on each of the dimensions.[12]

IRRE staff members reviewed the rank order of schools resulting from these ratings and reported that the lineup accurately reflected their own assessments of the extent of implementation among the schools. This independent assessment reassured the evaluators about the underlying validity of the ratings. At the same time, readers are cautioned that the ratings are more useful in a relative than in an absolute sense and that small differences in ratings may have little or no meaning. It would be difficult to conclude, for example, that a school rated 2.9 implemented FTF more completely than one rated 2.8. On the other hand, a school rated 3.1 can safely be assumed to have implemented the initiative at a higher level than one rated 2.5. In general, the larger the differences and the more consistently they show up across groups of schools, the greater the confidence that should be placed in the ratings' meaningfulness.

Overall Averages

The top row of Table 2.2 shows that the average score for all schools across the individual structural and functional dimensions was 2.9, indicating both considerable progress and considerable room for improvement. The right-hand column shows that scores for individual

[11]More specifically, 1 indicated no implementation of the dimension; 2 indicated that implementation had begun but was relatively rudimentary; 3 indicated good implementation but with room for growth; and 4 indicated that implementation had reached a high level. The evaluators also used midpoint ratings of 1.5, 2.5, and 3.5.

[12]The methodology used to develop the ratings is described more fully in Appendix A.

By design, the evaluators measured dimensions related to the seven "critical features" that figure prominently in IRRE's early descriptions of the initiative and are, therefore, of substantive importance to the initiative. Potential dimensions were eliminated from consideration if it was known in advance that there was little variation on these dimensions among schools. For example, it was known that virtually all high school SLCs (except for special "catch-up" academies for ninth-graders and for communities geared toward English language learners) covered all four years of school and that all were thematic (at least in name).

Variation in implementation also occurred among the SLCs within an individual school, but it would have been prohibitively costly to collect information on all the SLCs. Moreover, since program impacts could not be evaluated at the SLC level, it would not have been possible to relate variation in SLC implementation of FTF to variation in program impacts.

The First Things First Evaluation

Table 2.2

Extent of Implementation of First Things First: Across Schools

Measure	Average Score Across All Schools	Range of Scores for Individual Schools
Average score across all dimensions	2.9	2.3 - 3.4
Average score across structural dimensions	3.4	2.7 - 4.0
Average score across functional dimensions	2.7	2.1 - 3.4
Average score across all dimensions		
All high schools	2.8	2.3 - 3.4
All middle schools	3.0	2.6 - 3.4
Average score across structural dimensions		
All high schools	3.3	2.7 - 3.8
All middle schools	3.6	2.9 - 4.0
Average score across functional dimensions		
All high schools	2.7	2.1 - 3.4
All middle schools	2.7	2.4 - 3.2
Average score across all dimensions for		
All 2001 cohort schools	2.9	2.3 - 3.4
2001 cohort high schools	2.8	2.3 - 3.4
2001 cohort middle schools	2.9	2.6 - 3.4
Average score across all dimensions for		
All 2002 cohort schools	3.0	2.7 - 3.2
2002 cohort high schools	2.9	2.7 - 3.0
2002 cohort middle schools	3.0	2.8 - 3.2
Average score across all dimensions for		
Houston schools	3.1	2.7 - 3.4
Riverview Gardens schools	2.6	2.3 - 2.8
Greenville-Weston High School	2.6	NA
Shaw High School	3.0	NA

SOURCE: MDRC analysis of field research reports.

NOTE: Ratings are on a scale of 1 to 4, where 1 indicates no implementation, 2 indicates beginning implementation, 3 indicates good implementation with room for growth, and 4 indicates a high level of implementation.

schools varied widely, ranging from 2.3 to 3.4. Reasons for such variation are discussed in a later section of this chapter.

Structural and Functional Dimensions

The second and third rows of Table 2.2 show average scores and ranges of school-specific scores, first for the five structural dimensions that were measured and then for the ten functional dimensions. Both the averages and the ranges provide quantitative evidence for the proposition that it is considerably easier to change the structure of schools than it is to change how administrators, teachers, and students behave within the new structures. This is not to say that changing structure is easy. The discussion in a later section makes it clear that schools experienced real problems in implementing some of the structural changes called for by the program (establishing "pure" classes within SLCs and reducing student-teacher ratios, for example), especially at the outset. But relative to efforts to change how people interact within those structures, structural changes can be achieved fairly readily when they are promoted and supported by principals and other key figures.

High Schools and Middle Schools

The second panel of Table 2.2 provides evidence that implementation proceeded more readily in middle schools than in high schools. As discussed below, some structural features (for example, purity) proved easier to put in place in middle schools than in high schools, and other factors (also considered later in the chapter) may further help to explain these differences.

Earlier- and Later-Starting Schools

A plausible hypothesis was that schools in the 2001 cohort would score higher on the implementation measure than their later-starting counterparts because their extra year of experience with the initiative would result in more complete adoption. On the other hand, schools in the 2002 cohort might have benefited from IRRE's greater clarity and prescriptiveness about the program model. The bottom panel of Table 2.2 shows average scores and ranges of scores for the 2001 cohort and the 2002 cohort schools, and for high schools and middle schools separately within each implementation cohort. The overall scores for the two cohorts — 2.9 for the earlier-implementing and 3.0 for the later-implementing schools — are very similar, as are scores for the earlier- and later-implementing high schools and middle schools. Indeed, in this small sample, the earlier-starting group includes schools with both the highest and the lowest implementation scores.

This suggests that time per se does not result in better implementation. What matters is the use that schools make of that time (or fail to make of it). As discussed below, in one district, a complete turnover of district personnel meant that the initiative failed to move forward for

over a year. Because schools in the two cohorts had attained very similar average levels of implementation by the end of the research period, no further effort is made to distinguish between the two implementation cohorts in the remainder of this chapter.

Variation Among Expansion-Site Districts and Schools: Factors Influencing Implementation

Implementation unfolds not in the abstract but in specific settings, influenced by a variety of factors that play out in unique ways. This section explores differences among schools and districts in their implementation of FTF, and it examines reasons for this variation.

School Ratings

As noted earlier in the chapter, overall implementation ratings for each school were derived by averaging the ratings given to various dimensions. Table 2.3 shows the rating of each school; in the interest of confidentiality, schools are identified only by letter in this table and in the rest of this report.

Two points are critical. First, while some schools had achieved more complete implementation than others, all had made sufficient progress for the researchers to conclude that the schools should be included in the impact evaluation. Second, these ratings represent implementation at a given point in time — the middle of the 2003-2004 academic year for the Houston and Riverview Gardens schools, the end of that year for the Mississippi sites. Much can change in only a few months, and, under the direction of a new area superintendent in one district, things largely fell apart at two schools between the middle of the 2003-2004 school year and the start of the next year. If the ratings had been conducted at the end of the year, the schools would have received considerably lower scores.

District Ratings

The top panel of Table 2.3 also shows average scores across all the schools within each district participating in the scaling-up demonstration. (Greenville and Shaw had only one school each.) Two districts (Houston and Shaw) had notably higher average scores than the other two (Greenville and Riverview Gardens). Reasons for this disparity are explored below.

What Makes a Difference?

The answer is twofold and relatively easy to state — although far from easy to ensure: district and school leadership and intensive technical assistance delivered by IRRE staff and consultants. Implementation benefited most when district officials, the principal, the SIF, other

The First Things First Evaluation

Table 2.3

Extent of Implementation of First Things First: Individual Schools and Districts

Measure	Average Score	Range of Scores for Individual Schools
Average score across all dimensions for		
Houston schools	3.1	2.7 - 3.4
Riverview Gardens schools	2.6	2.3 - 2.8
Greenville schools	2.6	NA
Shaw schools	3.0	NA
Average score across all dimensions for all high schools		
School E	3.4	NA
School F	3.0	NA
School G	2.7	NA
School H	2.3	NA
School I	2.6	NA
School J	3.0	NA
Average score across all dimensions for all middle schools		
School S	3.4	NA
School T	3.1	NA
School U	3.2	NA
School V	2.8	NA
School W	2.6	NA
School X	2.8	NA

SOURCE: MDRC analysis of field research reports.

NOTE: Ratings are on a scale of 1 to 4, where 1 indicates no implementation, 2 indicates beginning implementation, 3 indicates good implementation with room for growth, and 4 indicates a high level of implementation.

school administrators, and SLC coordinators were aligned in support of the initiative and exerted the pressure needed for the FTF changes to be put in place. The combination of supports and pressures from leadership that were supplied by IRRE was also critical for success. Table 2.4 summarizes the experiences of the five school districts treated in this report with regard to leadership and to IRRE involvement.

The First Things First Evaluation

Table 2.4

Leadership and Technical Assistance in the First Things First Sites

Kansas City, Kansas

- **Leadership** – Consistent support from the superintendent and other central office leaders from FTF's inception. District personnel provide both support and pressure for effective implementation at FTF schools, overcome district- and school-level resistance, launch major initiative to improve instruction.

- **Technical Assistance** – Key IRRE personnel visit the site at least bi-monthly to advise, provide sounding-board for district leadership, monitor implementation, maintain clarity of vision, and push for continuous progress.

Houston, Texas

- **Leadership** – School district as a whole divided into regions ("districts"). Some district superintendents support FTF; a leadership change in one district derails progress of FTF schools. Turnover in leadership of four of seven study schools.

- **Technical Assistance** – Regular visits by IRRE staff. District-paid FTF director serves as liaison between district and FTF schools. In last implementation year, IRRE appoints a former SIF as an IRRE employee, to coordinate activities across FTF schools.

Riverview Gardens, Missouri

- **Leadership** – New superintendent in second year of FTF implementation, complete turnover in central office staff. New administration shows little interest in or commitment to FTF, and implementation progress at schools is stalled. Administration becomes more favorable toward FTF at end of study period. Turnover in leadership of all three study schools.

- **Technical Assistance –** On-site part-time IRRE consultant in first year of implementation not seen as helpful by most school personnel. Regular visits by IRRE personnel in addition to on-site staff continue until new administration takes over. Initially IRRE not welcome in district; toward end of study period, IRRE invited back in and strongly supported.

Greenville, Mississippi

- **Leadership –** Superintendent expresses strong support for FTF. Initiative receives limited attention from central office personnel. Turnover in leadership of high school.

- **Technical Assistance –** IRRE consultant makes regular visits to site during planning year and first two implementation years, is liked and respected by local personnel. School and district leaders report insufficient assistance in last study year.

(continued)

Table 2.4 (continued)

Shaw, Mississippi

- **Leadership –** Superintendent expresses support for FTF. Central office lacks resources to provide significant assistance.
- **Technical Assistance –** IRRE consultant makes regular visits to site during planning year and first two implementation years. In last year, Delta-based consultant hired by IRRE, seen by school personnel as catalyst for positive change.

Leadership

While the importance of leadership appears obvious, the way it operates is not. For one thing, leadership is not a stable commodity. All but one of the scaling-up schools experienced turnover in the position of the principal, the SIF, or both over the course of program implementation. Turnover necessarily disrupts relationships and procedures, but, in the FTF instance, it was not necessarily negative; in some cases, new appointees filled the positions much more effectively than had their predecessors. What seems to matter is not necessarily that leadership remain unchanging but that the requisite mix of pressure and support remain in force through the transition period and into the new administration.

Even when turnover was not an issue, at some point in their development, the majority of schools with higher implementation ratings experienced a leadership vacuum — the absence of a strong critical player at one or more levels. Effective leadership at some levels helped compensate for weak leadership at other levels. Thus, for example, a SIF was able to push change forward in one school headed by a weak principal — and conversely, at another school, a strong principal was able to take over the functions of a weak SIF. At yet another school, where both the principal and the SIF were largely occupied with other matters, an IRRE consultant propelled instructional improvement. IRRE's use of the term "system leadership" to describe what is called for seems apt: Not every person in a leadership position need be a strong promoter and enforcer of change, but there must be a critical mass of personnel who together succeed in accomplishing the tasks at hand.

But if change does not require that every leader be a champion of the reform, the FTF experience seems to suggest that change will not occur when leaders are actively opposed to the reform, especially when they occupy positions of considerable authority and cannot easily be dislodged (as can, say, an SLC coordinator). The remainder of this section elaborates on the roles of various leaders in the reform process.

District Personnel

The experience of Kansas City, Kansas, where district officials embraced FTF as the district's school reform initiative and closely monitored its unfolding, illustrates how a high level of support from the district personnel can foster implementation progress at the level of the individual schools. On the other hand, FTF's difficult course in Riverview Gardens suggests that if no district leaders display interest in or acceptance of the reform, school-level leaders cannot be expected to push for its implementation.

FTF was introduced into the Riverview Gardens district by a superintendent who was strongly supportive of its goals and means. She left the district for another position midway through the study period and was replaced by a new superintendent; this was followed by virtually complete turnover within the central office, so that none of the district personnel who had been involved in FTF's adoption or initial implementation remained in their positions. The new superintendent and his key administrators showed no interest in FTF and, citing other priorities, did not find the time to meet with IRRE staff. Moreover, they refused IRRE staff access to the schools and prohibited school staff members from communicating directly with IRRE rather than going through the central office. In this context, those principals and SIFs who had supported FTF were reluctant to press for its implementation and thus risk the disapproval of the central office.[13] The absence of an IRRE presence led teachers to wonder whether FTF still existed in Riverview Gardens; at the start of the 2003-2004 school year, a number of teachers commented to the field researcher that they didn't think their school was doing FTF any more.

Under pressure from the Missouri Department of Elementary and Secondary Education to improve instruction and raise test scores, district personnel looked to IRRE for assistance. As the 2003-2004 school year drew to a close, a rapprochement was beginning to occur. But the consequence of all this was that, for the better part of two years, there was no pressure either from outside or inside the district to implement FTF, and this is a major explanation for the fact that all three Riverview Gardens schools received relatively low implementation ratings.

In other districts participating in the demonstration, district personnel were generally more favorable toward FTF.[14] Their approval sent an important signal to staff in the schools. It meant that school-level personnel could move forward in implementing the initiative with the support of district leadership — and without fear of reprimand.

---

[13]It is worth noting that all three Riverview Gardens schools acquired new principals during the demonstration period.

[14]During the 2003-2004 school year, a change in administrators in one of the Houston district offices resulted in the appointment of a new district superintendent who was unsympathetic to FTF. In the final months of the study period, much of the considerable progress in implementing FTF that had been made by the two schools under his supervision came undone.

Principals

Just as district staff send a message to the principals they supervise, so principals send a critical message of support to administrators and teachers in their schools. The critical role of the principal in the start-up phase of the project was discussed in an earlier report.[15] That role continues to be important over time, and, ideally, the principal is someone who advocates for the initiative and propels it forward; minimally, that person understands it and makes decisions that do not subvert FTF tenets (as did the decision by two principals to reconfigure family advocacy groups midstream, in an effort to raise student test scores, discussed below). A school also benefits when its principal and SIF have a trusting, cooperative relationship — but principals must also be willing to reassign SIFs whose job performance is inadequate.

School Improvement Facilitators (SIFs)

Over the course of implementation, both the responsibilities of the SIF and the skills required to hold that position changed significantly. At the outset, interpersonal skills were paramount, since the position largely entailed guiding the structural changes and promoting positive interactions among SLC members. Once SLCs were in place and functioning reasonably effectively, however, attention shifted to instructional improvement, and the SIF's abilities as an *instructional* leader were front and center.

The FTF experience suggests that there are several reasons to consider hiring SIFs from outside the schools. For one thing, a single person with the requisite mix of interpersonal and instructional skills cannot readily be found at all schools. For another, because SIFs must be willing to exercise pressure, not just to support, they cannot be overly concerned with being liked by colleagues who are also friends. Finally, coming from outside the school may help protect SIFs from principals who want to turn them into administrative assistants, testing coordinators, assistant principals, and the like.

Other School Administrators

The demonstration experience suggests that FTF implementation can be strengthened by providing an important role to other school administrators, counselors, and the like. In one school, for example, each assistant principal was assigned to an SLC and attended its meetings regularly. (Interestingly, there is little evidence that the presence of school administrators deterred teachers from expressing their thoughts and opinions.) These individuals were also trained and enlisted to help monitor classroom instruction and guide the SLC's instructional improvement efforts.

[15]Quint, 2002.

SLC Coordinators

SLC coordinators played a crucial role in ensuring that SLC meetings were useful and productive. The skills required to be a coordinator were not always readily found, and as the demonstration progressed, IRRE recognized the need to provide coordinators with additional training. Some SLCs have also found it useful to have co-coordinators.

The Institute for Research and Reform in Education (IRRE)

IRRE's critical importance as an agent of change cannot be overstated. IRRE has consistently provided sites with a plan for change, a vocabulary for talking about it, a set of processes for implementing the key components, and a variety of tools for monitoring progress. Equally important, IRRE has given administrators positive recognition when called for, advice and a sounding board when requested, and a push when that seemed required.

It is not coincidental that the higher-rated schools were in locations marked by the presence of an on-site IRRE technical assistance provider for at least part of the last year of the study. These individuals were able to provide consistent, ongoing guidance and support, to follow up immediately on problems, and in general to keep the schools moving forward in their implementation efforts.

In contrast, at the lower-rated schools, IRRE's visits were much more occasional — in part because of the initial coolness of the new Riverview Gardens administration toward FTF, in part because IRRE is a small organization and has had to make hard choices about resources. During the 2002-2003 academic year, it opted to focus more attention on a new large urban district with large numbers of students who could benefit from the reform, and, the following year, it added schools in yet another large urban district to the roster of partner schools. An IRRE consultant in Greenville first moved to New Orleans and then left the position altogether. Interviewed toward the end of the 2003-2004 school year, Greenville administrators expressed gratitude for IRRE's assistance in earlier years, but they also said that they felt "cut loose" as the study period drew to a close. Attending an IRRE-sponsored conference of all the FTF sites toward the end of the year left them feeling reinvigorated and eager to move forward with the hard business of school change.

Implementing the Key Components

This section discusses the implementation of structural and functional dimensions associated with the three key components of FTF: small learning communities, the Family Advocate System, and efforts to improve instruction. Table 2.5 shows the average score across all schools, along with the range of scores, for each dimension.

The First Things First Evaluation

Table 2.5

Extent of Implementation of Key Dimensions of First Things First

Implementation Dimension	Average Score Across All Schools	Range of Scores for Individual Schools
Small learning communities		
Structural		
Purity of classes	3.5	2.5 - 4.0
Adequacy of common planning time	3.4	2.0 - 4.0
Functional		
Personalized relationships	3.0	3.0 - 3.0
Staff decision-making	2.5	1.5 - 4.0
Staff accountability	2.5	1.5 - 3.5
All small learning community dimensions	2.8	2.3 - 3.4
Family Advocate System		
Structural		
Presence of Family Advocate	4.0	4.0 - 4.0
Functional		
Relationships with students and families	2.7	1.5 - 3.5
Meaningful activities during Family Advocate Period	3.1	1.0 - 4.0
All Family Advocate System dimensions	3.2	2.3 - 3.8
Instructional improvement		
Structural		
Extended time in English Language Arts, Math	3.2	1.0 - 4.0
Reduced ratios	2.3	1.0 - 3.5
Functional		
Alignment of curriculum and standards	3.2	2.0 - 4.0
Active learning	2.8	2.0 - 4.0
Clear, high academic tandards	2.8	2.0 - 3.5
Theme-related instruction	2.3	1.0 - 3.0
Development of teacher professional learning communities	2.6	2.0 - 4.0
All instructional improvement dimensions	2.8	2.0 - 3.4

SOURCE: MDRC analysis of field research reports.

NOTE: Ratings are on a scale of 1 to 4, where 1 indicates no implementation, 2 indicates beginning implementation, 3 indicates good implementation with room for growth, and 4 indicates a high level of implementation.

Examples abound of the key role played by leadership at all levels — from school district superintendents to SLC coordinators — as well as by IRRE, in shaping how program elements were put in place. In addition, the experiences of the scaling-up schools may suggest other lessons for parties interested in mounting initiatives in which these components figure prominently.

Small Learning Communities

Within the first year of their implementation, SLCs had become the major organizing principle of the FTF schools, the context within which important interactions between teachers and students and among members of each group took place and within which key decisions were often made. Before FTF was implemented, the three earlier-implementing middle schools had all had positive experiences with teacher-student clusters, and teachers' familiarity with the SLC concept may help to explain the general acceptance of this program element. It seems likely that SLCs took hold quickly not only because teachers were predisposed to view them favorably but also because SLCs provided teachers and students with a new sense of belonging, and because members of both groups quickly came to see the benefits of mutual familiarity.

Implementing the Structural Dimensions of SLCs

Purity of classes and *adequacy of common planning time for SLC staff meetings* are two key structural dimensions associated with SLC functioning.

Purity of Classes

Achieving "purity" in setting up the class schedule — that is, creating classes in which core-subject classes are limited to teachers and students in the same SLC — is an important element of SLC functioning. The experience of the 2001 cohort schools provided an important lesson: Scheduling classes within SLCs is hard, especially the first time. Part of the difficulty is inherent in the program design, which limits scheduling options: A ninth-grader, for instance, cannot be assigned to any ninth-grade English class but only to those classes taught by an English teacher in the student's SLC. But in preparing for the first implementation year, school personnel added to the problem: Despite IRRE's technical assistance, entreaties, and warnings, some counselors and others responsible for scheduling failed to understand the full complexity of what needed to be done, and many thought that, as in the past, they could leave the job of scheduling until the two weeks before school started. Only then did they learn that the computerized scheduling programs with which they were familiar could not easily be adjusted to meet the initiative's special requirements. Some schools had to resort to hand-scheduling students at the eleventh hour, placing students into SLC classes when they could do so easily but into any available class when difficulties arose.

The message hit home: It is critical to take care of scheduling early, well before school starts in the fall. In preparing for the second year, the 2001 cohort schools followed this precept, and by all accounts that year went far more smoothly than had the first. The 2002 cohort schools, for their part, also handled scheduling early, so that their first implementation year was much easier than that of their predecessors

At another school, the principal was initially an obstacle to SLC purity. She was especially proud of her school's wide range of electives and reluctant to limit access to these electives to the students in the SLC where the courses were housed. During the first implementation year, ensuring SLC purity took a distinct second place to preserving open enrollment in the electives, but, by the second year, the principal was persuaded that maintaining purity was a more critical objective, and students' schedules were arranged with this goal in mind.

The data in Table 2.5 indicate that, by the 2003-2004 school year, schools had figured out how to arrange schedules so as to achieve a high level of purity: The average rating for all sites was 3.5, with scores for individual schools ranging from 2.5 to 4.0.[16] Data not shown in the table indicate that it was easier to achieve SLC purity in middle schools than in high schools: The average rating across the six middle schools on this dimension was 3.9, while that for the six high schools was 3.1.[17] Some students in the upper grades of high school needed certain courses in order to graduate, so it was a priority to ensure that they took these courses, whether within their SLC or not. Moreover, upper-level courses in high school — especially in science and social studies — are frequently electives. To make offering such courses, along with Advanced Placement courses, more feasible, it was necessary to draw on students from across SLCs. Middle school students, in contrast, take essentially the same classes, making scheduling within SLCs considerably less problematic.

Adequate Common Planning Time for SLC Members

IRRE prescribed that SLC members have at least 180 minutes a week of common planning time for SLC meetings. The average rating of 3.4 on this dimension (Table 2.5) indicates that, by the 2003-2004 school year, schools had made good progress in meeting this goal.[18] Three schools achieved ratings of only 2.0. At these schools, common planning time was drastically reduced when the schools were found to be out of compliance with state standards requir-

[16]Ratings were based on an examination of the class rosters of 20 English and math classes in each school, to determine the number of number of students in each class who did not belong to the teacher's SLC.

[17]Teacher survey responses support this finding. Across the high schools, only 45 percent of teachers of core-subject classes said that "all" or "most but not all" of their classes contained only students in their SLCs. For middle school teachers, the corresponding percentage was 80 percent.

[18]The field research ratings reflect whether a full 180 minutes of common planning time was scheduled but also whether — according to the researchers — teachers and others reported that meetings were frequently canceled.

ing that teachers have 250 minutes of individual planning time per week. To meet this standard, district and school leaders sacrificed the time available for group meetings.

Implementing the Functional Dimensions of SLCs

While having adequate time for meetings is important, the field research indicates that there was considerable variation — often among SLCs within the same school — in how that time was used. Although the content of SLC discussions was not rated as part of the analysis, the field researchers sat in on many of these meetings and were able to trace their evolution over time. At the outset of the demonstration, a good deal of meeting time was spent talking about individual students' conduct and performance, calling and holding meetings with parents, reviewing information and directives handed down by administrators, arranging field trips, and preparing for SLC award ceremonies. Teachers spent relatively little time engaged in a critical examination of their own instructional practices and those of their colleagues. This is not surprising, both because the teachers had little prior experience working together and had first to establish an atmosphere of trust and cooperation and because IRRE had not yet provided SLCs with guidelines on how to use common planning time for instructional ends.

As implementation moved forward — partly because many SLCs felt that they had licked major discipline problems, partly because teachers felt more comfortable with each other, and partly in response to IRRE's guidance — teachers in most SLCs spent more time discussing instruction and student achievement. Thus, some three-quarters of teachers who were surveyed in the spring of 2004 said that they discussed ways to make instruction more engaging for students and reviewed SLC students' progress against performance targets in "a lot" or "some" of their SLC meetings. The role of SLCs in promoting instructional improvement is discussed in a later section of this chapter.

Three main factors appear to have influenced the content and style of SLC functioning: the leadership skills of the SLC coordinator, the extent to which the SIF and others monitored staff meetings, and the extent of decision-making authority that the SLC possessed. Effective SLC leaders listened to their colleagues and did not try to impose their own views on others, but they also kept members focused on big issues rather than trivia, helped guide their peers in conducting new tasks (for example, observing other teachers' classes), and generally helped provide a sense of forward movement. In contrast, in SLCs whose coordinators lacked the requisite skills, discussions often seemed aimless, and tensions sometimes festered. SLC meetings also benefited from consistent high-level attention from SIFs and school administrators, such as assistant principals, who helped keep discussions on track.

SLC discussions were also more focused and productive when members had important matters to discuss. The extent of *decision-making* conducted by SLCs, the *development of close*

relationships between students and teachers and among members of each group, and the degree of *accountability for student outcomes* that teachers assumed are among the functional dimensions of SLCs considered below.

Personalized Relationships

The fostering of strong relationships between teachers and students and among members of each group is the raison d'etre of SLCs. The field research indicates that SLCs appear to have been effective in enabling teachers and students to know each other better and to forge closer relationships. Schools made good progress toward this goal, as the average rating of 3.0 indicates (Table 2.5). While interviews with teachers and others occasionally revealed concern about a potential loss of schoolwide identity, for many students and teachers, SLCs brought a new sense of belonging.

Student and teacher surveys supply additional evidence about the role of the SLCs in nurturing the development of close personal relationships. Asked to rate relationships between teachers and students in their schools, 30 percent of the teachers described these as "excellent" or "very good"; 45 percent, in contrast, gave similarly positive ratings to relationships between teachers and students in their own SLCs.[19] For their part, across the schools, 65 percent of the students expressed strong or moderate agreement with the statement "Being in an SLC lets me get to know my teachers." The proportion of students similarly agreeing with "Being in an SLC means being in classes with students I know" was somewhat smaller (52 percent), perhaps because SLCs — while much smaller than the schools in which they were lodged — could still contain as many as 350 students and fall within IRRE's guidelines.[20]

Staff Decision-Making

According to the FTF model, SLCs, acting within broad guidelines, have autonomy to make or participate in decisions related to academics, student discipline, the scheduling of classes, the hiring of new staff, the allocation of funds and space, and a variety of other topics. The summary score for this dimension — 2.5 — suggests that while schools in general had made some progress, teachers could have held much more decision-making power than they actually possessed. The scores range from 1.5 to 4.0, indicating that there was considerable

[19]Similarly, asked to rate relationships among teachers generally in their schools, one-third of the teachers described these relationships as "excellent" or "very good"; when asked to rate relationships among teachers in their SLC, this proportion rose to 61 percent.

[20]During the course of the demonstration, IRRE raised the upper limit from 250 to 350 students. This change was made because, over time, it became clear that the smaller the number of students in an SLC, the more likely it was that teachers would have to teach students from more than one SLC. Increasing the number of students in each SLC helped to ensure greater purity.

variation among schools in this regard, depending in large part on the principal's willingness to share power. A few principals gave teachers the authority to make a wide range of decisions; others wanted to retain control over particular areas.[21] Another consideration was that some teachers seemed uninterested in a greater role in decision-making, perhaps because they felt that they had enough to do in coping with the new structural and instructional changes.

Staff Accountability

Accountability was judged to be higher in schools where SLCs established numerical goals for student improvement and reviewed progress against these goals. The school ratings also took into account whether teachers' conversations about students in SCL meetings and the attitudes they expressed in interviews emphasized ways that they could help students learn or, in contrast, stressed students' poor preparation, disadvantaged backgrounds, uncaring parents, or other factors that militated against academic achievement (and reduced their own responsibility for poor outcomes). The average rating of 2.5 on this dimension (Table 2.5) indicates that, as with staff decision-making, schools had made progress on accountability but that there was also much room for improvement.

Program developers hoped that, over time, teachers would hold not only themselves but other members of their SLCs accountable for student achievement and would confront their colleagues whose students repeatedly experienced poor outcomes. Such confrontations did occur, although rarely: In one instance, for example, an SLC decided to take away one teacher's family advocacy students because the other teachers judged that the individual was not performing this role effectively. For the most part, however, hopes that teachers would take action to confront incompetent teachers proved unfounded — and perhaps the idea was unrealistic from the start. Field research interviews suggest that teachers in the SLCs were well aware that some of their colleagues were not very good teachers and indicate that occasionally other SLC members offered help to a colleague who appeared to be struggling in the classroom. But teachers felt strongly that it was the responsibility of school administrators — the principal and assistant principals — to assist and, if necessary, to dismiss teachers who were doing a poor job. Thus, while teachers sometimes resented the presence of weak colleagues, they resented even more

[21]Thus, for example, at many schools, SLCs were allocated small discretionary budgets of $3,000 or so a year, and, in response to the teacher survey, 57 percent of the teachers said that they participated "greatly" or "somewhat" in deciding how these funds were used. At a number of schools, however, principals told interviewers that, given straitened financial conditions, they were reluctant to give SLCs authority over funds that the teachers did not need or might misuse. These principals said that they were responsive to teachers' individual requests for funding — and, in fact, few teachers complained that their requests went unheeded. Principals often had virtually exclusive say over the hiring of new staff members — sometimes because hiring decisions were typically made during the summer, when teachers were on vacation, but, in at least one case, because the principal was proud of her prowess in selecting good teachers.

that administrators seemed willing to tolerate bad teachers rather than do the hard work of helping them improve or getting rid of them.

The Family Advocate System

The analysis measured implementation of the Family Advocate System along one structural dimension — *ensuring that all students had advocates* — and two functional dimensions: *establishing close teacher-student relationships* and *making meaningful use of the Family Advocate Period.*

Implementing the Structural Dimension of Family Advocacy

The average rating of 4.0 shown in Table 2.5 suggests that all schools accomplished the most basic task associated with the Family Advocate System: *ensuring that all students had advocates.*[22] That said, both the field data and the student survey indicate that there was less continuity in the identity of the family advocate than had been anticipated. Fewer than half — 42 percent — of all students who had had a family advocate for two years said that the same person had served in that role both years. This relatively low percentage reflects staff turnover, occasional transfers of staff and students among SLCs, and SLC reorganization; in two schools, too, principals decided to reassign family advocates in the hope that doing so would raise scores on state high-stakes tests.[23]

Implementing the Functional Dimensions of Family Advocacy

Family Advocate Relationships

Across all schools, the extent to which advocates had formed close relationships with their students was rated at 2.7 (Table 2.5). This average conceals a good deal of variation, however; the scores ranged from 1.5 (at a school where the principal, by his own admission, paid

[22]It is somewhat troubling that only 74 percent of the students responded on the student survey that they had an advocate. (The remainder were evenly split between those who said that they did not have one and those who weren't sure.) In some schools, advocates were known as "home room teachers" or by other designations, so it seems plausible that many students who responded that they did not have a family advocate were simply unfamiliar with the term.

[23]The two principals concluded that students would perform better if they took these tests in a more psychologically comfortable environment, along with other students whom they knew well. Accordingly, they dissolved the existing family advocacy groups, which included students in all grades, and established new groups of students who were all in the same grade and would therefore be tested together. It is not surprising that students' ratings of their relationships with their family advocates were lower at these two schools than at other schools participating in the demonstration. To anticipate a finding that is reported in Chapter 4, the reassignments did not appear to result in a more positive change in test scores than was achieved in comparable schools.

little attention to family advocacy and the component largely existed in name only) to 3.5 (at three schools where this component was especially well developed). Data not shown in the table indicate that relationships between students and advocates were more highly developed at the middle schools than at the high schools (average ratings were 3.0 and 2.25, respectively). According to the field researchers, middle school teachers saw their younger students as still malleable and thus more likely to benefit from positive interactions with caring adults.[24]

Students generally valued their interactions with their advocates. Thus, between 71 percent and 75 percent of all students responded on the survey that their advocate was either "very important" or "sort of important" in giving them someone to talk to when needed, helping them do better on schoolwork, and recognizing their accomplishments, and the majority saw their advocates as helpful in a variety of other ways. About two-thirds (64 percent) said that it was "very true" or "sort of true" that they felt comfortable approaching their family advocate even if they didn't have a specific problem to discuss, and an equal percentage agreed that their family advocate was someone whom they could work with to help them achieve in school. The role of the family advocate may well have been especially important for the large minority (42 percent) of students who said that there were no other adults in the school who had been helpful to them in the same way as their family advocates.

Field research interviews indicate that teachers' attitudes toward this component varied. Some appreciated the component's usefulness; others resented the extra responsibilities that advocacy entailed, or they felt ill-equipped to take on a role that they felt properly belonged to guidance counselors. Observations also suggest that the attitude of the SLC coordinator could sway the opinions of teachers in the SLC and could result in greater or less attention being given to this program element.

For many teachers, their role as family advocates provided a sense of accomplishment. Some 80 percent of teachers who served as advocates felt that they had made "a lot" or "some" progress in giving students a sounding board when they needed one, in helping them succeed academically, and in many other ways. And about two-thirds of the teachers said that it had been "very easy" or "fairly easy" to develop close relationships with the students for whom they served as advocates.[25]

---

[24]At one high school in Houston, family advocates took on responsibility for helping students select their courses. The process served to educate the teachers about graduation requirements — a topic about which they had largely been unaware beforehand — and led to more teacher-student interaction about academic matters.

[25]Interestingly, teachers and students differed in their estimates of the frequency with which conversations between students and advocates took place outside of class. Thus, 77 percent of the teachers, compared with 58 percent of the students, reported that such conversations took place weekly or more often. The reasons for this disparity are unclear.

Establishing relationships with students' families was harder. Almost two-thirds of the teachers reported that it was "very difficult" or "fairly difficult" to reach students' parents by telephone (sometimes because a family's phones line had been disconnected) and that it was even harder to conduct the two meetings a year with students and their parents, as called for by FTF guidelines. Well into the second semester of the school year, a substantial proportion of advocates at all schools said that there were families with whom they had not yet established contact. Teachers cited difficulties communicating with parents whose jobs made them unavailable during school hours, whose negative prior experiences with the school system made them reluctant to follow up with staff, or who did not speak English. Teachers also felt frustrated and disheartened by what they perceived as lack of parental involvement

The Family Advocate Period

The Family Advocate Period is a time in the weekly schedule devoted to students' meeting with their family advocate. Schools had autonomy to determine the length of the period and the time of day that it would be scheduled, resulting in considerable variation across sites. For example, at one school, the Family Advocate Period occupied one 35-minute period a week, while at other schools, the period occurred daily.

Schools also had the latitude to use the Family Advocate Period as administrators saw fit, as long as the resulting activities were consistent with the goals of the component. Teachers' initial uncertainty about how to use the time often led them to use it as a homeroom period or as a time to complete their own paperwork. In response, IRRE developed and distributed to all schools a guide that suggested a number of techniques for using the period effectively.[26] The average rating of 3.1 on this dimension (Table 2.5) suggests that, at most schools — aided by experience and the IRRE guide — teachers had learned to make good use of the period (although the school where family advocacy barely existed received a rating of 1.0 because it never implemented the period in the first place). About half the teachers reported on the survey that they frequently used the period for individual student conferences; the period was also used for team-building and goal-setting activities, and for homework help, along with administrative tasks.

Instructional Improvement

As noted above, two structural changes pertaining to language arts and mathematics classes — *increased instructional time* and *reduced student-teacher ratios* — have been key

[26]Institute for Research and Reform in Education, 2002. In addition to containing suggestions about the Family Advocate Period, IRRE's *Guide for Family Advocates* included forms, questions, and an "action plan" that could be used to generate discussion during family conferences and a sample resource directory that advocates could consult when referring students and families to appropriate local service agencies.

aspects of FTF since the program's inception. Over time, the functional attributes of instructional improvement have been spelled out with greater specificity and clarity.

Implementing the Structural Dimensions of Instructional Improvement

Increased Instructional Time in Language Arts and Math Classes

Because so many students in low-performing schools have weak basic skills in English and math, the FTF model calls for students to receive more instructional time in these two critical areas. As the rating of 3.2 indicates (Table 2.5), schools generally succeeded in arranging schedules so that students received more instruction in these two subjects than in others, and sometimes more instruction than in the past. (Almost all the schools had adopted block scheduling before FTF was introduced.) For example, one school assigned students to an extra half-block of language arts or math instead of to study hall.

On the other hand, the experience of a high school whose new principal was initially unsympathetic to FTF shows how an administrators' lack of commitment to the program model could undo earlier arrangements. During the first year of FTF's implementation, instruction in Algebra and English 1 was double-blocked, so that students took these classes every day. When the new principal took over during the second implementation year, he noted that scores on the state's high-stakes test had not improved during the previous year (although the students who had been double-blocked were not the ones tested), and he decided that math and language arts classes should revert to a single-blocked schedule, with students taking the classes only every other day.

Reduced Ratios

To ensure that students receive individual attention in language arts and math classes, the FTF model established a 15:1 student-teacher ratio in these subjects. The average rating of 2.3 (Table 2.5) indicates that schools had considerable difficulty meeting this rigorous standard. They were generally able to reduce class sizes below pre-FTF levels, however. Thus, the classroom observation study found that the average number of students enrolled in the language arts and math classes that were observed fell from 26 to 20 between the planning year and the second implementation year for the earlier-starting Houston schools; in Riverview Gardens, average class size dropped from 21 to 16 over the same period.[27]

[27]As IRRE conceived of it, reduced student-adult ratios were to be achieved not only by reducing class size but also by having other personnel assist the primary teacher in the classroom. These auxiliary personnel were to be "qualified" but not necessarily certified in the subjects they were teaching; indeed, to fulfill this role, schools could turn to people in the community rather than to other teachers.

(continued)

While schools made progress in increasing instructional time and in reducing student-teacher ratios, it proved well nigh impossible to achieve both goals at once, for students in all grades and in both subject areas. (Only one school managed to accomplish this feat.) Most schools lacked the personnel to cover the additional number of classes that across-the-board reductions in class size would have entailed, especially when these classes were longer than in the past. Instead, administrators had to make hard choices — to reduce class size only for ninth- and tenth-graders, for instance, on the grounds that more individualized attention early on would increase their probability of later educational success, or to focus on language arts but not on math.[28] One high school gave SLCs the choice of reduced ratios or more instructional time, with the result that some SLCs chose one alternative, and some the other.

Implementing the Functional Dimensions of Instructional Improvement

Alignment of Curriculum with State and Local Standards

The area of instructional improvement in which schools achieved the highest rating concerned the alignment of curriculum with state and local knowledge standards and the use of assessment methods mirroring those used on state high-stakes tests. The average rating on this dimension was 3.2, and the large majority of schools had ratings of 3.0 or higher (Table 2.5). While some schools had begun to align curriculum with standards before the inception of FTF, IRRE served as the catalyst for further work in this area and suggested a process for undertaking the self-scrutiny involved in better connecting curricula with prescribed content knowledge. Progress also stemmed from the fact that teachers and administrators strongly supported activities to ensure alignment, seeing such alignment as critical to students' achieving higher scores on the state tests.

Interestingly, because curricula are discipline-specific, discussions about alignment have occurred principally in the academic departments rather than in the SLCs. IRRE had originally seen the SLC as the principal venue for work to improve instruction across disciplines and among SLC teachers of the same subjects.[29] As the demonstration has evolved, however, IRRE

The expectation that other people would help out in the classroom did not materialize. Some schools tried to provide teachers with assistance by pairing them with administrators or counselors but found that these arrangements were unreliable because the staff members were often called away to handle pressing situations. The idea of bringing in people from the community was simply a nonstarter; principals and others did not have the time to recruit and train community members to assume classroom responsibilities.

[28]It also bears noting that reducing student-teacher ratios was not always within a school's control. One school experienced an unanticipated influx of new students at the beginning of the school year but could not at that point hire new teachers. Another school lost teachers when enrollment turned out to be smaller than expected.

[29]FTF's structural changes also encourage teachers to identify themselves with the SLC, a student-focused entity, rather than with the subject-centered academic department.

has come to recognize that departments play a critical role in instructional improvement efforts and has worked to define that role more coherently. Observations suggest that departments have become the locus for discussions of instructional *content*, while SLCs are the venue in which teachers talk about instructional *methods* that are applicable across subject areas.

Active Learning

The extent of active learning was judged by the degree to which teachers planned for and used the Kagan cooperative learning structures in their lessons and discussed their use in SLC meetings. The 2.8 rating across schools (Table 2.5) indicates that teachers were making use of these strategies, but not consistently — a conclusion that MDRC also reached through a classroom observation study that unfolded between 2000-2001 and 2003.[30] The average also masks a good deal of variation among the individual schools and among different teachers within the same school. Some teachers noted that the techniques captured students' interest; others felt that they were gimmicky or could not be used effectively for their particular subject or were too hard to plan for.

Administrators generally encouraged use of the active learning strategies. In Riverview Gardens, a high-level central office official announced that all teachers were to use these structures in every class every day — an edict that predictably aroused teachers' resentment but may also have led them to use the methods more than they would have otherwise. Principals, SIFs, and other administrators sometimes dropped in on classes to monitor use of the methods or asked to see lesson plans showing their use.

High, Clear Academic Standards

MDRC's classroom observation study found that even when teachers practiced the Kagan strategies, the level of instruction was often low. Using a modified version of Bloom's well-known taxonomy to measure the cognitive processes and types of knowledge transmitted in the classes that were observed, the researchers found that, in the second implementation year, stu-

[30]See Estacion, McMahon, and Quint, 2004. The study concluded that there was a marked increase over time in the proportion of classes in which students worked in pairs and small groups. Nonetheless, in half the classes observed during the earlier-starting schools' second implementation year, however, no paired or small-group learning took place.

In response to survey questions, teachers and students gave radically differing accounts of the extent of small-group instruction that occurred. Thus, 65 percent of the teachers of language arts and math classes, but only 27 percent of students in these classes, reported that students worked in small groups or pairs "in almost every class" or in "a lot of classes." Conversely, only 12 percent of the teachers but 34 percent of the students said that the teacher lectured for more than half the class period "in almost every class" or "in a lot of classes." Perhaps most tellingly, 53 percent of the students but only 36 percent of the teachers said that it was "very true" or "sort of true" that it was easy to "tune out" and not take part in the class.

dents were called on to memorize facts and apply procedures far more often than they were asked to analyze and evaluate information.[31] The peer observations and discussions of student work that were introduced by IRRE's Director of Instructional Supports were intended to make instruction more challenging. At the same time, discussions of grading standards were meant to ensure that teachers shared a common view of what constitutes high-quality work and that these standards would be clear to students.

Across the schools, the average rating on this dimension was 2.8 (Table 2.5) — again, with considerable variation among schools — indicating that while schools had made a solid start, there was much room for improvement. The field research suggests one reason for limited progress in this regard: Some teachers were skeptical about their students' capacity to take on challenging work. In SLC meetings and in individual interviews, teachers asserted that students' disadvantaged backgrounds and poor prior preparation made it difficult to offer instruction at a high level.

At the same time, there is suggestive evidence that teachers were thinking more about what constitutes high-quality instruction. On both the planning-year and the 2004 survey, teachers were asked the extent to which high, clear, and fair academic standards existed throughout their schools. At half the schools, the proportion of teachers reporting that these standards were in place fell by several percentage points. It seems highly unlikely that standards actually fell or became less clear or more unfair. Instead, it appears that teachers were grappling with the issues and looking more critically at their own classroom practice and that of their colleagues.

Theme-Related Instruction

In the minds of IRRE planners, one benefit of thematic SLCs was that teachers could relate the subjects they were teaching to areas in which students had expressed interest. The average score of 2.3 (Table 2.5) — the lowest score accorded to any of the dimensions associated with instructional improvement — indicates that, despite the broad nature of the SLC themes (such as "Performing Arts" or "Science and Technology"), teachers experienced considerable difficulty in incorporating the theme of their SLC into their classes. Field research interviews and observations suggest that while pairs of teachers occasionally collaborated to develop a theme-related unit and from time to time core-subject teachers made reference to their SLC's theme in their lessons, these practices were uncommon. Instead, the task of communicating the theme of the SLC fell to the elective teachers. Student surveys suggest that student engagement may have suffered as a consequence: While 78 percent of the students said that learning about a specific theme was "very important" or "sort of important" to them, only 59 percent agreed that being in an SLC gave them the chance to learn about this theme.

[31] See Bloom et al., 1956; Anderson and Krathwohl, 2001.

The Development of Teacher Professional Learning Community

Some of the indicators of the existence of a professional learning community among teachers — such as regular peer observations and discussions of student work and instructional methods — are also used to gauge the dimensions of curricular alignment, active learning, and high and clear academic standards. But the concept of a teacher professional learning community goes beyond specific practices to denote an ethos of ongoing collegial consultation and what might be called a "culture of continuous improvement." The mean score on this dimension was 2.6 (Table 2.5), but individual school scores ranged from 2.0, indicating that such a culture was in its infancy, to 4.0, indicating that it was well developed. According to field research reports, at School E — the school with the 4.0 rating — administrators and teachers alike were working together to build a common vision of good teaching.

Chapter 3

Changes in Support and Engagement Among Teachers and Students

This chapter turns to psychological outcomes associated with the post-implementation stage of the First Things First (FTF) theory of change: feelings of support and engagement among both staff and students. These outcomes are shown as Boxes C1 and C2 in the theory of change diagram shown in Chapter 1 (Figure 1.1). In this theory, changes in support and engagement are "intermediate outcomes." They are at once the result of program implementation and the psychological wellspring of the desired long-term outcomes of improved student attendance, persistence, and achievement.

FTF seeks to create an environment in which teachers experience interpersonal and instructional support from their colleagues, administrators, and others. According to the theory, these feelings of support lead teachers to develop increased feelings of competence, autonomy, relatedness to important others, and engagement — that is, the willingness to do the utmost to meet their students' needs. An analogous process exists for students. Increasing the support that students receive from adults and peers also induces students to develop positive beliefs about themselves and school and to display greater engagement with academics. For students, engagement entails both a belief that doing well is personally important and a set of behaviors and feelings that back up that belief and put it into practice (for example, trying hard, preparing for class, paying attention, taking responsibility, and avoiding anger and blame when setbacks occur).

This chapter focuses on outcomes at the FTF expansion sites outside Kansas City, Kansas; for easy comparison, the Kansas City findings are reprised briefly in a box on the next page. Items measuring support and engagement for teachers and students that were developed by the Institute for Research and Reform in Education (IRRE) were incorporated into the staff and student surveys administered during the planning year and subsequent implementation years at the scaling-up schools; the items for students are shown in Table 3.1, and those for staff appear in Table 3.2. The items form scales with scores that can hypothetically range from a low of 1 to a high of 4. This chapter examines changes on these scales in two ways. First, it shows changes in *average* scores over time; this approach has the advantage of making use of all the available data. Second, it shows changes in the *proportions* of teachers and students whose scores are considered especially high or low on each measure; these changes in proportions may be of

Support and Engagement in Kansas City, Kansas

The Youth Development Strategies, Inc., (YDSI) evaluation of First Things First (FTF)* examined changes over time in teacher and student support and engagement in the Kansas City, Kansas, schools, using survey items very similar to the ones used in the MDRC study to measure these constructs. The analytic approach differed, however, with the YDSI study focusing on changes in the likelihood that teachers and students would experience high or low levels of support or engagement.

Student surveys were administered each year from 1998 through 2003, and staff surveys were administered each year from 1998 through 2002. (YDSI opted not to present teacher support and engagement findings for schools that had been implementing FTF only two years, instead limiting its analysis to teachers in the first two clusters of schools mounting the initiative.)

Student reports of support from teachers

The YDSI study reported consistently positive findings with respect to students' feelings of being liked and cared about by their teachers. There were significant increases in the likelihood that Kansas City, Kansas, students attending middle school and high school would perceive their teachers as highly supportive and decreases in the likelihood that they would report the lowest levels of teacher support. These improvements in teacher-student relationships were apparent from the first year of implementation on.

Student engagement

There was also an increase in the likelihood that secondary school students would see themselves as highly engaged in school, but this did not become evident until the third year of implementation. In contrast, there was a steady decrease in the likelihood that students would report a low level of engagement from the first implementation year forward.

Teacher reports of support

In the Kansas City evaluation, teacher support is conceptualized as involving two subconstructs: support from colleagues and support from building and district leadership. The only significant improvement in support from colleagues, evident in the second implementation year, was in the proportion of middle and high school teachers who reported low levels of such support. However, after the first year implementation, there was a marked increase in the likelihood of secondary teachers reporting high levels of support from building and district leadership, and a decrease in the likelihood of reporting low system support.

Teacher engagement

After one year of implementation, there was an increase in the likelihood of teachers reporting a high level of engagement and a decrease in the likelihood of reporting a low level of engagement.

In summary, there were positive effects on the support and engagement measures among both students and teachers. These findings stand in marked contrast to those for the expansion sites, where students generally reported higher levels of support from their teachers but findings were otherwise mixed.

---

*Gambone et al., 2004.

particular interest to educators and others.[1] Two years of follow-up data are available for all schools; three years of follow-up data are available for schools that began implementing FTF in 2001. Tables in the chapter present results for groups of schools; changes in average scores for individual schools are shown in Appendix B.

The First Things First Evaluation

Table 3.1

Survey Items Measuring Support and Engagement Among Students

Teacher support
My teachers like to be with me.
My teacher likes the other kids in my class better than me.
My teacher interrupts me when I have something to say.
My teachers are fair with me.
My teachers' expectations for me are way off base.
My teachers aren't fair with me.
My teachers don't make clear what they expect of me in school.
My teachers care about how I do in school.
Overall engagement
I work very hard on my schoolwork.
I don't try very hard in school.
I pay attention in class.
I often come to class unprepared.
When something bad happens to me in school…
I get angry at the teacher.
I try to see what I did wrong.
I say it was the teacher's fault.

SOURCE: The 1999 measurement report for First Things First evaluations.

NOTE: Scale scores range from 1 (the lowest possible outcome) to 4 (the highest possible outcome).

[1]To maximize comparability with the Kansas City, Kansas, findings, the authors use the same criteria for defining scores in the high and low categories as were used in the recently issued YDSI evaluation (Youth Development Strategies, Inc., 2004).

The First Things First Evaluation

Table 3.2

Survey Items Measuring Support and Engagement Among Teachers

Support

- Staff get professional development support from the central office.
- Administrators help staff get what they need from the central office.
- Job expectations are made clear in this school.
- Administrators support staff decision-making about students.
- Staff get resources from the central office to support work with students.
- Staff get support from administrators to do what they need to do.
- Excellence in teaching is expected.
- The central office supports staff for educational innovations they want to try.

Overall engagement

Behavioral and emotional engagement

- I look forward to going to work.
- My job has become just a matter of putting in time.
- When I am teaching I feel discouraged.
- When I am teaching I feel happy.

Reaction to challenge

- When I see something about the system that I think is not good for kids…
  - I let somebody else deal with it.
  - I talk to all people involved.
  - I ignore it.
- If I didn't like the way a staff member was handling a student…
  - I would talk to a staff member and try to straighten it out.
  - I would ignore it.

Collective engagement

- Staff don't give up when difficulties arise.
- Staff do what is necessary to get the job done right.
- Staff go beyond the call of duty to do the best job they can.

SOURCE: The 1999 measurement report for First Things First evaluations.

NOTE: Scale scores range from 1 (the lowest possible outcome) to 4 (the highest possible outcome).

An important caution is in order: Any changes in these observed *outcomes* cannot be said to indicate the *impacts* of FTF. Evaluators use the term "impacts" (or its cousin, "effects") when there is a causal relationship between an initiative and outcomes — that is, when the initiative caused the change in outcomes. As is discussed more fully in Chapter 4, making such statements about causality requires the existence of a *counterfactual:* an estimate of what would have happened if the initiative had not been in place. In this evaluation, the best counterfactual for establishing what would have happened without FTF on the measures of teacher and student support and engagement is the change over the same period that occurred in schools resembling the FTF schools that did not mount the intervention. Because the study design did not call for teacher and student surveys to be administered in non-FTF schools, there is no counterfactual against which changes in support and engagement in the FTF schools can be assessed.[2] Thus, this chapter addresses changes in these outcomes over time but cannot determine whether, or to what extent, FTF was responsible for these changes, whether positive or negative.

The chapter's key findings follow:

- There was little change in the average level of support reported by teachers in either middle or high schools, or in the proportions of teachers reporting high or low levels of support, between the planning year and the second implementation year.
- Over the same time period, average scores on a measure of teacher engagement increased significantly for high school teachers; for middle school teachers, the increase in average scores was not statistically significant. The proportions of teachers in both middle and high schools who displayed low engagement decreased significantly.
- High school and middle school students both reported significantly higher levels of support from their teachers in the second implementation year than in the planning year. The percentage of both high school and middle school students receiving low support decreased over time.
- Mean student engagement scores present a more mixed picture: a decrease in engagement for all high school students, an increase for all middle school students, and little change among middle school students in the 2001 cohort schools.

[2]Consequently, if, for example, schools like the FTF schools registered a downward trend in student engagement, then no change or even a slight negative change in engagement levels in the FTF schools could be interpreted as a positive impact of the reform.

- Expressed in terms of effect sizes, most of the differences — including those that are statistically significant — are quite small.
- The pattern of findings for the expansion sites shows a number of differences from the pattern in Kansas City, Kansas, where support and engagement outcomes generally improved for both students and teachers at both the middle and the high school level.

The rest of this chapter describes these findings in further detail and then reflects on their meaning.

Support and Engagement Among Teachers

As noted above, support and engagement are analyzed in two different ways in this chapter. In both sets of analyses, two years of follow-up data are available for teachers in all schools, while three years of data are available for teachers in the 2001 cohort schools. The analyses further distinguish between teachers in high schools and those in middle schools.

Table 3.3 presents average scores on the measures of teachers' experiences of support and engagement between the planning year and the last year for which follow-up data are available. Table 3.4 displays the percentage of teachers each year between the planning year and the second implementation year, across all expansion-site high schools and middle schools, who registered especially high or low scores on each outcome. Table 3.5 resembles Table 3.4 but extends this analysis through the third implementation year for teachers in the 2001 cohort schools.

Support

Table 3.3 indicates that, compared with the planning-year levels, there were no statistically significant changes in the level of support registered by either high school or middle school teachers in either the second or third year of follow-up.[3] Nor were there any statistically

[3]In this chapter and Chapter 4, differences are described as "statistically significant" if they are unlikely to have arisen by chance. Three levels of statistical significance are identified: Differences are significant at the 10 percent level if the probability that they arose by chance is 1 in 10 or less, at the 5 percent level if the probability that they arose by chance is 1 in 20 or less, and at the 1 percent level if the probability that they arose by chance is 1 in 100 or less.

In comparing mean scores for both teachers and students, the analysis employed an independent group difference design. In fact, there is considerable, but incomplete, overlap in the groups of teachers and students who responded to the annual surveys that are the source of data on support and engagement. The use of an independent group difference design somewhat overstates the standard errors associated with these means and, therefore, somewhat understates the statistical significance of differences that may be found.

The First Things First Evaluation

Table 3.3

Teachers' Average Scale Scores: Support and Engagement

Schools	Planning Year	Year 1	Year 2	Year 3	Effect Size Year 2	Effect Size Year 3
Support						
All high schools	2.74	2.72	2.77	NA	0.05	NA
2001 cohort high schools	2.75	2.78	2.77	2.73	0.03	-0.03
All middle schools	2.80	2.74	2.78	NA	-0.04	NA
2001 cohort middle schools	2.87	2.75	2.82	2.87	-0.09	0.00
Engagement						
All high schools	3.00	3.00	3.07	NA	0.15 **	NA
2001 cohort high schools	3.01	3.04	3.07	3.05	0.13 *	0.08
All middle schools	3.02	3.00	3.07	NA	0.12	NA
2001 cohort middle schools	3.04	2.98	3.09	3.10	0.12	0.15

SOURCES: MDRC calculations based on 2001, 2002, 2003, and 2004 First Things First staff surveys.

NOTES: Responses are limited to classroom staff only.

Scale scores range from 1 (the lowest possible outcome) to 4 (the highest possible outcome).

Statistical significance levels are indicated as *** = 1 percent; ** = 5 percent; * = 10 percent. Statistical significance is indicated for differences between the planning year and the second and third years of implementation.

"Effect size" is a metric used to describe the magnitude of a difference. Effect sizes between 0 and 0.32 may be considered small.

The size of the sample used to measure changes in support and engagement among teachers ranged from 553 to 580 across all high schools and from 354 to 370 across all middle schools between the planning year and the second implementation year.

The size of the sample used to measure changes in support and engagement among teachers ranged from 316 to 364 across 2001 cohort high schools and from 147 to 164 across 2001 cohort middle schools between the planning year and the third implementation year.

significant changes over time in the proportion of either high school or middle school teachers displaying high or low levels of support, as is evident from Tables 3.4 and 3.5.

Engagement

Table 3.3 shows that mean levels of engagement increased between the planning year and the second implementation year; the increase was statistically significant for high school teachers but not for middle school teachers. Mean engagement scores rose and then fell again for high school teachers in the 2001 cohort, so that the level of engagement reported for this group at the end of the follow-up period was no longer significantly higher than it had been during the planning year.

The data in Table 3.4 indicate that the percentages of both middle schools and high schools in the low category of engagement decreased significantly by the second implementation year. The percentage of both high school middle school teachers in the high-engagement category rose, but this increase was statistically significant only for middle school teachers. Finally, Table 3.5 shows that the proportion of highly engaged high school teachers increased significantly after FTF had been in place for three years at their schools; otherwise, no significant differences between the planning year and the third follow-up year were observed.

Support and Engagement Among Students

Table 3.6 is analogous to Table 3.3. It shows average scores on measures of support and engagement registered by students at all high schools and middle schools from the planning year through the second implementation year. It also shows scores for students at schools in the 2001 cohort through the third implementation year. Table 3.7 is akin to Table 3.4: It displays the percentage of students each year between the planning year and the second implementation year who had especially high or low scores on the scales of support from teachers and engagement. Finally, Table 3.8 parallels Table 3.5: It extends the analysis through the third implementation year for students in the 2001 cohort schools.

Support from Teachers

As Table 3.6 makes clear, both middle school and high school students registered significantly higher average levels of support from their teachers during the last year of follow-up than they had during the planning year. Table 3.7 indicates that, between the planning year and the second implementation year, there was a significant increase in the proportion of high school students receiving high support from teachers. In addition, the proportion of students in the low-support category declined significantly after two years for both high school and middle

The First Things First Evaluation

Table 3.4

Percentage of Teachers in High and Low Categories of Support and Engagement: All Schools

	Planning Year	Year 1	Year 2
Support			
All high schools			
High	13.2	12.6	11.1
Low	71.3	69.5	70.2
All middle schools			
High	12.1	9.8	13.6
Low	67.9	71.4	67.0
Engagement			
All high schools			
High	14.8	15.6	17.9
Low	52.0	51.4	46.7 *
All middle schools			
High	15.1	15.7	19.7 *
Low	55.2	52.6	44.5 ***

SOURCES: MDRC calculations based on 2001, 2002, 2003, and 2004 First Things First staff surveys.

NOTES: Responses are limited to classroom staff only.

High support is the percentage of teachers scoring 3.5 or higher on the scale measuring support; low support is the percentage of teachers scoring 3.0 or lower on the scale measuring support. High engagement is the percentage of teachers scoring 3.5 or higher on the scale measuring engagement; low engagement is the percentage of teachers scoring 3.0 or lower on the scale measuring engagement.

Statistical significance levels are indicated as *** = 1 percent; ** = 5 percent; * = 10 percent. Statistical significance is indicated for differences between the planning year and the third year of implementation.

The size of the sample used to measure changes in support and engagement among teachers ranged from 553 to 580 across all high schools and from 354 to 370 across all middle schools between the planning year and the second implementation year.

The First Things First Evaluation

Table 3.5

Percentage of Teachers in High and Low Categories of Support and Engagement: 2001 Cohort Schools

	Planning Year	Year 1	Year 2	Year 3
Support				
2001 cohort high schools				
High	13.5	14.7	11.7	11.0
Low	68.7	66.9	68.9	69.8
2001 cohort middle schools				
High	14.0	9.3	13.4	14.3
Low	61.6	71.5	65.7	68.3
Engagement				
2001 cohort high schools				
High	14.4	17.6	18.8	21.8 ***
Low	50.8	48.2	45.7	49.6
2001 cohort middle schools				
High	15.3	14.0	21.5	22.2
Low	53.1	56.4	39.5	47.9

SOURCES: MDRC calculations based on 2001, 2002, 2003, and 2004 First Things First staff surveys.

NOTES: Responses are limited to classroom staff only.

High support is the percentage of teachers scoring 3.5 or higher on the scale measuring support; low support is the percentage of teachers scoring 3.0 or lower on the scale measuring support. High engagement is the percentage of teachers scoring 3.5 or higher on the scale measuring engagement; low engagement is the percentage of teachers scoring 3.0 or lower on the scale measuring engagement.

Statistical significance levels are indicated as *** = 1 percent; ** = 5 percent; * = 10 percent. Statistical significance is indicated for differences between the planning year and the third year of implementation.

The size of the sample used to measure changes in support and engagement among teachers ranged from 316 to 364 across 2001 cohort high schools and from 147 to 164 across 2001 cohort middle schools between the planning year and third implementation year.

school students. The same patterns were found after three years among students at schools in the 2001 cohort (Table 3.8). All these changes were in the desired direction.

Engagement

Student engagement changed in different ways for high school and middle school students. For high school students, this change was in the "wrong" direction: Table 3.6 shows a decrease in average levels of engagement among students at all high schools between the planning year and the second follow-up year, and among students in the 2001 cohort high schools

The First Things First Evaluation

Table 3.6

Students' Average Scale Scores: Support from Teachers and Engagement

Schools	Planning Year	Year 1	Year 2	Year 3	Effect Size Year 2	Effect Size Year 3
Support from Teachers						
All high schools	2.75	2.79	2.81	NA	0.10 ***	NA
2001 cohort high schools	2.78	2.85	2.87	2.86	0.16 ***	0.14 ***
All middle schools	2.73	2.76	2.76	NA	0.05 **	NA
2001 cohort middle schools	2.83	2.82	2.81	2.87	0.03	0.07 ***
Engagement						
All high schools	3.27	3.22	3.22	NA	-0.11 ***	NA
2001 cohort high schools	3.34	3.27	3.27	3.26	-0.16 ***	-0.17 ***
All middle schools	3.18	3.21	3.23	NA	0.10 ***	NA
2001 cohort middle schools	3.29	3.21	3.26	3.29	-0.06	0.00

SOURCES: MDRC calculations based on 2001, 2002, 2003, and 2004 First Things First student surveys.

NOTES: Scale scores range from 1 (the lowest possible outcome) to 4 (the highest possible outcome).

Statistical significance levels are indicated as *** = 1 percent; ** = 5 percent; * = 10 percent. Statistical significance is indicated for differences between the planning year and the second and third years of implementation.

"Effect size" is a metric used to describe the magnitude of a difference. Effect sizes between 0 and 0.32 may be considered small.

The size of the sample used to measure changes in support from teachers and engagement among students ranged from 7,209 to 7,877 across all high schools and from 5,438 to 5,699 across all middle schools between the planning year and the second implementation year.

The size of the sample used to measure changes in support from teachers and engagement among students ranged from 4,535 to 4,615 across 2001 cohort high schools and from 2,322 to 2,472 across 2001 cohort middle schools between the planning year and the third implementation year.

The First Things First Evaluation

Table 3.7

Percentage of Students in High and Low Categories of Support from Teachers and Engagement: All Schools

	Planning Year	Year 1	Year 2
Support from Teachers			
All high schools			
High	22.4	24.2	24.7 ***
Low	36.4	33.6	32.7 ***
All middle schools			
High	24.7	24.1	25.9
Low	40.2	37.0	36.4 ***
Engagement			
All high schools			
High	20.3	16.9	17.6 ***
Low	31.1	35.5	36.3 ***
All middle schools			
High	16.5	17.1	17.9 *
Low	37.4	36.0	34.6 ***

SOURCES: MDRC calculations based on 2001, 2002, 2003, and 2004 First Things First student surveys.

NOTES: High support is the percentage of students scoring 3.25 or higher on the scale measuring support from teachers; low support is the percentage of students scoring 2.5 or lower on the scale measuring support from teachers. High engagement is the percentage of students scoring 3.75 or higher on the scale measuring engagement; low engagement is the percentage of students scoring 3.0 or lower on the scale measuring engagement.

Statistical significance levels are indicated as *** = 1 percent; ** = 5 percent; * = 10 percent. Statistical significance is indicated for differences between the planning year and the third year of implementation.

The size of the sample used to measure changes in support from teachers and engagement among students ranged from 7,209 to 7,877 across all high schools and from 5,438 to 5,699 across all middle schools between the planning year and the second implementation year.

The First Things First Evaluation

Table 3.8

Percentage of Students in High and Low Categories of Support from Teachers and Engagement: 2001 Cohort Schools

	Planning Year	Year 1	Year 2	Year 3
Support from Teachers				
2001 cohort high schools				
High	24.3	27.6	29.5	28.7 ***
Low	34.8	29.5	29.4	30.6 ***
2001 cohort middle schools				
High	29.4	27.2	29.1	30.8
Low	34.0	33.3	32.6	30.2 ***
Engagement				
2001 cohort high schools				
High	24.7	18.8	20.6	19.6 ***
Low	26.4	30.7	32.4	33.2 ***
2001 cohort middle schools				
High	22.0	16.7	18.4	19.8 *
Low	28.5	35.0	31.1	29.0

SOURCES: MDRC calculations based on 2001, 2002, 2003, and 2004 First Things First student surveys.

NOTES: High support is the percentage of students scoring 3.25 or higher on the scale measuring support from teachers; low support is the percentage of students scoring 2.5 or lower on the scale measuring support from teachers. High engagement is the percentage of students scoring 3.75 or higher on the scale measuring engagement; low engagement is the percentage of students scoring 3.0 or lower on the scale measuring engagement.

Statistical significance levels are indicated as *** = 1 percent; ** = 5 percent; * = 10 percent. Statistical significance is indicated for differences between the planning year and the second year of implementation.

The size of the sample used to measure changes in support from teachers and engagement among students ranged from 4,535 to 4,615 across 2001 cohort high schools and from 2,322 to 2,472 across 2001 cohort middle schools between the planning year and the third implementation year.

between the planning year and the third follow-up year. The data in Tables 3.7 and 3.8 point to a decrease in the percentage of high school students with high engagement and an increase in the percentage with low engagement.

In contrast, when data for all middle school students are examined through the first two years of follow-up, changes in the "right" direction are evident. As seen in Table 3.6, average engagement scores for students at all middle schools rose significantly between the planning year and the second follow-up year. Table 3.7 indicates that, over the same time period, there was an increase in the percentage of middle school students in the high-engagement category and a decrease in the percentage in the low-engagement category.

The findings for middle-school students in the 2001 cohort schools are somewhat different. These students started off with higher average levels of engagement than students in the 2002 cohort schools. During the first year of implementation, their engagement scores declined; scores then increased again, so that, by the end of the follow-up period, they were back to the planning-year levels. In the third year, the percentage of middle school students in 2001 cohort schools reporting high engagement was significantly lower than it had been during the planning year.

Interpreting the Findings

Why scores on the measure of support among teachers changed so little is not clear. One possibility is that four of the eight items tapping teachers' feelings of support relate to support presented by the central office. In two districts, the central office may have seemed a rather distant and uninvolved presence to teachers in the study schools. And while the central office in a third district is located next door to one of the middle schools, interviews with teachers in that district suggest that the turnover in district leadership left many teachers there feeling uncertain about the new administration's intentions with regard to FTF.

On the other hand, teachers' levels of engagement changed in the desired direction. In particular, the proportions of middle and high school teachers displaying low engagement decreased — perhaps because teachers who were initially feeling "burned out" were reinvigorated by FTF and the energy of their colleagues. Too, it may be that — precisely in the face of the stress that inevitably accompanies implementation of an initiative requiring major changes — teachers recognized and were buoyed by the effort that they and their colleagues had made.

It seems likely that the small learning community (SLC) structure and the family advocacy component left students feeling better known and more cared about than had been true in the past. But widespread increases in students' feelings of support from their teachers were not matched by similar increases in engagement, especially among high school students. The implementation findings in Chapter 2 suggest one possible explanation for this disparity. They

indicate that, in this relatively early implementation period, teachers generally made more progress in establishing personalized relationships with the students in their SLCs and in implementing the Family Advocate System than they did in improving instruction. With respect to this last key element of FTF, teachers made increased use of active learning strategies aimed at increasing student involvement, but the lessons they taught were not very challenging and seldom included thematic content. For high school students to be deeply involved in their learning, instructional improvement may have to progress further than was the case at this relatively early point in program implementation.

To determine the impacts of FTF, Chapter 4 examines changes in student attendance, persistence, achievement, and other outcomes both at the FTF schools and at a set of comparison schools. The findings in the present chapter suggest that positive major changes in these outcomes may not be evident at this early stage if — as the FTF theory of change posits — increased support and engagement among students are both preconditions for improved student performance. But changes over time in the measures of teacher engagement hint at the possibility that as teachers become more invested in delivering challenging instruction, student engagement and performance will also increase.

Chapter 4

The Impacts of First Things First on Student Outcomes

This chapter assesses the success of First Things First (FTF) at improving outcomes for students. To do so, it examines the *impacts* of the school reform on these outcomes. "Impacts" are defined as changes in outcomes that were caused by the reform, above and beyond other changes that occurred. The first section of the chapter describes how these impacts were measured. The following sections then present impact findings for each site. The findings suggest that:

- Middle and high school students in Kansas City, Kansas, registered large gains on a wide range of academic outcomes that were sustained over several years and were pervasive across the district's schools; similar gains were not present in the most comparable schools in the state. The improvements occurred over the course of eight years of substantial effort by the school district and by the Institute for Research and Reform in Education (IRRE) to implement FTF as the district's central educational reform.[1] Findings include increased rates of student attendance and graduation, reduced student dropout rates, and improved student performance on the state tests of reading and mathematics. The measured impacts on student test scores reflected double-digit increases in the percentage of students who scored at levels deemed "proficient" by the state and double-digit reductions in the percentage of students scoring at levels deemed "unsatisfactory."

- There were limited signs of early positive impacts at some of the reform's expansion sites, which had been implementing FTF for two to three years. But given the widely varying implementation experiences of these sites and the short follow-up period available for their evaluation, it is not yet clear whether the expansion sites will replicate the positive findings for Kansas City.

Estimating Impacts

In principle, the impact of FTF on a student outcome equals the *difference* between what the outcome was after the school reform was under way and what it would have been without the reform. In practice, one can estimate this difference by comparing the change over

[1]Similar conclusions were reached in a recently completed evaluation by Gambone et al. (2004).

time in a student outcome for schools that adopted the reform (FTF schools) with the corresponding change for similar comparison schools that did not adopt it (the "counterfactual").[2] Variants of this approach were used for each of the sites in the evaluation. Thus, all impact estimates represent the observed improvement of FTF schools relative to the observed improvement of their comparison schools. Appendix C provides more detail on how impact estimates were obtained for each site, by describing the student outcome measures used, how comparison schools were chosen, and the statistical models used.

The Basic Approach

Ideally the time-series design used to produce impact estimates for the present report should have data on consistently measured student outcomes for multiple pre-intervention baseline years, multiple post-intervention follow-up years, multiple FTF schools, and multiple comparison schools.[3] The first step in estimating impacts from these data is to measure the change at FTF schools in the outcome for a given follow-up year relative to its average level during the baseline period. This represents how student performance changed in the presence of FTF. The next step is to measure the corresponding change for comparison schools. This provides an estimate of how student performance would have changed at the FTF schools in the absence of the reform. The *difference* between these two changes is an estimate of the impact of FTF — what the initiative caused to happen.

To be more concrete, consider the following hypothetical example. Assume that during a three-year baseline period before FTF was launched at a site, 50 percent of the tenth-grade students at its FTF schools and 55 percent of the tenth-grade students at its comparison schools passed their high-stakes state test in mathematics. Also assume that, during the third year of FTF implementation at the site, 70 percent of the students at the FTF schools and 60 percent of the students at the comparison schools passed the test. Hence, within three years of launching the reform, there was a 20 percentage point improvement at the FTF schools and a 5 percentage point improvement at the comparison schools. The difference — 15 percentage points — is an estimate of the improvement caused by the reform.

---

[2]For a description of this approach — which is referred to as "short interrupted time-series analysis" — see Bloom (2003).

[3]Multiple baseline years help to provide a reliable benchmark of pre-intervention outcomes by averaging random year-to-year fluctuations in student outcomes. Multiple follow-up years help to provide the elapsed time needed for a reform to be implemented and thus to begin to take effect. Multiple FTF schools help to provide a reliable measure of change over time in the presence of the reform. This reliability stems from (1) the ability of multischool averages to reduce random year-to-year fluctuations in student outcomes and (2) their ability to "dampen the shocks" that can occur at a single school due to idiosyncratic local events, such as a change in principal. For the same reasons, multiple comparison schools can help to provide a reliable basis for estimating the change over time in student outcomes that would have occurred without the reform.

Figure 4.1 illustrates how the basic evaluation design was adapted to accommodate the constraints and data availability of each site. These adaptations represent "variations on a theme." The discussion begins with Houston, the site where the most complete version of the evaluation design was implemented, and it continues through the other sites, where specific local constraints prevented use of some aspects of the ideal design.

The Evaluation Design for Houston

Panel A of Figure 4.1 illustrates the Houston design. Houston had individual student data on outcomes during three pre-intervention baseline years (denoted in the figure by dark rings) for all FTF and comparison schools, two post-intervention follow-up years (denoted by solid vertical lines) for all these schools, and a third post-intervention year (denoted by a dashed vertical line) for the one high school and one middle school that launched FTF a year before the others.

There were three FTF high schools with five to eleven comparison schools each and four FTF middle schools with three to fifteen comparison schools each.[4] Comparison schools were chosen from the Houston Independent School District to match the past test scores of students at each FTF school as closely as possible. In addition, as described in Appendix C, available data on the demographic characteristics and past test scores of individual students were used to statistically adjust for compositional shifts over time in the background characteristics of schools' student populations (which were not substantial). The impacts of FTF in Houston were thus estimated as the differences between changes over time in the adjusted mean outcomes for its FTF schools and corresponding changes for its comparison schools. The primary limitation of this design is that there were only two years of follow-up information for the full sample of schools, which may not have provided enough time for them to implement all the structural and instructional changes thought to be needed to produce impacts on student achievement.

The Evaluation Design for Riverview Gardens

Panel B in Figure 4.1 illustrates the evaluation design for Riverview Gardens, Missouri. This design is based on three years of pre-intervention baseline data and three years of post-intervention follow-up data. Thus, it looks much like the evaluation design for Houston. However, the Riverview Gardens design differs from Houston's in three important ways.

First, because FTF was put in place at all secondary schools in the Riverview Gardens school district, its comparison schools had to be selected from other urban districts in Missouri. There were eight high schools and twelve middle schools in the comparison groups for this site.

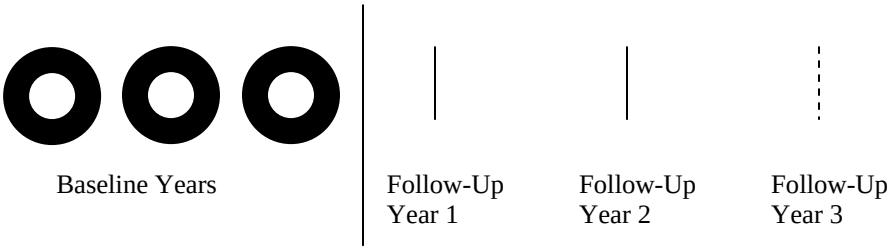
[4]Comparison schools were matched separately to each FTF school in Houston. However, given the similarities among FTF schools, their comparison groups comprise many of the same schools.

The First Things First Evaluation

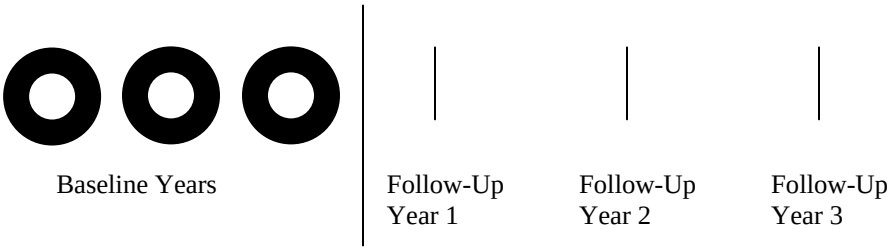
Figure 4.1

Design Diagrams for the Impact Analysis

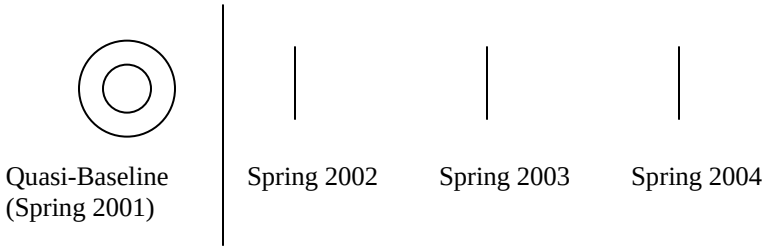
A. Houston, Texas
Student-level data, regression-adjusted for demographics and pretest (3 FTF high schools and 10 to 11 comparison schools; 4 FTF middle schools and 3 to 15 comparison schools)



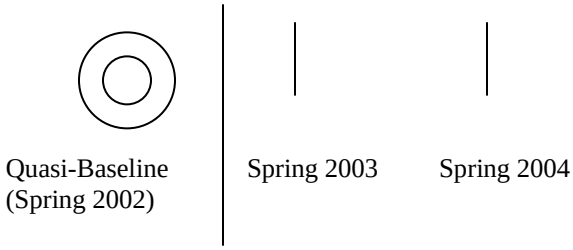
B. Riverview Gardens, Missouri
Aggregate-level data, no regression adjustments (1 FTF high school and 8 comparison schools; 1 FTF composite middle school [Central and East combined] and 12 comparison schools)



C. Kansas City, Kansas
Student-level data, regression-adjusted for demographics (4 FTF high schools and 7 comparison schools; 8 FTF middle schools and 9 comparison schools)



D. Shaw and Greenville, Mississippi
Aggregate-level data, no regression adjustments (2 FTF high schools and 4 to 10 comparison schools)



Although these comparison schools were subject to the same state-level forces as were the FTF schools, they were subject to different district-level forces. Hence, they were not as close a match in these regards as were the comparison schools for Houston.

Second, consistent outcome data are available only for each school, not for individual students.[5] It therefore was not possible to control statistically for compositional shifts in the student populations over time. Fortunately, aggregate data for the schools suggest that these shifts were not large.

Third, Riverview Gardens has only one FTF high school, and its two middle schools were treated as one "composite" school for the analysis.[6] Hence, there is less statistical power to estimate the impacts of FTF in this site than there is for sites whose findings are based on the average results for several schools (although even these sites are limited by their small numbers of schools). As illustrated later in this chapter, this means that the impacts of FTF in Riverview Gardens (with respect to one high school or one middle school) must be much larger than those in other districts (with respect to the average for two to eight schools) in order to be "statistically significant" and thus identifiable with confidence. This is a specific example of the more general fact that there is very little statistical power or precision for studying the impacts of an educational intervention at a single school or at a very small number of schools.

The Evaluation Design for Kansas City, Kansas

Unlike the first two evaluation designs, the one for Kansas City, Kansas, did not include pre-intervention baseline years, because the state test for this site (its primary source of outcome data) was changed recently. Thus, consistently measured state test scores are available only for years *after* FTF was launched.

To deal with this situation, the first administration of the new state test (spring 2001) was used as the point of reference, or benchmark, for gauging future improvements.[7] This "quasi-baseline" year is represented by a light ring in Panel C of Figure 4.1. State test scores for spring 2002, 2003, and 2004 were used to construct outcome measures for the three follow-up years (denoted by vertical lines in the figure).

[5]Individual student data are available for the most recent years in the analysis but not for earlier years.

[6]All Riverview Gardens middle school students are treated as though they came from a single "composite" school, because of changes in the number of middle schools in the district and the allocation of students to these schools.

[7]The first administration of the new test was in spring 2000. However, its sample is not comparable to that for later years because the test was not administered to a large proportion of students who had special needs, and accommodations for these students were not standardized in ways that they were subsequently.

There were four FTF high schools and eight FTF middle schools at the site. For some of these schools, the quasi-baseline year was the first year of full FTF implementation; for others, it was the second year of implementation; and for others, it was the third implementation year. For all these schools, spring 2001 test scores were used to establish a quasi-baseline level, and scores for subsequent springs were used to measure follow-up outcomes.

Because FTF was implemented throughout the Kansas City, Kansas, school district, comparison schools were selected from other urban areas in the state that were subject to the same state-level influences but different district-specific factors.[8] Unfortunately, given the Kansas City schools' high concentration of minority students who exhibited especially low performance on state tests, these schools could not be matched closely. Nevertheless, comparison groups of seven high schools and nine middle schools were identified. In addition, available data on individual student background characteristics were used to adjust for changes over time in these characteristics (which were not large).[9]

The impacts of FTF on student performance in Kansas City, Kansas, were therefore estimated as the *difference* between the statistically adjusted changes in test scores for FTF schools and comparison schools. The methodology — as applied in Kansas City — is biased against finding positive results, for two reasons. First, having only one quasi-baseline year reduces the reliability of the impact estimates relative to designs that have more stable multiyear baselines and thereby reduces the chances of obtaining impact estimates that are statistically significant. Second, using a post-intervention year as a quasi-baseline "subtracts out" any early impacts that might have been produced.[10] Nevertheless, as argued later in this section, the improvements in students' outcomes in Kansas City, Kansas, were so large, multifaceted, pervasive, and sustained — *and* the improvements in the comparison schools were so much smaller — that the findings provide strong evidence of FTF's central role in producing academic progress.

---

[8]Unlike in Houston, the high school comparison group in Kansas City, Kansas, was the same for all FTF high schools, and the middle school comparison group was the same for all FTF middle schools. Further refinements required to produce separate matches for each school in Kansas City were judged not to be warranted, given how difficult it was to find close matches for any schools.

[9]Unlike in Houston, pretest scores for individual students (their scores on standardized tests for earlier grades) were not available in Kansas City, Kansas.

[10] The difference between the FTF and comparison schools in district-specific factors does not necessarily bias the findings in any particular direction. Further, the fact that the comparison schools started out at a higher level of academic performance than the Kansas City, Kansas, schools (though they still fell below the statewide average) does not necessarily bias the results in a particular direction. Making significant improvement in very low-performing schools may be harder (because of institutional factors) or easier (because there is more "room" for improvement) than in somewhat higher-performing schools. At any rate, the analysis compares deviations from past trends in the two groups of schools, so the baseline differences in the levels of academic performance do not directly feed into the analysis.

The Evaluation Design for the Delta Region of Mississippi

The final evaluation design (see Figure 4.1, Panel D) is for FTF schools in the Delta Region of Mississippi. For a number of reasons, this design for measuring impacts is the weakest among the sites. Because a new Mississippi state test was administered for the first time in spring 2002, this year had to be used as the quasi-baseline year, which left only two years of follow-up, 2003 and 2004. Thus, the conservative nature of the quasi-baseline approach described above for Kansas City was exacerbated by the very short follow-up period for schools from the Delta Region of Mississippi. In addition, outcome data are available only for schools as a whole, not for individual students, so it was not possible to adjust for compositional shifts in the student population (although the shifts were not large). Also, for one FTF high school from this site, it was impossible to find comparison schools that were similar in *both* algebra and English II, because of the enormous differences in students' scores in these two subjects.[11] Furthermore, the school provided a very small sample, with only about 50 students tested each year in a given subject. For all these reasons, the findings for Mississippi are suggestive only.

Presenting and Interpreting the Results

The following sections present the results of the FTF impact analysis. Unlike the preceding section — where the order of sites was determined by how closely their evaluation designs approximated the desired approach — the following sections first discuss results for the original site in Kansas City, Kansas, and then present findings for the reform's expansion sites.

Findings are presented in summary bar charts, which illustrate the change in an outcome from the baseline (or quasi-baseline) period to each follow-up year, separately for FTF schools and their comparison schools. The estimated impact, which is the difference between these two changes, is listed at the top of each pair of bars. In addition, Appendix D presents tables that disaggregate the bar charts' average findings for all schools at a site into findings for individual schools. This makes it possible to assess how pervasive the observed impacts were.

Before examining the findings, it is important to note that, for several reasons, they represent a conservative test of the effectiveness of FTF. As discussed earlier, there are particular aspects of the Kansas City analysis that make it a conservative estimate. But other factors apply across all sites. First, many initiatives may have been undertaken by the comparison schools to improve their learning environments and, thereby, to increase the achievement of their students (although there was little information available about how intensive, extensive, or successful these efforts were). Thus, improvements at the FTF schools were compared with any improve-

[11]For this reason, separate comparison groups of six high schools for English II and four high schools for algebra were selected for this FTF high school. A single comparison group of ten high schools was selected for both subjects at the other FTF high school.

ments produced by initiatives that may have been implemented by the comparison schools. Consequently, the question addressed by the impact findings in this report is "By how much more did FTF improve student outcomes than would have occurred due to other efforts that would have existed without the reform initiative?"

Second, by necessity, single schools or only a few schools are the foci of the study at each site, implying a quite limited ability to detect impacts that might have been produced by FTF. Only very large effects would be statistically significant and thus could be detected with confidence. Thus, if positive impacts are estimated but are not statistically significant, it is impossible to tell whether these findings represent true impacts of the reform or random errors in estimates of these impacts. Given this situation, it is especially important to base conclusions on the overall pattern of findings across sites, not just on findings for specific sites or specific schools.

Third, for all sites except Kansas City, Kansas, the follow-up period for the present evaluation comprises only one to three years after FTF was launched. Given the complex nature of whole-school reforms like FTF — which require making a series of structural, attitudinal, behavioral, and instructional changes — a number of years are needed to complete implementation and for students to be sufficiently exposed to the changes that they can benefit from them. Although it might be reasonable to expect large improvements in certain student behaviors during the first few years of a vigorous and successful implementation, it is probably too soon to expect large improvements in student achievement, which is thought by many to be much more difficult to change.

Fourth, for Kansas City, Kansas, and the Delta Region of Mississippi, the benchmark used to gauge improvement is a quasi-baseline year after implementation of FTF began. Thus, if any impacts were produced before or during the quasi-baseline year, they would be "netted-out" of the analysis and, thus, would not be attributed to the initiative.

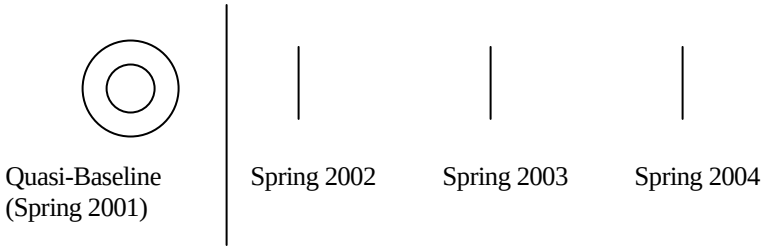
Before proceeding, one further caution is important. This has to do with the limits of the evaluation design with respect to determining what observed changes were actually caused by FTF. As noted, there are clear limits to this ability, and the limits vary considerably across sites, given the data available and the nature of their comparison schools. Thus, as conclusions are drawn throughout the chapter, they are never based just on the strength of the evaluation design and its associated statistical analysis. Instead, they are based on the full pattern of findings about how (and how successfully) FTF was implemented; on whether FTF influenced the various student outcomes that it targeted (such as attendance and dropout rates as well as test scores); and, in some cases, on whether the magnitudes of the estimated impacts are so large (and unusual) that, on their face, they provide compelling evidence that FTF must have caused them. In other words, the report bases its conclusions on the weight of the preponderance of the evidence. Acknowledging these methodological complications and the need to arrive at conclusions based on

the preponderance of the evidence (especially in Kansas City), findings about the effects of FTF are labeled "impacts" for ease of presentation.

FTF in Kansas City, Kansas

A recent long-term study of FTF in Kansas City found that the school reform initiative produced large, pervasive, and sustained impacts on a number of student outcomes.[12] The present study confirms these findings, using a somewhat different analytic approach. It also adds to the previous study by (1) using recently available data to document that the large impacts produced by FTF continued for another year (2003-2004), (2) reporting impacts on student outcomes using measures that may be more readily interpretable by policymakers (based on changes in the percentage of students whose outcomes met certain threshold conditions), and (3) measuring improvements in the FTF schools against those for a small group of comparison schools (instead of for the state as a whole).

FTF was initially launched in Kansas City, where schools have, by far, the highest concentration in the state of economically disadvantaged, minority, and low-achieving students. Because the reform was implemented at all comprehensive high schools and middle schools in the district, comparison schools had to be chosen from other urban districts in the state. The evaluation design that was used to measure impacts on student achievement for this site (illustrated in Figure 4.1 and repeated below) is based on data for one quasi-baseline year and three follow-up years.



Outcomes measures for the analysis focus on student scores in reading and math for the Kansas State Assessment plus rates of attendance, dropout, and graduation.[13] Test scores are

[12]Gambone et al., 2004.

[13]Since 1995, the Kansas City district also administered the Metropolitan Achievement Test (MAT-7) each spring in grades 1 through 11 in reading and math. This was done in response to a state mandate that all school districts in Missouri "triangulate" their measures of student performance based on three testing regimes: (1) local assessments customized to district standards and benchmarks, (2) a state test aligned with state standards and benchmarks, and (3) a nationally normed test chosen from a list of acceptable alternatives. During the past several years, the MAT-7 was moved from a spring administration to a fall administration in order to

(continued)

reported by the state in five categories: unsatisfactory, basic, proficient, advanced, and exemplary. In keeping with the categories used by Gambone and colleagues, the present analysis reports test outcomes two ways: as the percentage of students whose scores were proficient or above (representing the top three categories in the state scoring system) and as the percentage of students with unsatisfactory scores (representing the bottom category in the state scoring system.)[14] Improvements in student scores are represented by *increases* in the percentages of students whose scores are proficient or above (referred to hereafter as "proficient") and *decreases* in the percentages of students whose scores are unsatisfactory.

The pattern of results obtained from a series of analyses suggests that:

- For high schools, FTF produced sustained "double-digit" improvements in reading achievement both in terms of increasing the percentage of students whose scores were proficient and reducing the percentage whose scores were unsatisfactory. FTF also improved achievement in math, although by a smaller margin and with less consistency. In addition, FTF improved rates of student attendance, dropout, and graduation.
- For middle schools, FTF produced large improvements in reading scores, math scores, and attendance rates.

Appendix D, showing findings for individual schools, illustrates that these impacts were pervasive across schools in the district.

High School Results

This section presents estimates of the impacts of FTF on high schools in Kansas City. Findings for student achievement are presented first, followed by a discussion of findings for other student outcomes.

"make room" for the new state test. In addition, it was reduced to grades 5, 8, and 11. District officials, school principals, and teachers now pay little attention to the test, and it has become largely irrelevant to the operation of Kansas City schools. Additionally, because this test was not administered by the comparison schools, it could not be used to measure the impacts of FTF.

[14]Because the present analysis combines the top three (of five) state scoring categories to define "proficient" and uses the bottom state category to define "unsatisfactory," the remaining state category (which is between these two extremes) is not included. Thus, the percentage of students scoring proficient or unsatisfactory does not sum to 100 percent, and a change in the percentage proficient does not automatically translate into a corresponding change in the opposite direction for the percentage unsatisfactory.

Achievement

In Kansas City, Kansas, the state high school assessment is administered to tenth-graders in math and to eleventh-graders in reading. Table 4.1 lists the percentage of students whose reading or math scores were unsatisfactory each year for the FTF schools, their comparison schools, and the state. These findings provide part of the "raw material" for the present analysis.[15]

The First Things First Evaluation

Table 4.1

High School State Assessment Test Scores for First Things First Schools, Comparison Schools, and the State: Kansas City, Kansas

	Percentage Unsatisfactory			
	Quasi-Baseline Year	Follow-Up Years		
	Spring 2001	Spring 2002	Spring 2003	Spring 2004
11th-grade reading				
FTF schools	52.8	52.9	41.1	31.0
Comparison schools	22.5	29.0	22.5	20.0
State	15.3	17.0	13.3	11.5
10th-grade math				
FTF schools	74.6	65.5	69.8	56.5
Comparison schools	40.6	43.4	42.4	32.3
State	26.6	27.3	26.5	20.6

SOURCE: MDRC calculations from individual student records from a statewide data file.

NOTE: Because FTF was implemented before the administration of the new test, spring 2001 represents the third year of implementation for one school, the second year of implementation for another school, and the first year of implementation for two other schools.

As can be seen, during the quasi-baseline year, 52.8 percent of the eleventh-grade FTF students had unsatisfactory reading scores, and 74.6 percent of the tenth-grade FTF students had unsatisfactory math scores. Corresponding rates for comparison schools were 22.5 percent and 40.6 percent, respectively. The differences between the FTF schools and the comparison

[15]Each section on findings for high schools or middle schools includes a simple table that describes student performance over time with respect to one (and occasionally two) outcomes. These findings are presented selectively in order to reduce them to a manageable number.

schools illustrate the fact that the especially high concentration of economically disadvantaged, minority, and low-performing students in Kansas City made it very difficult to find closely matched comparison schools. Corresponding rates for the state as a whole (15.3 percent and 26.6 percent) were even further removed from those for Kansas City.

The results in Table 4.1 (plus those presented below) are so striking that, even given the limitations of the evaluation design and the preexisting differences between the FTF schools and comparison schools, they provide strong evidence that FTF improved student achievement in Kansas City high schools substantially. According to these findings, FTF schools experienced a dramatic decline in their percentage of students with unsatisfactory test scores (from 52.8 percent to 31.0 percent for reading and from 74.6 percent to 56.5 percent for math). At the same time, there was no corresponding change for reading scores at the comparison schools and a moderate delayed improvement for math. Even smaller changes were observed for the state as a whole, which had much less margin for improvement. Thus, between 2001 and 2004, FTF high schools closed a large portion of their "performance gap" relative to other schools.

Figure 4.2 presents estimates of the impacts of FTF on reading achievement, obtained using the analytic approach described in Appendix C. The following is a step-by-step explanation of how to interpret these findings, along with those in similar figures used throughout this chapter.

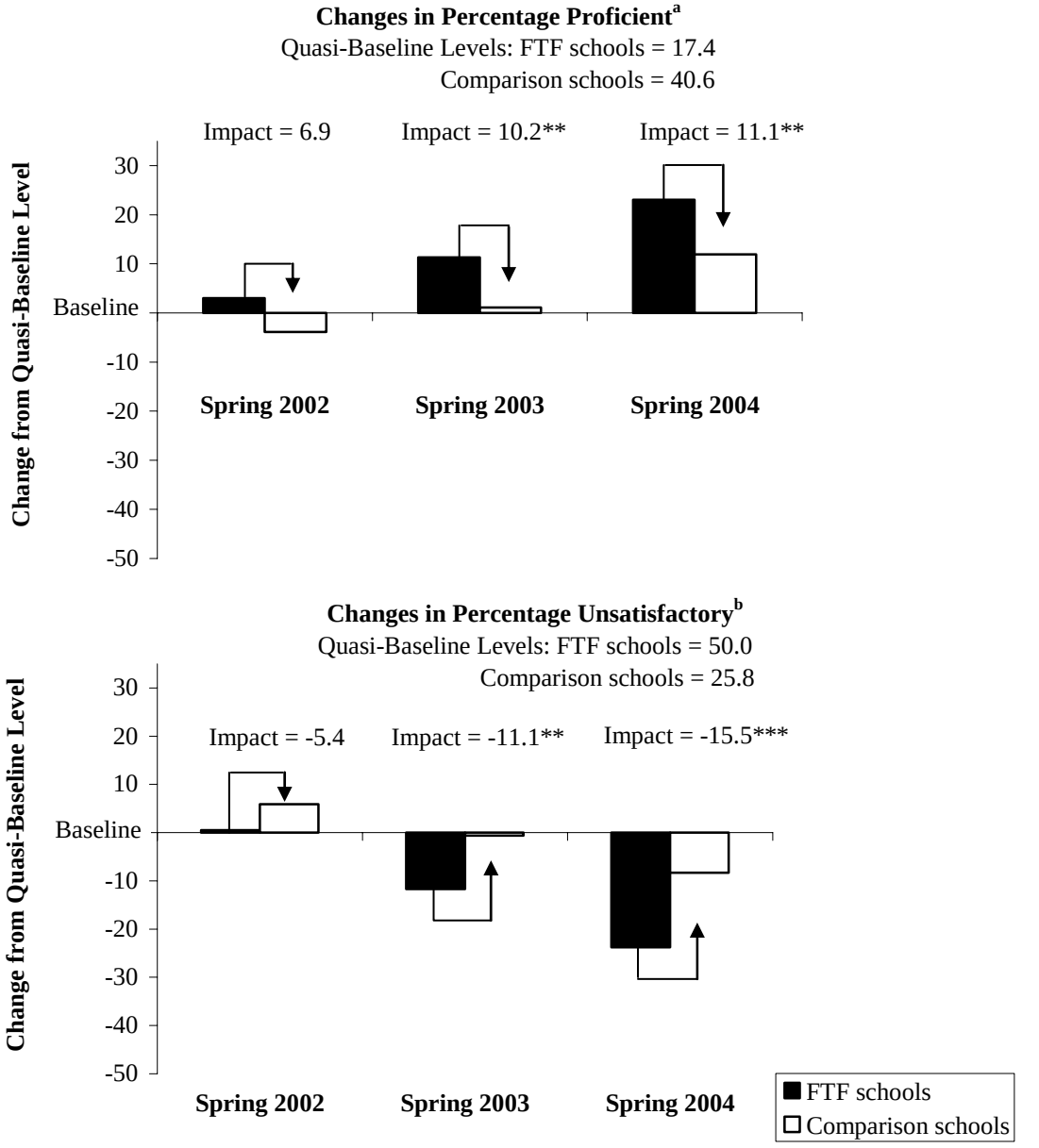
Each dark bar in the figure represents the percentage point change for FTF schools from the baseline period (the quasi-baseline year in Kansas City) to a given follow-up year. The light bars represent corresponding changes for comparison schools. The difference between these two changes for a given outcome and follow-up year — which is listed above each pair of bars — is the estimated impact of FTF. The statistical significance of this estimate is denoted by the number of stars next to it (with an absence of stars indicating an absence of statistical significance). The top panel in the figure examines the extent to which FTF increased rates of proficient performance. The bottom panel examines the extent to which the reform reduced rates of unsatisfactory performance. Each panel reports below the title the quasi-baseline levels for the FTF schools and comparison schools to identify their respective starting points for the analysis. All baseline levels and changes from these levels reported in Figure 4.2 adjust for students' background characteristics, as described in Appendix C. The findings are similar but not identical to the findings reported in Table 4.1, which do not adjust for background characteristics.

Consider the findings for spring 2004 in the top panel of Figure 4.2. The large dark bar indicates that the percentage of students scoring proficient at the FTF schools increased by 23.1 points from the schools' quasi-baseline level of 17.4 percent. The smaller light bar denotes that the percentage of students scoring proficient at the comparison schools increased by 12.0 points from the quasi-baseline level of 40.6 percent for these schools. The difference between these

The First Things First Evaluation

Figure 4.2

Changes from Quasi-Baseline Levels in the Percentage of 11th-Graders Scoring Proficient or Unsatisfactory on the State Reading Test: Kansas City, Kansas



(continued)

Figure 4.2 (continued)

SOURCE: MDRC calculations from individual student records from a statewide data file.

NOTES: Sample includes 11th-grade students from four First Things First (FTF) high schools and seven comparison schools. Students in the sample consist of test-takers for whom administrative records exist between the 2000-2001 and 2003-2004 academic years.

"Proficient" is defined as the sum of the top three performance categories on the state test: exemplary, advanced, and proficient. "Unsatisfactory" refers only to the bottom category: unsatisfactory.

Each bar represents the "deviation from quasi-baseline," or the difference between the quasi-baseline level (average in spring 2001) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the quasi-baseline for FTF schools and the deviation from the quasi-baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics.

A two-tailed t-test was applied to differences in deviations from quasi-baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in the percentage of students scoring in the state's top three performance categories.

[b]The desired change in this measure is a decrease from baseline, which represents a decrease in the percentage of students scoring in the state's bottom performance category.

two changes indicates that FTF *increased* the percentage of students whose scores were proficient, by 11.1 percentage points beyond that which would have occurred without the reform. This impact estimate is statistically significant, and thus one can be confident that it represents a true change in student achievement, not just a random change due to the sampling of students or measurement error.[16]

Findings for spring 2004 in the bottom panel of Figure 4.2 tell a complementary story. The dark bar indicates that the percentage of students scoring unsatisfactory at the FTF schools declined by 23.8 percent from their baseline level of 50.0 percent. The light bar indicates that the percentage of students scoring unsatisfactory at the comparison schools declined by 8.3

[16]Given the statistical properties of the student outcome data and evaluation design used for the present analysis, there is adequate precision to detect average effects of FTF at the *four* high schools in Kansas City that are achievable and educationally meaningful. More specifically, the minimum detectable effects of the reform are a change of 12.7 points in the percentage of students scoring proficient and a change of 6.4 points in the percentage of students scoring unsatisfactory. These changes represent the smallest effects that, if produced, would have an 80 percent chance of being detected (80 percent statistical power) using a two-tail test of statistical significance at the 5 percent level. Impact estimates for *individual* schools that are, by necessity, reported for other sites have much less precision. Therefore, the evaluation is much less able to detect impacts produced by the reform at these sites.

from their baseline level of 25.8 percent. The difference between these two declines indicates that FTF *reduced* the rate of unsatisfactory performance by 15.5 percentage points beyond what would have occurred without the reform. This estimate is statistically significant.

The overall pattern of findings in the figure clearly indicates that FTF markedly improved reading achievement at Kansas City high schools. These findings represent a progression over time in which improvements began to emerge in 2002, were well-established by 2003, and continued to grow in 2004.

Figure 4.3 reports the impacts of FTF on student performance on the state math test. These findings suggest that the reform did, in fact, improve math performance. But the improvement for math was less pronounced and less consistent than that for reading. In addition, it seems to have been concentrated among the lowest-performing students, because observed impacts were almost solely in terms of reducing the percentage of scores that were unsatisfactory. For this outcome, impacts of –10.8 and –6.7 percentage points (which were statistically significant) were observed for the first two follow-up years, and those of –5.2 percentage points (which was not statistically significant) were obtained for the last follow-up year.

Attendance, Dropout, and Graduation

Although state test scores are available only since spring 2001 for Kansas City, Kansas, earlier information exists for other student outcomes. Because of this, the evaluation design for measuring FTF impacts on high school attendance, dropout rates, and graduation rates differs somewhat from that for measuring test-score impacts. The starting level for each of these other outcomes is its average value for three academic years, 1997-1998, 1998-1999, and 1999-2000. For schools that implemented FTF in 2000-2001, this represents a true pre-intervention baseline period. For schools that began implementation in 1998-1999 or 1999-2000, the period represents a mix of pre- and post-intervention quasi-baseline years. For all schools in the analysis, there is a *four-year follow-up period*, which begins in 2000-2001.

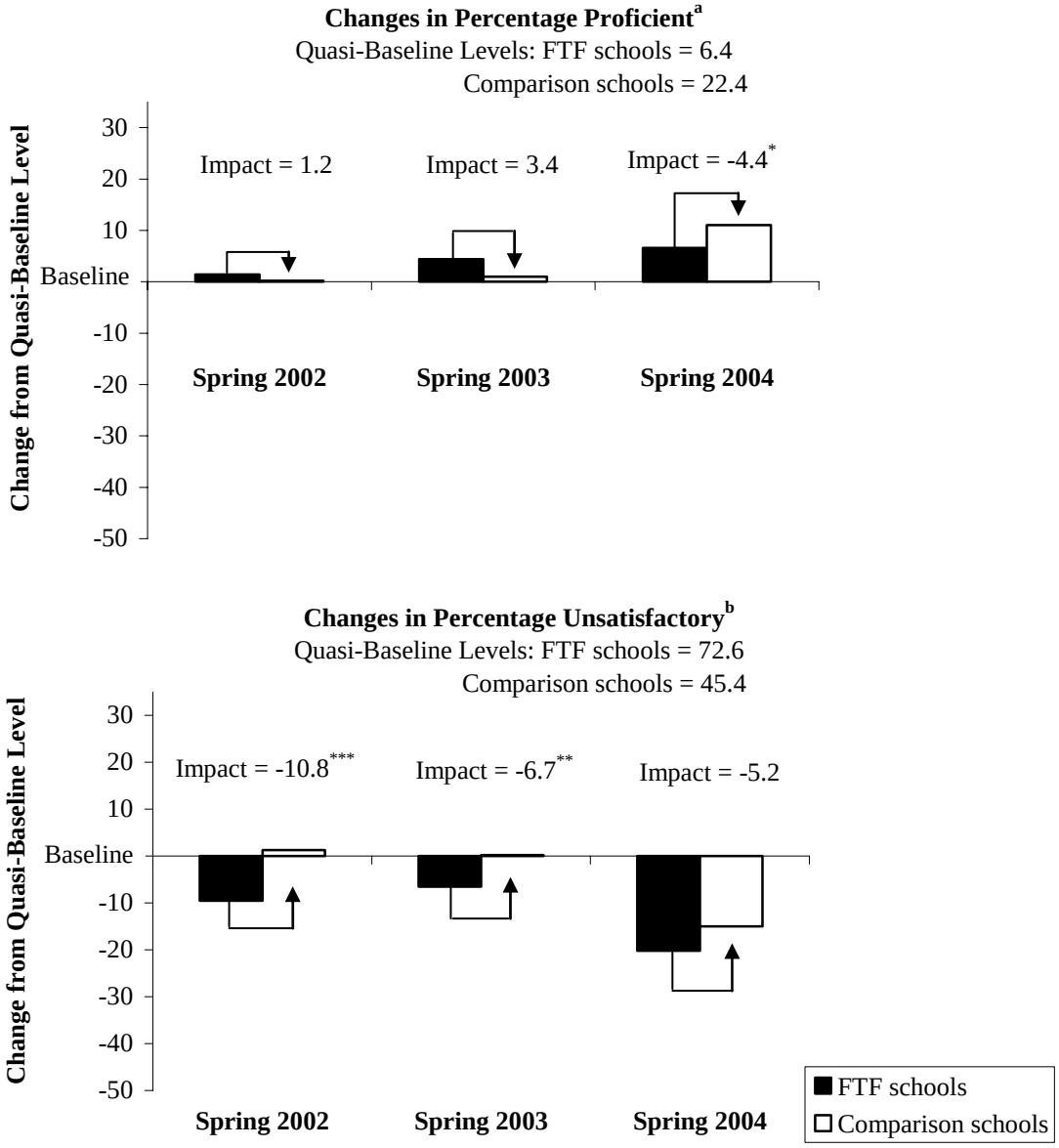
Data on rates of attendance, dropout, and graduation are available only at the school level, not for individual students.[17] Thus, for estimating impacts on these outcomes, it was not possible to adjust for shifts over time in students' background characteristics. However, because these shifts were not large, it is unlikely that adjustments for them would have changed the findings appreciably.

[17] Attendance, dropout, and graduation data for each school were taken from the Kansas State Department of Education Web site (www.ksde.org), as reported by schools.

The First Things First Evaluation

Figure 4.3

Changes from Quasi-Baseline Levels in the Percentage of 10th-Graders Scoring Proficient or Unsatisfactory on the State Math Test: Kansas City, Kansas



(continued)

Figure 4.3 (continued)

SOURCE: MDRC calculations from individual student records from a statewide data file.

NOTES: Sample includes 10th-grade students from four First Things First (FTF) high schools and seven comparison schools. Students in the sample consist of test-takers for whom administrative records exist between the 2000-2001 and 2003-2004 academic years.

"Proficient" is defined as the sum of the top three performance categories on the state test: exemplary, advanced, and proficient. "Unsatisfactory" refers only to the bottom category: unsatisfactory.

Each bar represents the "deviation from quasi-baseline," or the difference between the quasi-baseline level (average in spring 2001) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the quasi-baseline for FTF schools and the deviation from the quasi-baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics.

A two-tailed t-test was applied to differences in deviations from quasi-baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in the percentage of students scoring in the state's top three performance categories.

[b]The desired change in this measure is a decrease from baseline, which represents a decrease in the percentage of students scoring in the state's bottom performance category.

Figure 4.4 reports estimates of the impacts of FTF on rates of attendance, dropout, and graduation for the four follow-up years in the analysis. These findings indicate that the reform improved all three outcomes in Kansas City, Kansas.

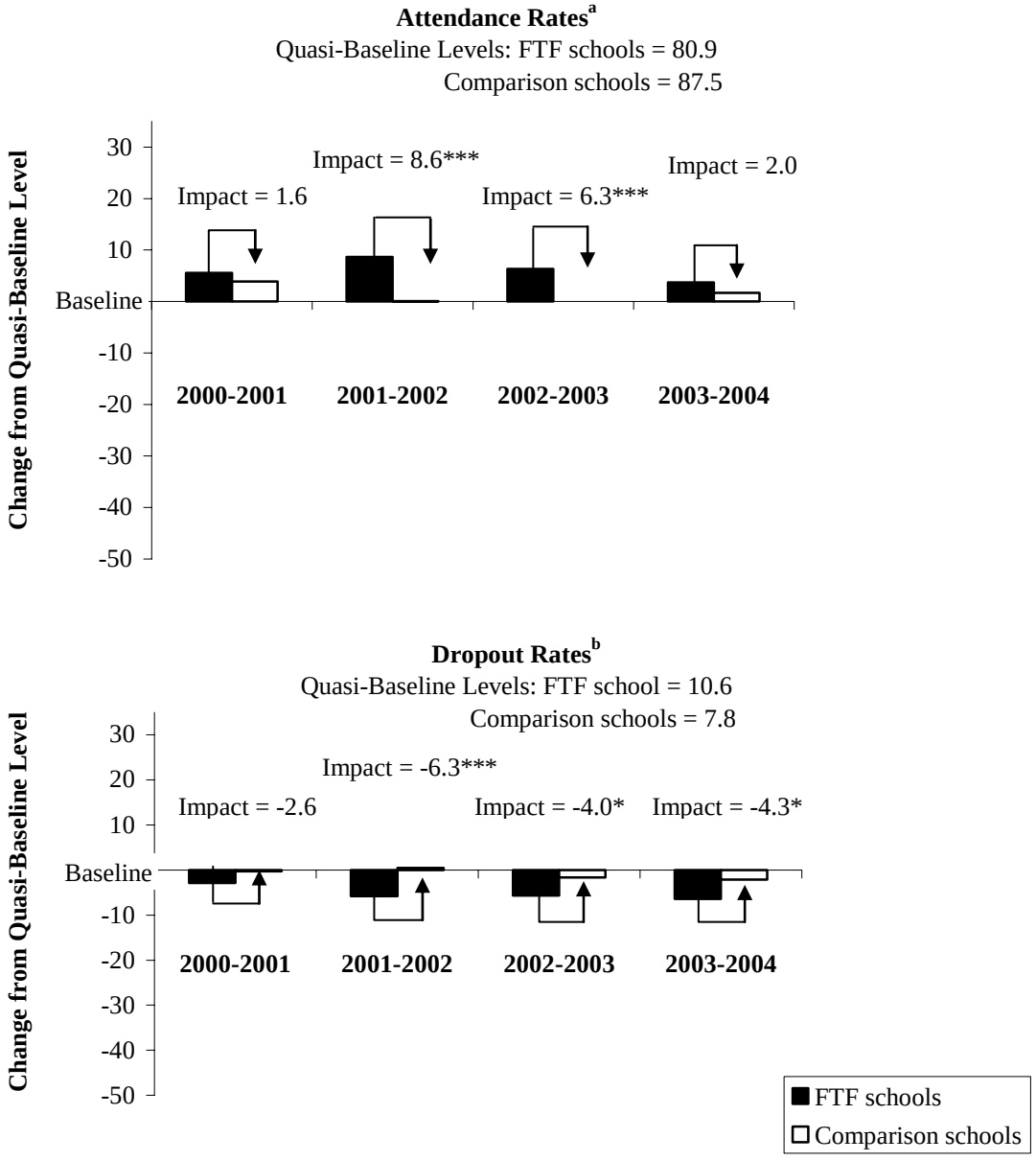
The top panel of the figure displays annual rates of student attendance.[18] During the three-year quasi-baseline period for the analysis, the average attendance rates were 80.9 percent for FTF schools and 87.5 percent for comparison schools. Thus, attendance problems were initially more serious for FTF schools. The positive dark bars in the figure indicate the subsequent increases that occurred in attendance rates for FTF schools. The smaller positive light bars indicate corresponding changes for comparison schools. The differences between these two sets of changes represent estimated impacts of FTF on attendance, which ranged from relative improvements of 1.7 to 8.6 percentage points (impacts that are statistically significant in two of four years). Thus, attendance improved by more at FTF schools than at comparison schools, although not by a consistent margin.

[18]Attendance is measured each year as the total number of days of student attendance reported for all grades divided by the total number of days of recorded student enrollment (multiplied by 100).

The First Things First Evaluation

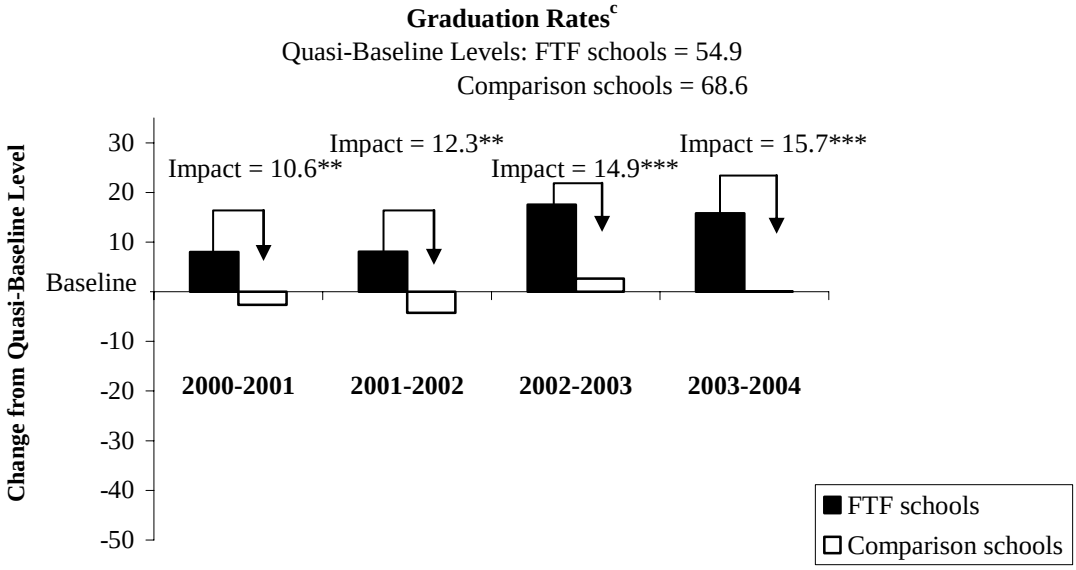
Figure 4.4

Changes from Quasi-Baseline Levels in High School Attendance, Dropout, and Graduation Rates: Kansas City, Kansas



(continued)

Figure 4.4 (continued)



SOURCE: MDRC calculations from school-level records of state data.

NOTES: Sample includes four First Things First (FTF) high schools and seven comparison schools.

Each bar represents the "deviation from quasi-baseline," or the difference between the quasi-baseline level (average of three prior school years) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the quasi-baseline for FTF schools and the deviation from the quasi-baseline for comparison schools.

A two-tailed t-test was applied to differences in deviations from quasi-baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in school-level attendance rates.

[b]The desired change in this measure is a decrease from baseline, which represents a decrease in school-level dropout rates.

[c]The desired change in this measure is an increase from baseline, which represents an increase in school-level graduation rates.

The second panel of Figure 4.4 is for student dropout rates.[19] According to data obtained from and calculated by the State of Kansas (which are reported annually by each school), the average baseline dropout rate for FTF schools was 10.6 percent, and the rate for comparison schools was 7.8 percent. The dark negative bars in the chart indicate the amount by which this rate declined during the follow-up period for FTF schools, and the light bars indicate the amount of corresponding change for comparison schools. The differences between these two changes for each follow-up year — which range from –2.6 to –6.3 percentage points (and are statistically significant in three of four years) — indicate the degree to which FTF reduced dropout rates.

The third panel of Figure 4.4 shows high school graduation rates.[20] During the three-year quasi-baseline period for the analysis, these graduation rates were, on average, 54.9 percent for FTF schools and 68.6 percent for comparison schools. Thus, initially, a much smaller percentage of entering students were graduating from FTF schools. The large positive dark bars in the chart indicate the large increases in graduation rates that occurred for subsequent ninth-grade cohorts at FTF schools. The small, fluctuating positive and negative light bars indicate the modest and inconsistent changes that occurred during the same period at comparison schools. The differences between these two sets of bars — which range from 10.6 to 15.7 percentage points (and are statistically significant) — indicate that FTF in Kansas City produced double-digit increases in graduation rates.

Middle School Results

This section examines the impacts of FTF on the state test scores and attendance of middle school students in Kansas City, Kansas.

Achievement

The Kansas State Assessment for middle schools is administered in math to seventh-graders and in reading to eighth-graders. Scores on these tests are reported in the same five categories used for high schools. Thus, the analysis of middle school test scores examines the extent to which FTF *increased* the percentage that were proficient (that is, were in one of the top three state categories) and *reduced* the percentage that were unsatisfactory (were in the bottom state category).

[19]Dropout rates are defined cumulatively for each entering cohort of ninth-grade students as they proceed (or not) through high school during the next four years. Thus, the dropout rate for 2003-2004 is defined to represent the cumulative rate for students who entered ninth grade in 2000-2001.

[20]Graduation rates represent the percentage of students in each entering ninth-grade cohort who graduated by the end of the summer four years later.

As a first step, Table 4.2 lists the percentage of middle school students each year whose scores were unsatisfactory. These findings are even more striking than the findings for high schools. As can be seen, test scores from FTF middle schools improved dramatically in both reading and math. For example, the percentage of scores that were unsatisfactory declined between the quasi-baseline year (2001) and the last follow-up year (2004) from 38.5 percent to 13.7 percent in reading and from 60.8 percent to 37.9 percent in math. Corresponding scores for comparison schools remained roughly constant during the first two follow-up years and then improved. This was also the case for the state as a whole. Thus, it appears that few changes in overall state test performance occurred between spring 2001 and spring 2003. However, in spring 2004, scores improved for comparison schools and the state. Nevertheless during all follow-up years, test scores improved by far more for FTF schools than for comparison schools or the state.

The First Things First Evaluation

Table 4.2

Middle School State Assessment Test Scores for First Things First Schools, Comparison Schools, and the State: Kansas City, Kansas

	Percentage Unsatisfactory			
	Quasi-Baseline Year	Follow-Up Years		
	Spring 2001	Spring 2002	Spring 2003	Spring 2004
8th-grade reading				
FTF schools	38.5	32.0	16.4	13.7
Comparison schools	23.4	21.9	24.3	14.4
State	11.1	11.3	8.8	6.5
7th-grade math				
FTF schools	60.8	55.6	49.3	37.9
Comparison schools	39.1	43.1	41.9	28.7
State	20.8	20.4	18.6	14.4

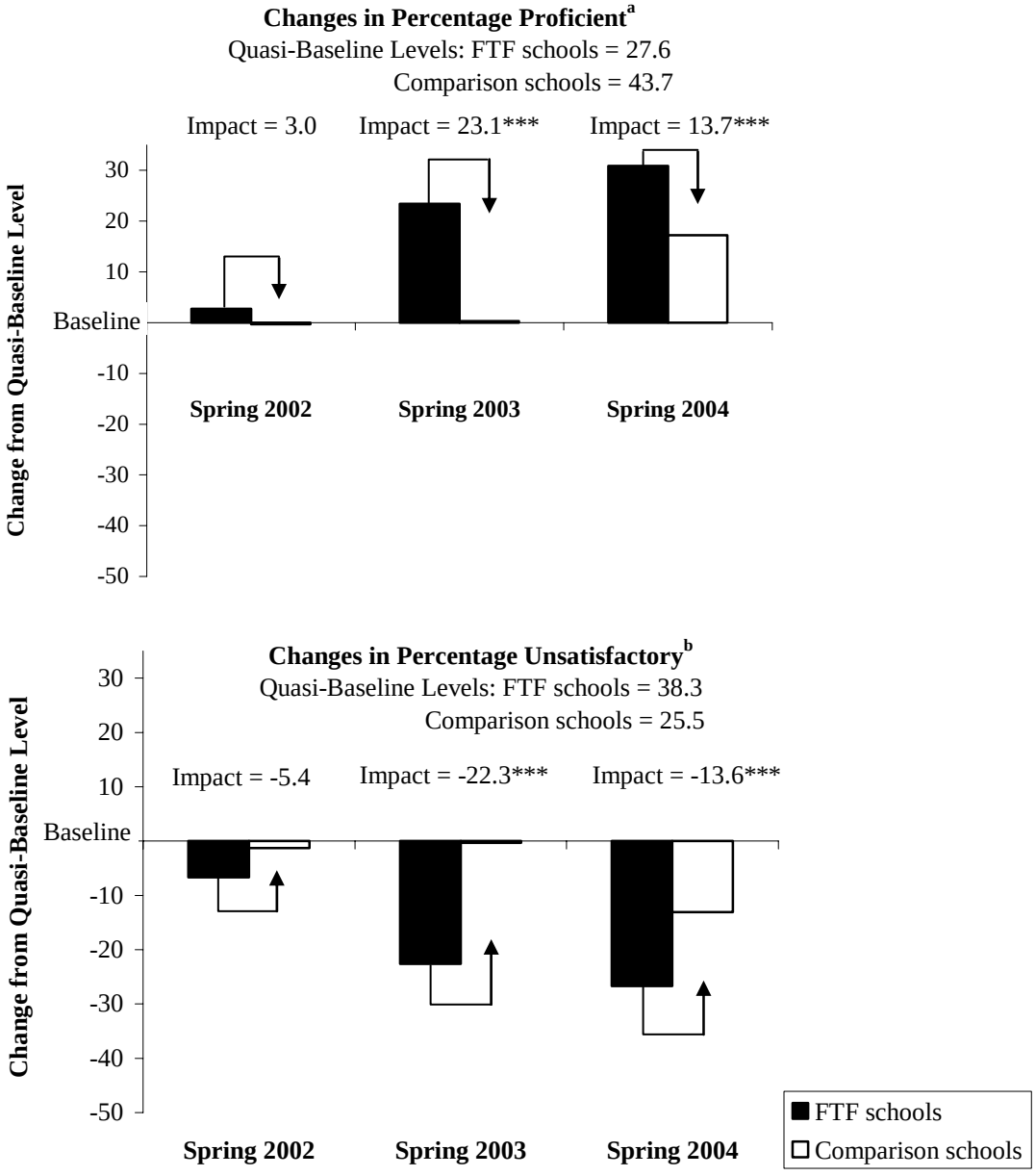
SOURCE: MDRC calculations from individual student records from a statewide data file.

NOTE: Because FTF was implemented before the administration of the new test, spring 2001 represents the third year of implementation for two schools, the second year of implementation for another two schools, and the first year of implementation for four schools.

The First Things First Evaluation

Figure 4.5

Changes from Quasi-Baseline Levels in the Percentage of 8th-Graders Scoring Proficient or Unsatisfactory on the State Reading Test: Kansas City, Kansas



(continued)

Figure 4.5 (continued)

SOURCE: MDRC calculations from individual student records from a statewide data file.

NOTES: Sample includes 8th-grade students from eight First Things First (FTF) middle schools and nine comparison schools. Students in the sample consist of test-takers for whom administrative records exist between the 2000-2001 and 2003-2004 academic years.

"Proficient" is defined as the sum of the top three performance categories on the state test: exemplary, advanced, and proficient. "Unsatisfactory" refers only to the bottom category: unsatisfactory.

Each bar represents the "deviation from quasi-baseline," or the difference between the quasi-baseline level (average in spring 2001) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the quasi-baseline for FTF schools and the deviation from the quasi-baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics.

A two-tailed t-test was applied to differences in deviations from quasi-baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in the percentage of students scoring in the state's top three performance categories.

[b]The desired change in this measure is a decrease from baseline, which represents a decrease in the percentage of students scoring in the state's bottom performance category.

Figure 4.5 reports statistical estimates of the impacts of FTF on reading scores. The large positive dark bars in the top panel of the figure and the large negative dark bars in the bottom panel illustrate the striking improvements that occurred at FTF schools during 2003 and 2004. These improvements were much larger than their comparison school counterparts (designated by light bars). The differences indicate that FTF increased the incidence of proficient scores and reduced the incidence of unsatisfactory scores by 13.6 to 23.1 percentage points beyond what would have occurred without the reform.[21]

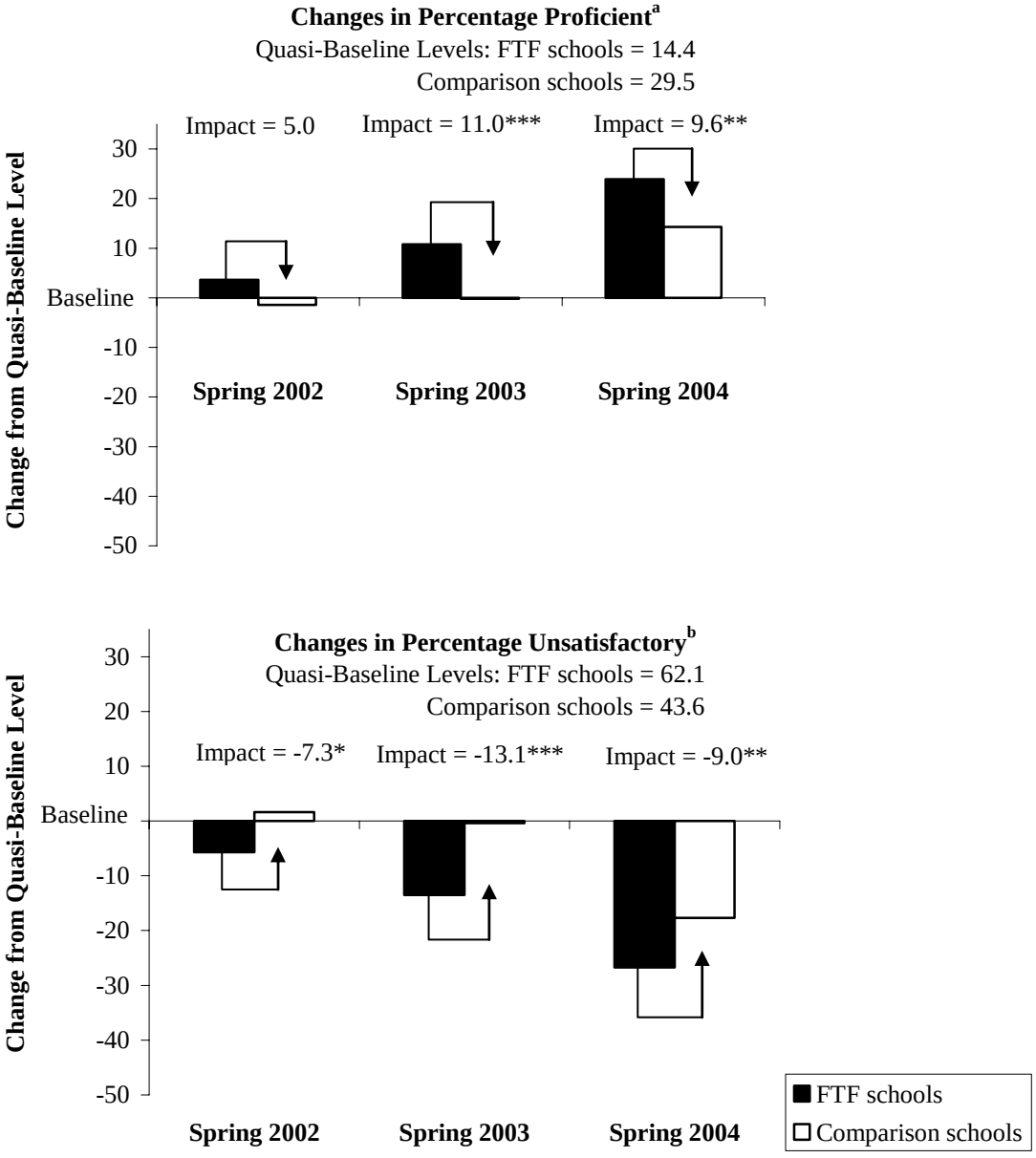
Figure 4.6 indicates that improvements in math scores that were caused by FTF emerged earlier than improvements in reading scores. In 2002, there was a 5.0 percentage point relative increase in the incidence of proficient scores and a 7.3 percentage point relative reduction in the incidence of unsatisfactory scores. These were followed by even larger relative improvements in 2003 and 2004. During these last two years, it appears that FTF increased the incidence of proficient scores and reduced the incidence of unsatisfactory scores by 9.0 to 13.1 percentage points more than would have occurred without the reform.

[21]The minimum detectable effects of FTF on reading scores were a change of 14.2 points in the percentage of scores that were proficient and a change of 10.0 points in the percentage of scores that were unsatisfactory.

The First Things First Evaluation

Figure 4.6

Changes from Quasi-Baseline Levels in the Percentage of 7th-Graders Scoring Proficient or Unsatisfactory on the State Math Test: Kansas City, Kansas



(continued)

Figure 4.6 (continued)

SOURCE: MDRC calculations from individual student records from a statewide data file.

NOTES: Sample includes 7th-grade students from eight First Things First (FTF) middle schools and nine comparison schools. Students in the sample consist of test-takers for whom administrative records exist between the 2000-2001 and 2003-2004 academic years.

"Proficient" is defined as the sum of the top three performance categories on the state test: exemplary, advanced, and proficient. "Unsatisfactory" refers only to the bottom category: unsatisfactory.

Each bar represents the "deviation from quasi-baseline," or the difference between the quasi-baseline level (average in spring 2001) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the quasi-baseline for FTF schools and the deviation from the quasi-baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics.

A two-tailed t-test was applied to differences in deviations from quasi-baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in the percentage of students scoring in the state's top three performance categories.

[b]The desired change in this measure is a decrease from baseline, which represents a decrease in the percentage of students scoring in the state's bottom performance category.

Attendance

Figure 4.7 presents findings for middle school attendance. Even though average rates of attendance were substantial during the three-year quasi-baseline period (90.5 percent for FTF schools and 90.1 percent for comparison schools), they improved somewhat thereafter — more so for FTF schools than for comparison schools. This produced a relative improvement of 2.5 percentage points during the last three follow-up years.

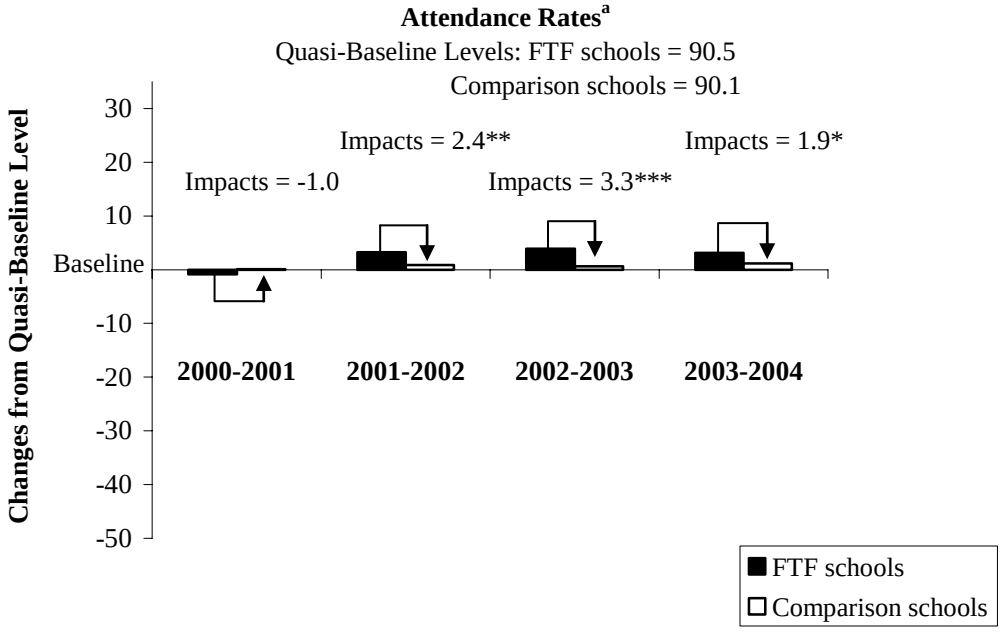
Kansas City Summary

Though application of the impact methodology in Kansas City faced real challenges, the breadth, consistency, pervasiveness across schools (as shown in Appendix D), and size of the effects found here — which are consistent with the theory of action and actual implementation of the initiative — provide a compelling case for the success of FTF in this initial site. The analysis now turns to the replication sites.

The First Things First Evaluation

Figure 4.7

Changes from Quasi-Baseline Levels in Middle School Attendance Rates: Kansas City, Kansas



SOURCE: MDRC calculations from school-level records of state data.

NOTES: Sample includes eight First Things First (FTF) middle schools and nine comparison schools.

Each bar represents the "deviation from quasi-baseline," or the difference between the quasi-baseline level (average of three prior school years) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the quasi-baseline for FTF schools and the deviation from the quasi-baseline for comparison schools.

A two-tailed t-test was applied to differences in deviations from quasi-baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in school-level attendance rates.

FTF in Houston, Texas

FTF was phased in at the Houston Independent School District in two waves. During the first year, the reform was launched at one high school and one middle school (the 2001 cohort schools). During the second year, FTF was launched at two more high schools and three more middle schools (the 2002 cohort schools). Because of this timing, follow-up years in the present analysis do not represent the same calendars years for all schools. Instead, for each school they are measured relative to the launch of FTF at that school.[22]

The evaluation design for Houston (presented in Figure 4.1 and repeated below) is based on data for three pre-intervention baseline years for all schools plus three follow-up years for the 2001 cohort school and two follow-up years for the 2002 cohort schools. (Appendix D shows findings for individual schools among the 2002 cohort.)

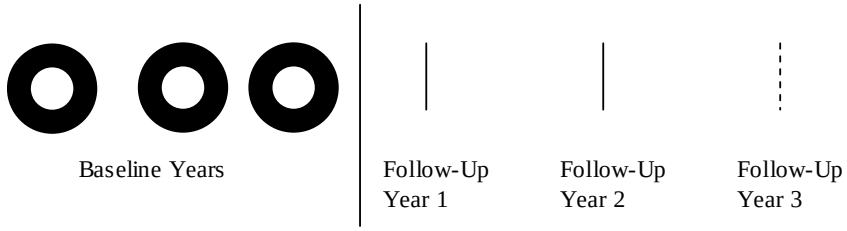
The impact analysis for FTF in Houston focuses mainly on students' performance in reading and math on the Texas state test[23] plus measures of student attendance and persistence (the extent to which they remained in school as opposed to dropping out or leaving the school district).[24] Impact findings are also presented for students' performance on the nationally normed Stanford Achievement Test (SAT-9), which is administered each year by the Houston district at the FTF schools and comparison schools.[25] These latter findings are given less emphasis, however, because the district deemphasized the SAT-9 when the new state test began.

[22]The first follow-up year for a school is the first year of implementation of FTF in that school.

[23]For many years before FTF was implemented in Houston, the state administered the Texas Assessment of Academic Skills (TAAS). In the spring of 2003, the state switched to the Texas Assessment of Knowledge (TAKS). Because the new test is more difficult and has higher standards than the previous test, fewer students were expected to pass it. Even fewer students were expected to pass the test when the threshold for doing so was raised again in 2004.

[24]Attendance rates in Houston high schools and middle schools are calculated for each student by dividing the total number of days present by the total number of days enrolled. Individual student rates are then averaged for all students at each school. Persistence rates in Houston high schools are calculated as the percentages of ninth-grade students who were still in school in the Houston school district at any time during the following academic year.

[25]The SAT-9 was administered through spring 2003. In spring 2004, the SAT-10 was administered. For the present analysis, SAT-10 normal curve equivalents (NCEs) were converted to SAT-9 NCEs, in order to make scores comparable across years. For ease of reference, the test is referred to as "SAT-9" throughout. The SAT-9 is administered to grades 6 through 11. For this analysis, eighth-grade scores are used to represent middle school outcomes and tenth-grade scores are used to represent high school outcomes.



.

Results of the analysis suggest that:

- FTF markedly improved student performance on the state test for the 2001 cohort high school, School E. This finding is consistent with a wide range of qualitative and quantitative indicators that point to School E's exceptionally strong record of implementation. Corresponding improvements were not observed for the other two FTF high schools or for other student outcomes.
- For middle schools, the findings do not indicate that FTF improved student outcomes at the Houston expansion site.

High School Results

The effects of FTF on high school student achievement, attendance, and persistence (from ninth grade to the next year) are described in this section.

Overview of Achievement

Table 4.3 provides a simple summary of the test performance of high school students from FTF schools, comparison schools, the Houston district, and the State of Texas (for the state test only). The top panel of the table presents results for the state test, reported as the percentage of students who passed the test. The bottom panel of the table presents results for the SAT-9, reported as the percentage of students who scored above the 50th percentile for all students nationally. Within each panel, test results are presented separately for the 2001 cohort FTF school (with its comparison schools) and for the 2002 cohort FTF schools (with their comparison schools), in order to maintain the distinction between their follow-up years (spring 2002-2004 for the early-cohort school versus spring 2003-2004 for the later-cohort schools). Given that baseline scores for the two tests are not highly correlated, a separate group of comparison schools (one matched on prior scores for the state test and one matched on prior scores for the SAT-9) was selected for each FTF high school, to estimate impacts on scores for each test.

As can be seen, passing rates on the state test for the FTF schools are lower than rates for the district, which, in turn, are lower than rates for the state as a whole. This highlights the fact that FTF in Houston was targeted to schools that had been experiencing especially serious

The First Things First Evaluation

Table 4.3

High School Test Results for First Things First Schools, Comparison Schools, the District, and the State: Houston, Texas

	Spring 1999	Spring 2000	Spring 2001	Spring 2002	Spring 2003	Spring 2004
	Percentage Passing TAAS/TAKS State Test					
10th-grade reading						
				3 Follow-Up Years		
2001 cohort FTF school	72.7	72.3	62.7	77.8	54.1	51.9
Comparison schools	76.1	81.7	78.8	91.3	55.7	60.8
					2 Follow-Up Years	
2002 cohort FTF schools		76.9	83.8	87.2	50.4	58.8
Comparison schools		80.9	78.1	90.3	52.7	55.1
District	82.9	86.4	85.9	92.6	62.1	66.6
State	88.0	90.0	90.0	94.0	72.0	75.0
10th-grade math						
				3 Follow-Up Years		
2001 cohort FTF school	70.0	73.8	72.2	73.9	50.6	35.4
Comparison schools	67.8	75.2	78.2	84.3	49.5	36.2
					2 Follow-Up Years	
2002 cohort FTF schools		73.5	82.8	87.3	47.9	36.1
Comparison schools		74.3	78.4	85.4	48.0	29.8
District	76.0	82.5	85.6	89.5	62.2	50.4
State	81.0	86.0	89.0	92.0	73.0	63.0
	Percentage At/Above 50th Percentile on Nationally Normed Test					
10th-grade reading						
				3 Follow-Up Years		
2001 cohort FTF school	29.6	29.7	12.8	13.2	11.5	21.4
Comparison schools	22.2	25.1	16.4	17.6	16.9	29.8
					2 Follow-Up Years	
2002 cohort FTF schools	21.8	24.0	18.1	19.3	16.5	21.7
Comparison schools	23.0	25.2	17.2	18.1	16.8	29.2
District	34.0	37.7	31.3	33.5	30.2	40.4
10th-grade math						
				3 Follow-Up Years		
2001 cohort FTF school	35.1	42.8	26.8	23.6	18.1	25.9
Comparison schools	25.0	34.6	26.5	25.5	18.9	29.9
					2 Follow-Up Years	
2002 cohort FTF schools	24.7	33.6	32.4	28.9	24.9	28.4
Comparison schools	26.6	35.1	27.8	26.2	19.5	29.3
District	37.4	46.9	42.4	40.7	33.7	42.1

(continued)

Table 4.3 (continued)

SOURCES: MDRC calculations from individual student records from the Houston Independent School District data file for FTF schools, comparison schools, and district results; TAAS and TAKS performance results reported on the Texas Education Agency website (www.tea.state.tx.us) for state results.

NOTE: FTF was implemented in one high school in the fall of 2001 (2001 cohort) and in two additional high schools in the fall of 2002 (2002 cohort). Therefore, spring 2002 is a follow-up year for one school and a baseline year for the other two schools.

Boxed areas represent follow-up years.

difficulties. A second overall pattern in the data is that passing rates for all groups of schools were rising during the baseline period. These rates dropped precipitously, however, when the new, more difficult state test was first administered, in 2003. Rates dropped even further in 2004 (for math), when the threshold for passing was raised again.

Two points are thus very important to note when viewing the performance of FTF in Houston through the lens of these state test results. First is the fact that lower pass rates on the new test relative to the old one do not imply a reduction in student achievement. Instead, they reflect an increase in the difficulty of the test and its standards for passing. Second is the fact that although changes in the test created considerable measurement error, it was still possible to use test results to identify large FTF impacts, if they existed.

The findings in Table 4.3 suggest that such a large impact was produced by the 2001 cohort high school, School E. The baseline passing rates for this school were consistently below those of its comparison schools. However, after FTF and the new state test were implemented, passing rates for the FTF school were similar (and in some cases almost identical) to those for its comparison schools. Thus, the FTF school closed a substantial preexisting "achievement gap." Findings for the other FTF high schools do not suggest similar impacts for them, at least not during the first two years of FTF implementation.

The SAT-9 results in the table also indicate that FTF schools and their comparison schools were among the lower-performing schools in Houston, which, in turn, was performing below the national average. But there is no sign in these data of an impact of FTF for either the 2001 cohort FTF school or the 2002 cohort schools. Results for the SAT-9 seem to fluctuate from year to year for the FTF schools, their comparison schools, and the district as a whole. But there is no clear change from baseline levels for any group.

Now consider the findings of a more detailed statistical analysis of these data plus those for other important student outcomes. These findings are presented first for the 2001 cohort high school and then for the two 2002 cohort high schools.

The 2001 Cohort High School: School E

Achievement

Figure 4.8 presents estimates of the impacts of FTF on state test scores for School E. As explained earlier, these estimates adjust for individual student differences in background characteristics and seventh-grade test scores in reading or math. Results for reading are presented in the top panel, and results for math are presented in the bottom panel. The two charts tell the same basic story, although the impact findings for reading are statistically significant, whereas those for math are not (because there is more year-to-year fluctuation in math scores and, thus, less statistical precision for their analysis).[26]

On average, during the three-year baseline period, 59.5 percent of the students at School E passed the state test in reading, whereas 72.6 percent of the students at the comparison schools did so. During the first follow-up year (2002), passing rates increased for both sets of schools, although those for School E rose by 4.9 percentage points more than those for its comparison schools (a difference that was not statistically significant). The next year, when the new, more difficult and demanding state test was administered for the first time, passing rates for both groups of schools declined substantially. However, those for School E declined by 12.5 percentage points *less* than did those for its comparison schools (a difference that was statistically significant). In the final year, as standards for passing the state test were raised yet again, passing rates at School E and its comparison schools dropped even further. However, once again, rates for School E declined by 8.8 percentage points less than did those for its comparison schools (a difference that was also statistically significant). Thus, it appears that as FTF was being implemented at School E, it made substantial progress toward closing the performance gap between its students and those from its comparison schools.

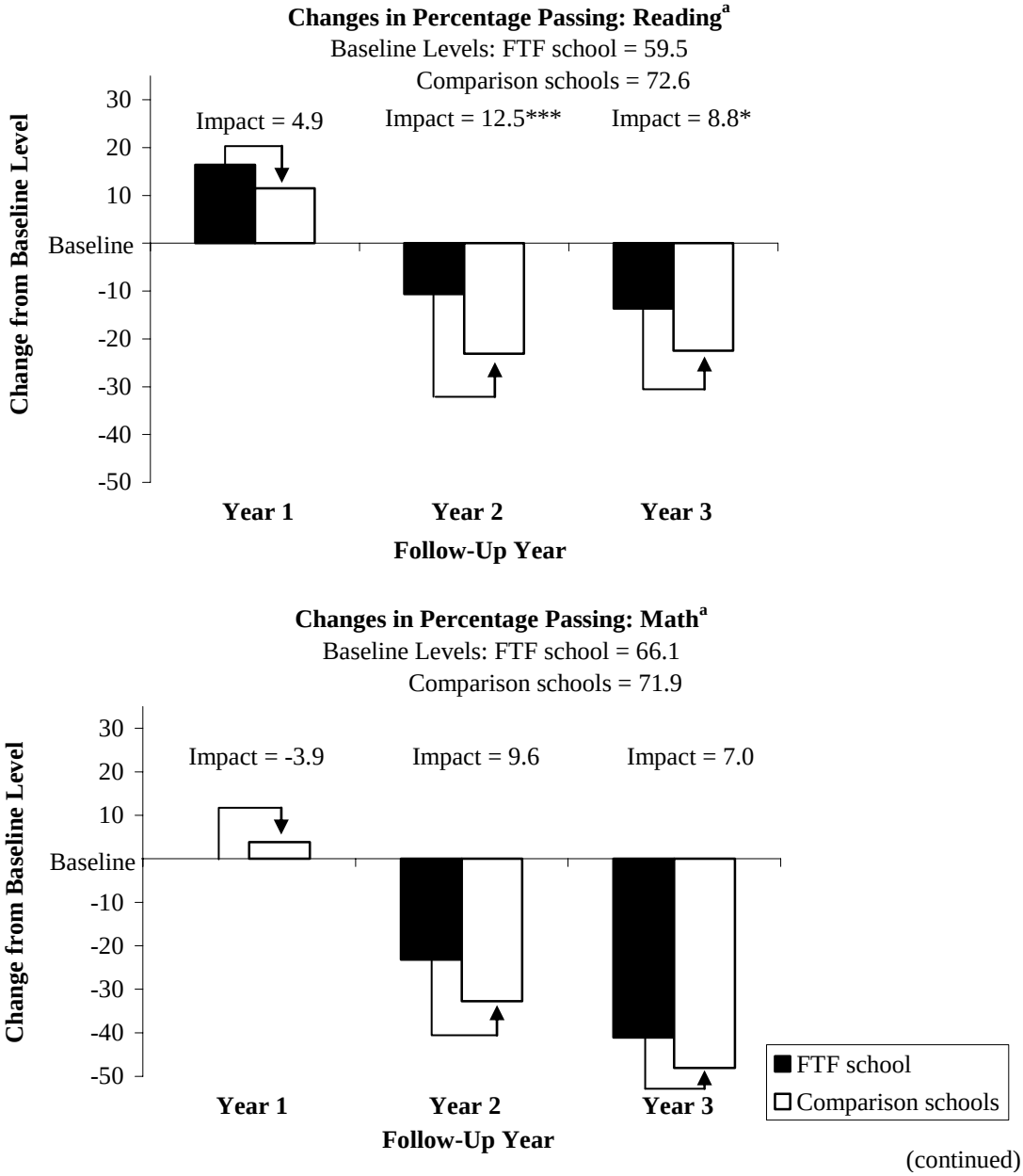
Findings for math in the bottom panel of Figure 4.8 suggest a similar conclusion. The average baseline passing rates for math at School E and its comparison schools were 66.1 percent and 71.9 percent, respectively. During the first follow-up year, there was no change in test scores at School E (hence, its bar in the chart for this year is not visible) and a slight rise in scores at the comparison schools. This resulted in a small relative decline for School E, of –3.9

[26]The minimum detectable effect for School E is a baseline-to-follow-up change of roughly 12 percentage points in reading and 18 percentage points in math. This means that impacts of this magnitude or larger are likely to be identified (with 80 percent statistical power), if they exist.

The First Things First Evaluation

Figure 4.8

Changes from Baseline Levels in the Percentage of 10th-Graders Passing the TAAS/TAKS in Reading and Math for the 2001 Cohort High School (School E): Houston, Texas



(continued)

Figure 4.8 (continued)

SOURCE: MDRC calculations from individual student records from the Houston Independent School District data file.

NOTES: Sample includes 10th-grade students from one First Things First (FTF) high school and five comparison schools. The sample consists of students for whom administrative records exist between the 1998-1999 and 2003-2004 academic years.

Each bar represents the "deviation from baseline," or the difference between the baseline level (average across three pre-implementation years) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the baseline for the FTF school and the deviation from the baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics and prior achievement.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in the percentage of students passing the state test.

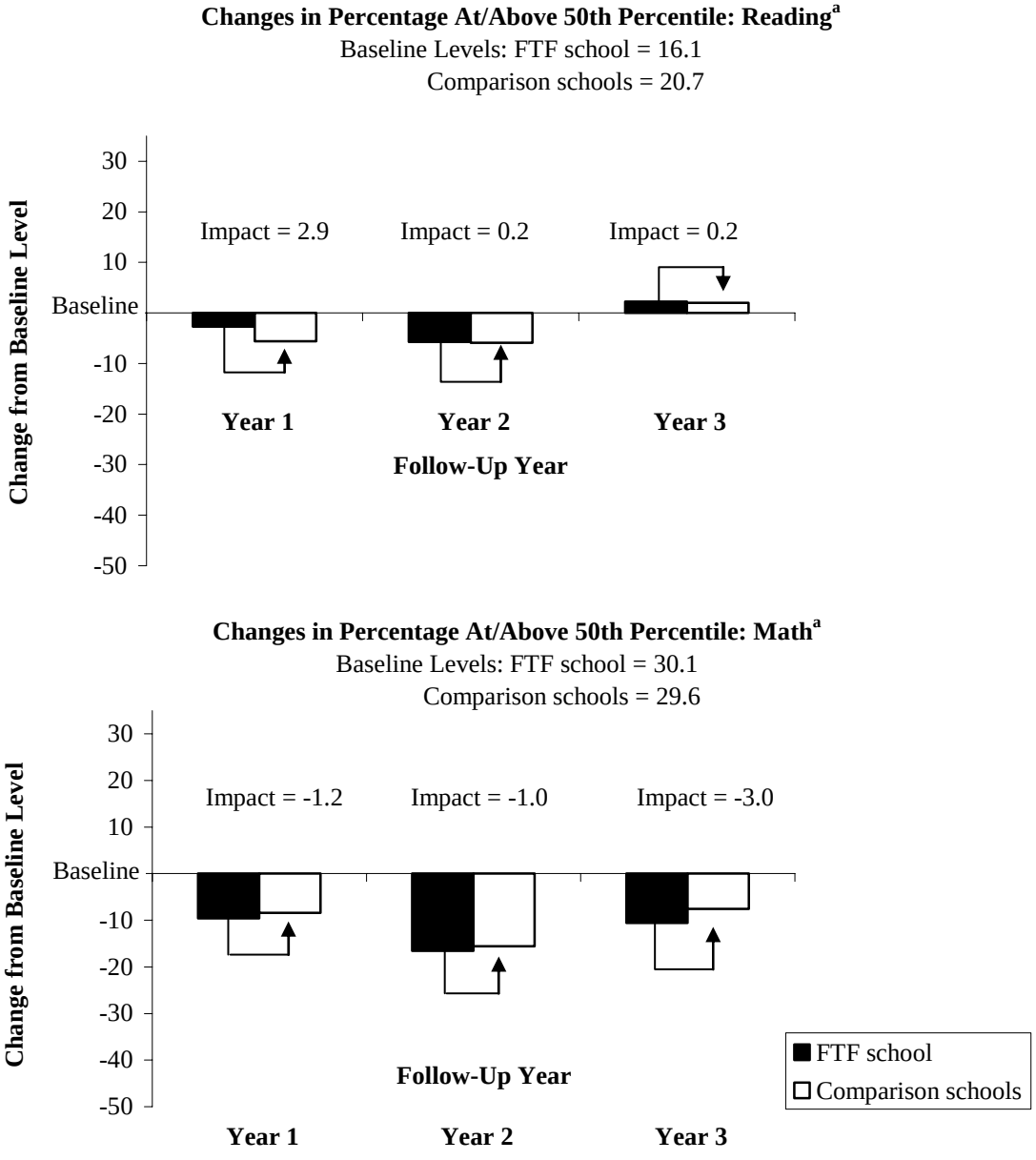
percentage points, which was not statistically significant. During the next two years, as the math test got harder and its standards rose, passing rates for both School E and its comparison schools fell, but those for School E fell by 9.6 and 7.0 percentage points less. Even though these differences were not statistically significant, it is nonetheless the case that School E closed its preexisting performance gap in math.

Figure 4.9 presents estimates of the impacts of FTF on SAT-9 scores for School E, with findings for reading in the top panel, and math in the bottom panel. During the baseline period, 16.1 percent of the students at School E and 20.7 percent of the students at its comparison schools scored at or above the 50th percentile nationwide in reading. After FTF was launched, this rate fluctuated for both sets of schools, with virtually no change in their relative positions. During the baseline period, 30.1 percent of the students at School E and 29.6 percent of the students at its comparison schools scored at or above the 50th percentile nationwide in math. After FTF was launched, these rates dropped for both groups of schools, with, once again, almost no change in their relative positions. Hence, there is no sign in the data that FTF increased student achievement on the SAT-9 at School E.

The First Things First Evaluation

Figure 4.9

Changes from Baseline Levels in the Percentage of 10th-Graders Scoring At/Above the 50th Percentile on the SAT-9 in Reading and Math for the 2001 Cohort High School (School E): Houston, Texas



(continued)

Figure 4.9 (continued)

SOURCE: MDRC calculations from individual student records from the Houston Independent School District data file.

NOTES: Sample includes 10th-grade students from one First Things First (FTF) high school and ten comparison schools. The sample consists of students for whom administrative records exist between the 1998-1999 and 2003-2004 academic years.

Each bar represents the "deviation from baseline," or the difference between the baseline level (average across three pre-implementation years) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the baseline for the FTF school and the deviation from the baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics and prior achievement.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in the percentage of students scoring at or above the 50th percentile on a nationally normed test.

Attendance and Persistence

Figure 4.10 presents estimates of the impacts of FTF on two other student outcomes for School E. The top panel of the figure presents findings for annual rates of student attendance for all grades. The bottom panel presents findings for annual rates of ninth-grade student persistence.

Baseline attendance rates at School E and its comparison schools were 87.8 percent and 89.9 percent, respectively. This implies that, on average, a day of class was missed every other week. After FTF was launched, these rates increased by varying amounts each year for School E and by somewhat less for its comparison schools. This resulted in relative improvements for School E of roughly 2 percentage points in follow-up Years 2 and 3, although these estimates are not statistically significant. The relatively small changes in employment rates that occurred are consistent with the limited margin for improvement that existed initially.

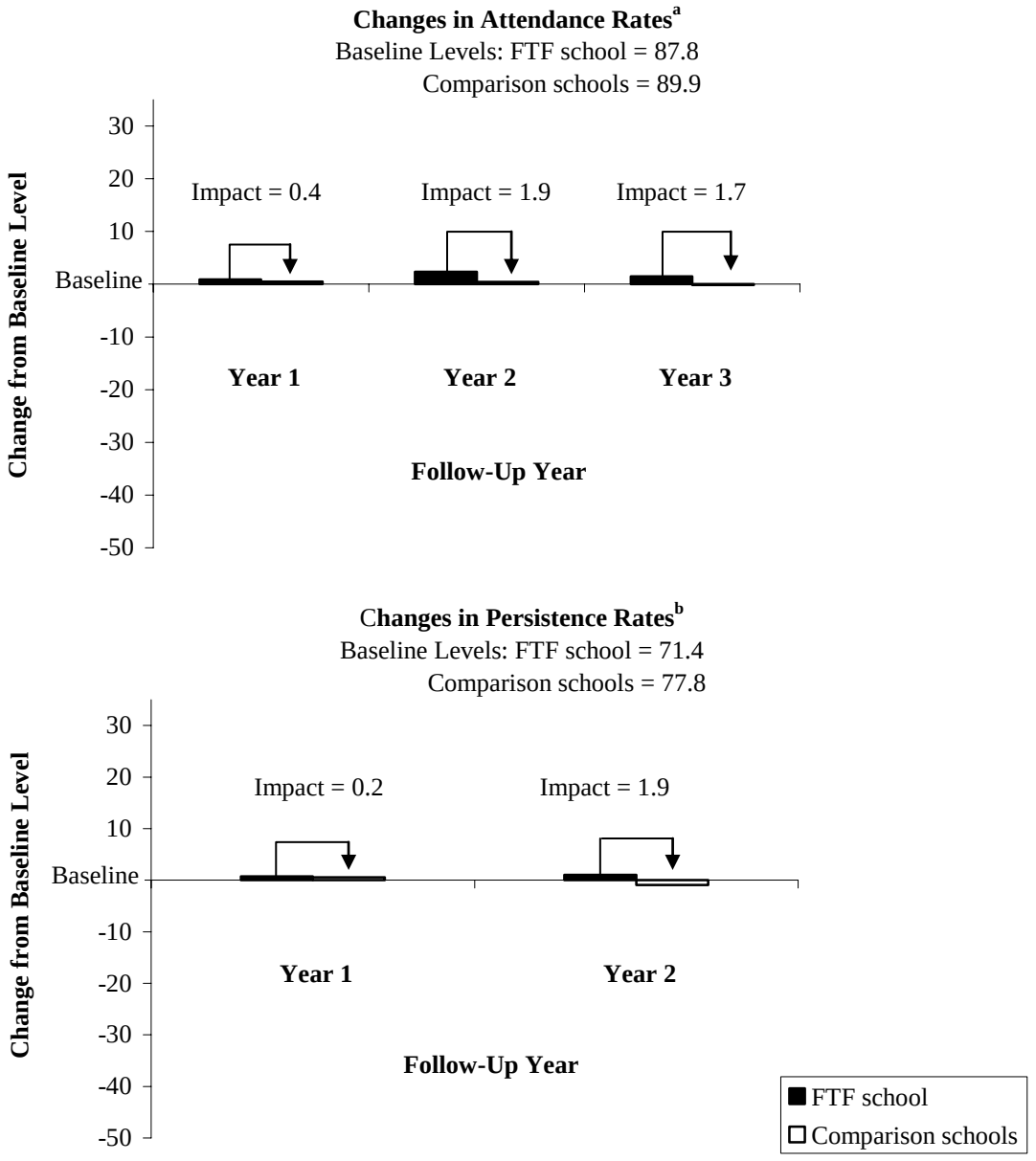
Baseline persistence rates were 71.4 percent and 77.8 percent at School E and its comparison schools, respectively. Thus, roughly three out of four entering ninth-graders remained in school after their first year of high school. After FTF began, there was very little change in this rate for either School E or its comparison schools. (Note that only two years of follow-up are available for this outcome, because it cannot be measured until the following school year ends for each ninth-grade cohort.)

The preceding findings suggest that although FTF may have helped School E produce large increases in state test scores, it did not produce demonstrable impacts on other student out-

The First Things First Evaluation

Figure 4.10

Changes from Baseline Levels in High School Attendance Rates and 9th-Grade Persistence Rates for the 2001 Cohort High School (School E): Houston, Texas



(continued)

Figure 4.10 (continued)

SOURCE: MDRC calculations from individual student records from the Houston Independent School District data file.

NOTES: Sample for attendance includes 9th- to 12th-grade students from one First Things First (FTF) high school and ten comparison schools; sample for persistence includes only 9th grade. The sample consists of students for whom administrative records exist between the 1998-1999 and 2003-2004 academic years.

Each bar represents the "deviation from baseline," or the difference between the baseline level (average across three pre-implementation years) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the baseline for the FTF school and the deviation from the baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in student attendance rates.

[b]The desired change in this measure is an increase from baseline, which represents an increase in the percentage of 9th-graders who are enrolled in school in the Houston school district the following year.

comes. The absence of impacts on the SAT-9 might be attributable to the substantial district-wide decline in emphasis placed on this test as the new state test became the overwhelming center of attention. The absence of impacts on attendance rates might be due to the fact that they were relatively high to begin with and, thus, had a limited margin for improvement. But there is no clear explanation for why the reform did not increase student persistence.

Nevertheless, the estimated impacts on the state test were quite large and are consistent with a wide range of qualitative and quantitative evidence indicating that, due to the vigorous efforts of an especially innovative and effective principal and his staff, FTF was implemented very well at this school. For example, according to reports by field researchers, School E is tied with one other school (the Houston 2001 cohort middle school, School S) for the highest overall rating of implementation success among FTF expansion sites. In addition, School E was the only school to register statistically significant increases in teachers' feelings of both support and engagement. Furthermore, detailed accounts by IRRE staff provide many specific examples of actions that were taken at this school that suggest a very high level of implementation quality. It therefore seems plausible that educational changes that were seen "on the ground" translated into improved performance by students on the single most important indicator of their success: the state's high-stakes test.

The 2002 Cohort High Schools

For the 2002 cohort high schools in Houston, Figures 4.11, 4.12, and 4.13 present estimates of the impacts of FTF on state test scores, SAT-9 scores, student attendance rates, and student persistence. These figures provide no evidence that, to date, the reform has improved student outcomes at these two schools.

Achievement

Figure 4.11 presents impact findings for the Texas state test. It shows that the FTF schools and comparison schools were well matched at baseline, especially for reading. During the first follow-up year, the passing rates for both groups of schools dropped precipitously with the onset of the new, more difficult state test. These rates dropped even further in math during the second follow-up year as the threshold for passing was raised yet again. There was very little difference and no consistent pattern, however, in the relative changes for the FTF schools and comparison schools.

Figure 4.12 (found on page 105) presents corresponding findings for student performance on the SAT-9. Once again, there was a close match on the baseline scores. But subsequent changes in performance were not large for this test, and there was no consistent pattern in the relative changes of the FTF schools and comparison schools.

Attendance and Persistence

The top panel of Figure 4.13 (found on page 107) illustrates that annual attendance rates were well matched at baseline for the FTF schools and comparison schools and changed very little for either group of schools during the first two years of FTF implementation. The bottom panel of the figure illustrates that ninth-grade persistence rates were also relatively well matched at baseline and that they also changed very little after the beginning of FTF. As noted earlier, findings for persistence are reported only for one follow-up year because of the lag in time required to measure this outcome.

Middle School Results

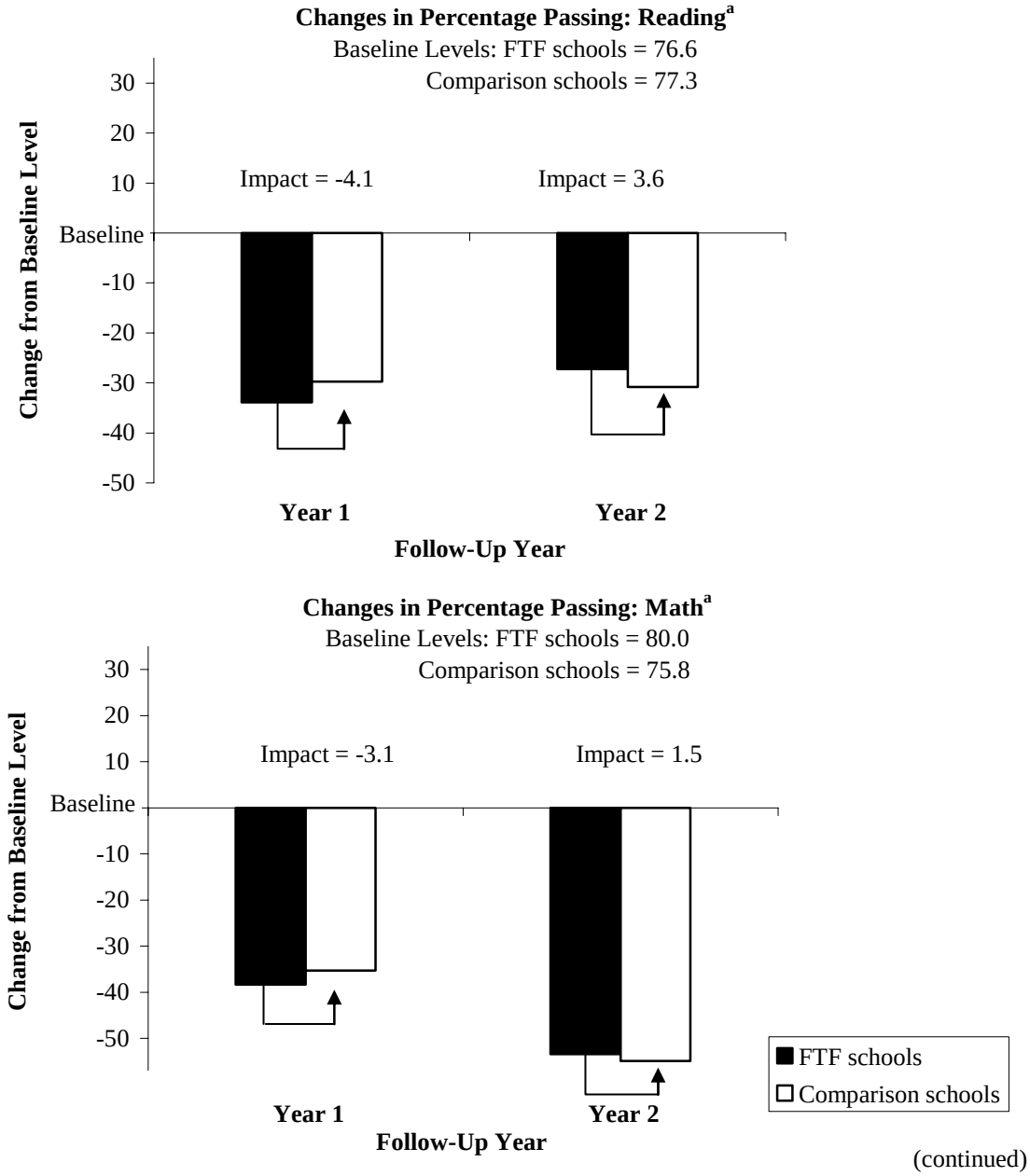
Table 4.4 (found on page 109) summarizes test scores for the 2001 cohort middle school (School S) and its comparison schools, the three 2002 cohort middle schools and their comparison schools, the Houston Independent School District, and the State of Texas (for the state test only). The top panel of the table summarizes findings for the state test in reading and math, which is administered to eighth-graders each year. The bottom panel summarizes findings for the SAT-9, which is also administered to eighth-graders each year.

Findings in the top panel of Table 4.4 indicate that passing rates for the state test were rising for all groups during the baseline period but declined sharply in 2003, when the new state

The First Things First Evaluation

Figure 4.11

Changes from Baseline Levels in the Percentage of 10th-Graders Passing the TAAS/TAKS in Reading and Math for the 2002 Cohort High Schools (Schools F and G): Houston, Texas



(continued)

Figure 4.11 (continued)

SOURCE: MDRC calculations from individual student records from the Houston Independent School District data file.

NOTES: Sample includes 10th-grade students from two clusters. Each cluster consists of a First Things First (FTF) high school matched with a group of between six and seven non-FTF schools. The sample consists of students for whom administrative records exist between the 1999-2000 and 2003-2004 academic years.

Each bar represents the "deviation from baseline," or the difference between the baseline level (average across three pre-implementation years) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the baseline for FTF schools and the deviation from the baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics and prior achievement.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in the percentage of students passing the state test.

test was implemented. Passing rates for math fell again in 2004, when the threshold for passing was raised again. Throughout this period, there is no sign that student performance in the FTF middle schools improved relative to student performance in the comparison schools.

Findings in the bottom panel indicate no systematic change over time in student performance for either cohort of FTF schools, their comparison schools, or the district as a whole. The rest of this section examines findings from a statistical analysis of test-score data plus information on student attendance.

The 2001 Cohort Middle School: School S

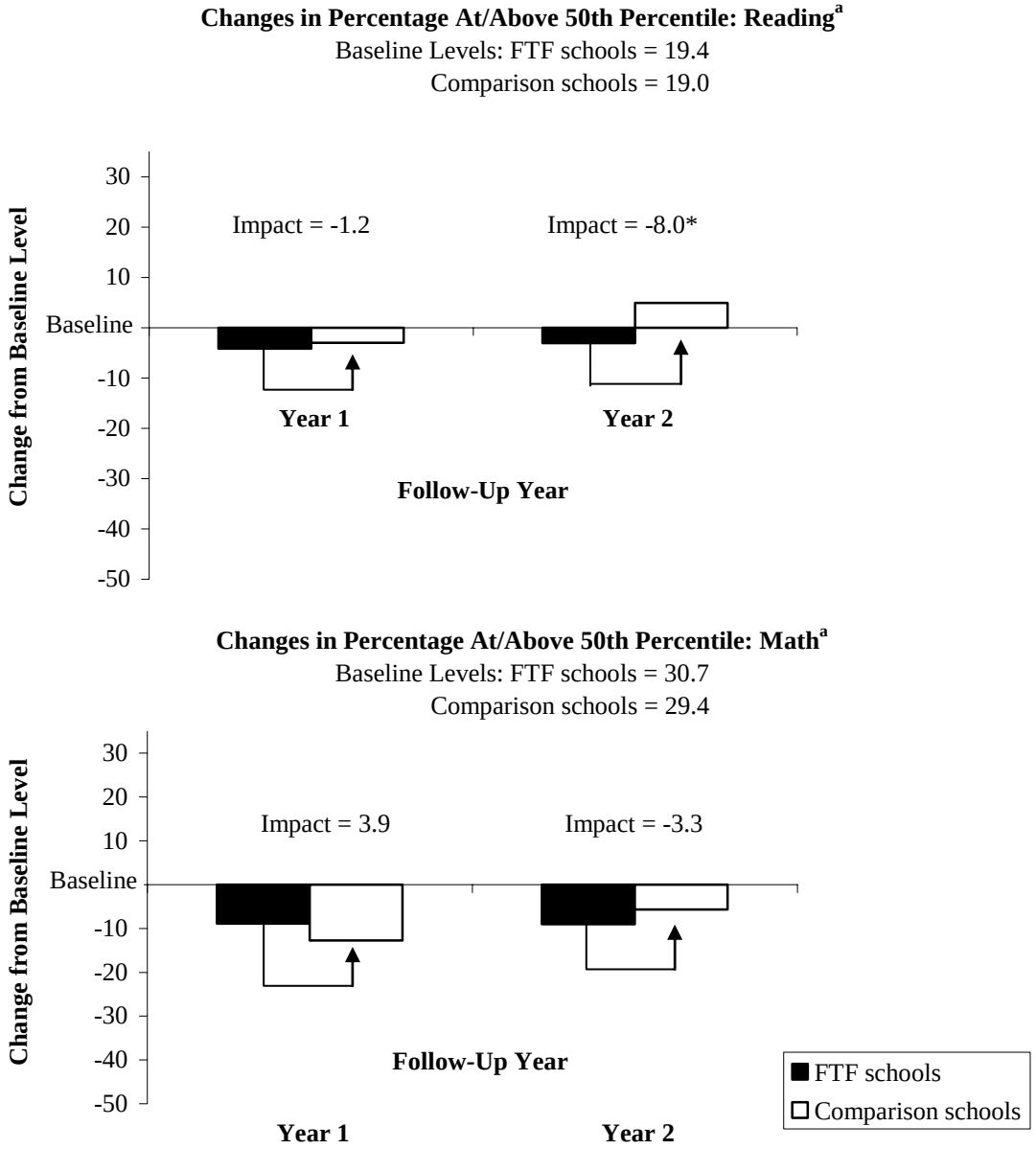
Achievement

Figure 4.14 (found on page 112) presents estimates for the 2001 cohort middle school, School S, of the impacts of FTF on passing rates for the state tests in reading and math. Baseline passing rates for School S and its comparison schools are quite similar. These rates changed very little in reading after FTF (and the new state test) was launched. And although passing rates dropped precipitously in math with the new state test, they did so very similarly for School S and its comparison schools. Hence, School S experienced no relative improvement in either subject.

The First Things First Evaluation

Figure 4.12

Changes from Baseline Levels in the Percentage of 10th-Graders Scoring At/Above the 50th Percentile on the SAT-9 in Reading and Math for the 2002 Cohort High Schools (Schools F and G): Houston, Texas



(continued)

Figure 4.12 (continued)

SOURCE: MDRC calculations from individual student records from the Houston Independent School District data file.

NOTES: Sample includes 10th-grade students from two clusters. Each cluster consists of a First Things First (FTF) high school matched with a group of eleven non-FTF schools. The sample consists of students for whom administrative records exist between the 1999-2000 and 2003-2004 academic years.

Each bar represents the "deviation from baseline," or the difference between the baseline level (average across three pre-implementation years) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the baseline for FTF schools and the deviation from the baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics and prior achievement.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in the percentage of students scoring at or above the 50th percentile on a nationally normed test.

Figure 4.15 (found on page 114) presents corresponding impact estimates with respect to student performance on the SAT-9. These findings also provide no indication that FTF caused student achievement to improve at School S. Impact estimates for reading were negative (suggesting a reduction in relative performance) but were very small and not at all statistically significant. Impact estimates for math were larger but bounced erratically from negative to positive over time.

Attendance

Lastly, Figure 4.16 (found on page 116) presents estimates of impacts on student attendance rates. These findings indicate that baseline attendance rates for both School S and its comparison schools were very high (96.2 percent and 93.5 percent, respectively) and, thus, had virtually no room for improvement. Indeed, the baseline rates for School S were so high that subsequently there was almost no place to go but down.

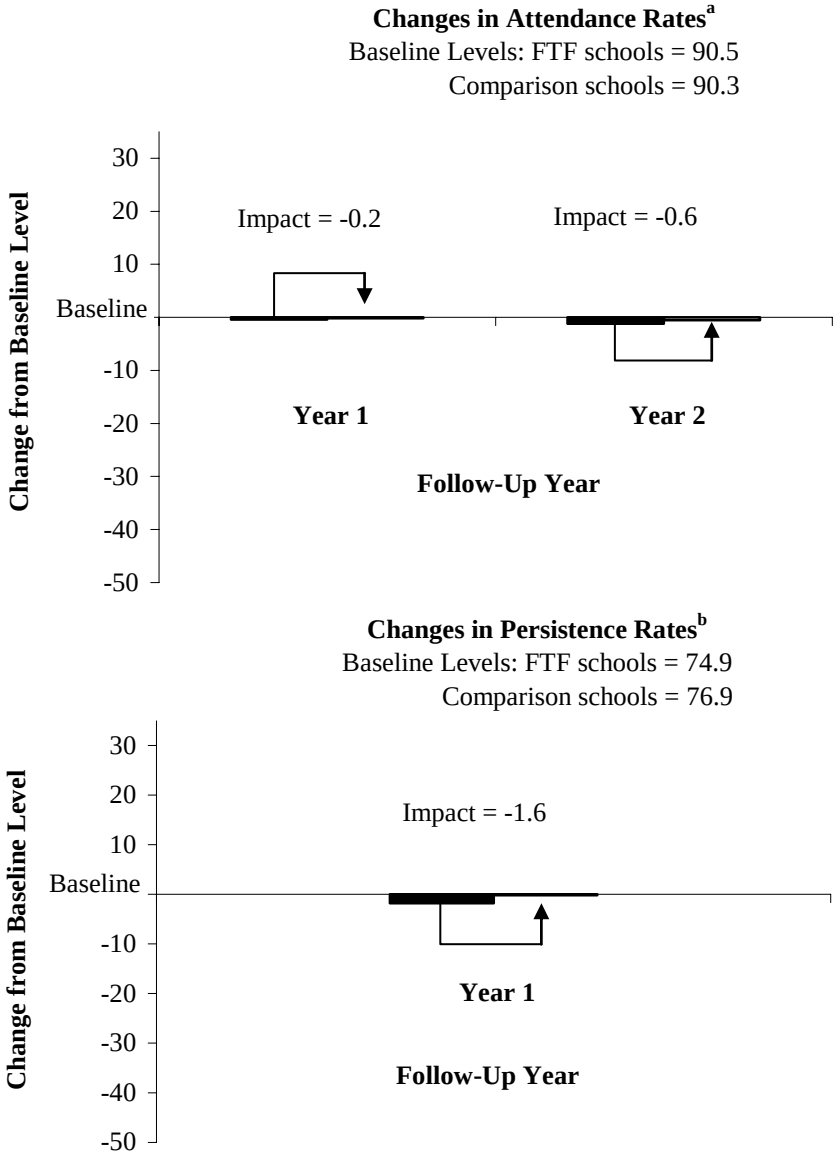
The 2002 Cohort Middle Schools

Figures 4.17, 4.18, and 4.19 (found on pages 117-121) present estimates of the impacts of FTF on student outcomes at the three 2002 cohort middle schools in Houston. These findings provide no systematic evidence that the reform initiative has improved student outcomes to date.

The First Things First Evaluation

Figure 4.13

Changes from Baseline Levels in High School Attendance Rates and 9th-Grade Persistence Rates for the 2002 Cohort High Schools (Schools F and G): Houston, Texas



(continued)

Figure 4.13 (continued)

SOURCE: MDRC calculations from individual student records from the Houston Independent School District data file.

NOTES: Sample for attendance includes 9th- to 12th-grade students from two clusters; sample for persistence includes only 9th grade. Each cluster consists of a First Things First (FTF) high school matched with a group of eleven non-FTF schools. The sample consists of students for whom administrative records exist between the 1999-2000 and 2003-2004 academic years.

Each bar represents the "deviation from baseline," or the difference between the baseline level (average across three pre-implementation years) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the baseline for FTF schools and the deviation from the baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics and prior achievement.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in student attendance rates.

[b]The desired change in this measure is an increase from baseline, which represents an increase in the percentage of 9th-graders who are enrolled in school in the Houston school district the following year.

Achievement

Figure 4.17 (found on page 117) illustrates that state passing rates for the FTF schools and comparison schools were very similar during the baseline period. These rates dropped for both schools, precipitously in math and less so in reading, with the onset of the new state test. In reading, it appears that the FTF schools fell slightly further behind their comparison school counterparts (a difference that was statistically significant in one of the two years). But in math, there was no consistent change in their relative performance. Figure 4.18 (found on page 119) illustrates that FTF schools and comparison schools were well matched at baseline in terms of their SAT-9 scores in both reading and math. Subsequently, there was an erratic pattern of changes in test scores for the two groups of schools — most of which represented declining performance, and none of which represented a relative improvement for the FTF schools.

The First Things First Evaluation

Table 4.4

Middle School Test Results for First Things First Schools, Comparison Schools, the District, and the State: Houston, Texas

	Percentage Passing TAAS/TAKS State Test					
	Spring 1999	Spring 2000	Spring 2001	Spring 2002	Spring 2003	Spring 2004
8th-grade reading						
				3 Follow-Up Years		
2001 cohort FTF school	70.8	78.9	91.1	93.7	84.3	84.5
Comparison schools	75.3	81.0	86.6	90.9	84.4	86.3
					2 Follow-Up Years	
2002 cohort FTF schools		82.8	87.2	90.7	82.2	80.0
Comparison schools		83.6	88.4	91.8	86.0	87.7
District	78.2	83.6	87.9	92.3	84.9	86.5
State	88.0	89.0	91.0	94.0	88.0	89.0
8th-grade math						
				3 Follow-Up Years		
2001 cohort FTF school	71.1	84.6	88.5	90.8	63.6	58.8
Comparison schools	72.3	80.9	87.3	90.3	57.5	49.6
					2 Follow-Up Years	
2002 cohort FTF schools		79.7	86.6	87.5	60.3	48.9
Comparison schools		82.7	87.2	90.5	58.8	50.1
District	74.2	82.6	87.4	90.5	60.6	54.2
State	85.0	90.0	92.0	92.0	72.0	66.0

	Percentage At/Above 50th Percentile on Nationally Normed Test					
	Spring 1999	Spring 2000	Spring 2001	Spring 2002	Spring 2003	Spring 2004
8th-grade reading						
				3 Follow-Up Years		
2001 cohort FTF school	22.7	24.0	34.6	27.6	34.1	35.3
Comparison schools	21.7	22.7	29.1	23.6	26.0	26.9
					2 Follow-Up Years	
2002 cohort FTF schools		32.9	36.6	27.8	31.8	26.5
Comparison schools		27.6	33.1	28.9	31.7	31.4
District	31.0	31.0	36.2	33.2	34.9	35.0
8th-grade math						
				3 Follow-Up Years		
2001 cohort FTF school	26.6	28.5	35.8	28.1	21.5	48.2
Comparison schools	22.6	23.4	27.5	22.4	23.6	32.4
					2 Follow-Up Years	
2002 cohort FTF schools		31.7	36.8	25.6	30.2	39.7
Comparison schools		26.6	29.8	27.5	26.6	32.5
District	30.1	31.7	33.7	30.3	30.6	40.1

(continued)

Table 4.4 (continued)

SOURCES: MDRC calculations from individual student records from the Houston Independent School District data file for FTF schools, comparison schools, and district results; TAAS and TAKS performance results reported on the Texas Education Agency website (www.tea.state.tx.us) for state results.

NOTE: FTF was implemented in one middle school in the fall of 2001 (2001 cohort) and in three additional middle schools in the fall of 2002 (2002 cohort). Therefore, spring 2002 is a follow-up year for one school and a baseline year for the other three schools.

Boxed areas represent follow-up years.

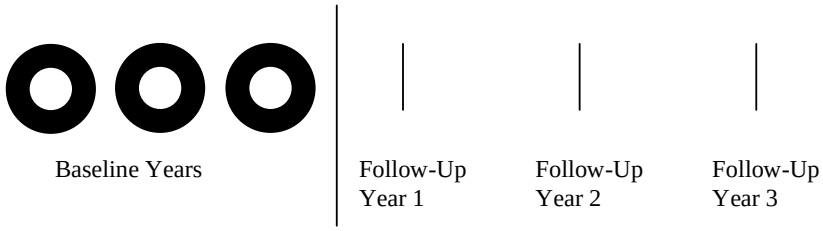
Attendance

Figure 4.19 (found on page 121) illustrates that attendance rates for the FTF schools and comparison schools were quite high and very closely matched during the baseline period (93.1 percent and 93.6 percent, respectively). Subsequent changes for both groups were small, and their differences were imperceptible.

FTF in Riverview Gardens, Missouri

As noted earlier, FTF was implemented at all secondary schools in the Riverview Gardens school district. Thus, comparison schools for the one high school in the district and a single "composite" sample from its two middle schools were selected from other school districts throughout Missouri. The evaluation design for this site (illustrated in Figure 4.1 and repeated below) uses school-level data on student outcomes for three pre-intervention baseline years and three post-intervention follow-up years. Impacts are measured as differences between baseline-to-follow-up changes in outcomes for FTF schools and comparison schools. Because school-level data had to be used for the analysis, it was not possible to adjust for changes over time in students' background characteristics. Fortunately, these changes were small.

The following discussion examines the impacts of FTF on student performance in communication arts and math on the state's high-stakes test, known as the Missouri Assessment Program (MAP). Impacts on rates of student attendance, dropout, and graduation are also examined. As is done for the Kansas test, scores on the Missouri test are reported by the state in five performance categories (from lowest to highest): step 1, progressing, nearing proficient, proficient, and advanced. The present analysis measures the impacts of FTF on the percentage of students



who scored in the bottom two of these categories, which represents the overwhelming majority of students from the site. Complementary findings for the highest two categories are reported in Appendix D; few students from Riverview Gardens scored in these categories.[27]

Results of the analysis suggest that:

- For both high schools and middle schools, FTF may have improved student performance in math, but this finding is not statistically significant. No evidence was found of impacts on other high school outcomes.

High School Results

Achievement

The MAP is administered each year in communications arts to eleventh-graders and in math to tenth-graders. Table 4.5 (found on page 122) summarizes these findings for the FTF high school, its comparison schools, and the State of Missouri. For both subjects, there is a pretty good baseline match between the FTF school and its comparison schools. In addition, proportionally many more students from these schools scored in the bottom two performance categories than did students from across the state. This illustrates, once again, the fact that FTF was explicitly targeted to schools that were experiencing difficulties.

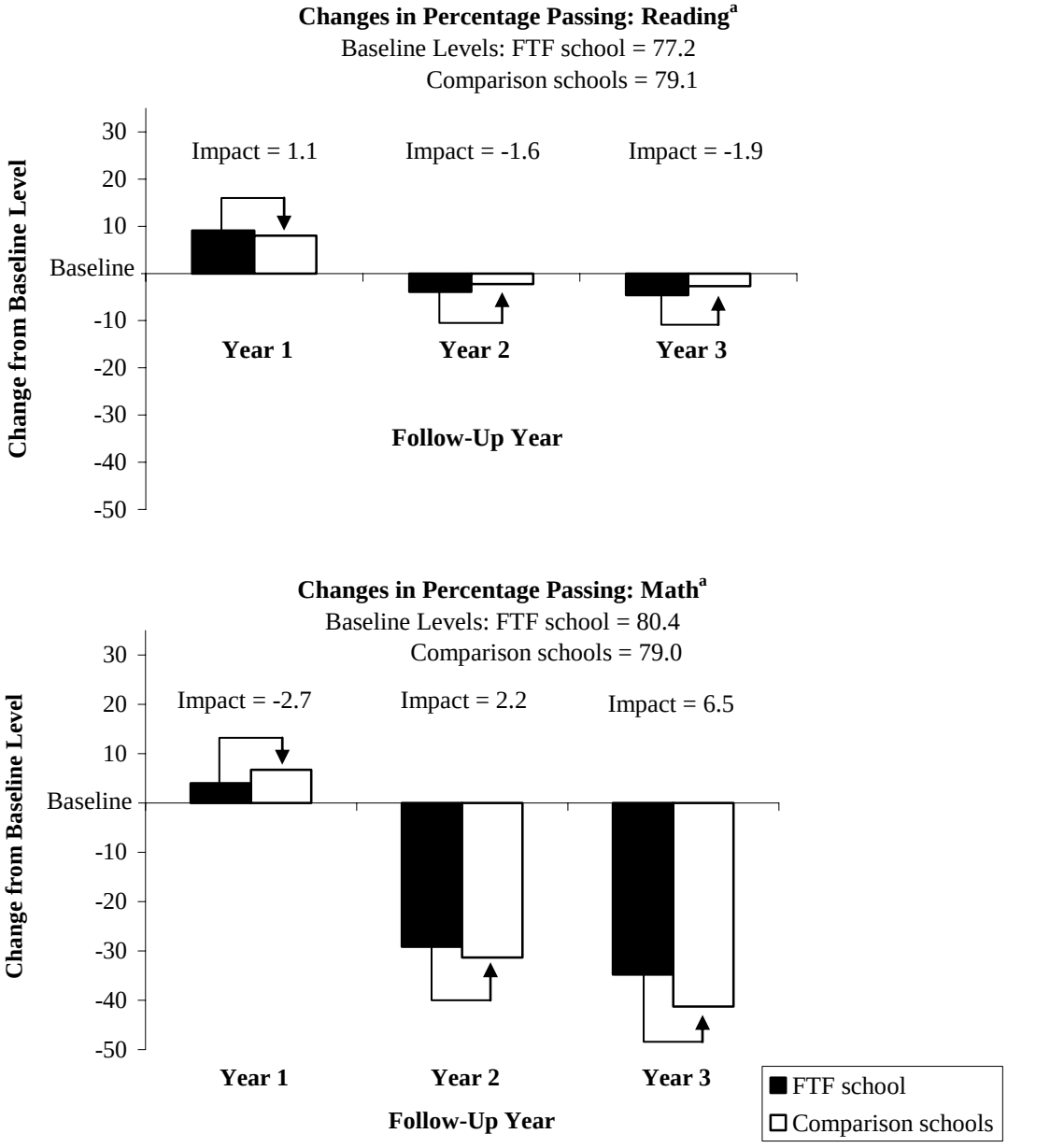
During the follow-up period, there was no systematic change in the relative performance of the FTF school in communication arts. However, the school did appear to experience a relative improvement in math.

[27]During the baseline period, only 5.6 percent and 0.4 percent of the high school students from Riverview Gardens and 11.2 percent and 1.1 percent of the middle school students scored in the upper two performance categories for communications arts and math, respectively.

The First Things First Evaluation

Figure 4.14

Changes from Baseline Levels in the Percentage of 8th-Graders Passing the TAAS/TAKS in Reading and Math for the 2001 Cohort Middle School (School S): Houston, Texas



(continued)

Figure 4.14 (continued)

SOURCE: MDRC calculations from individual student records from the Houston Independent School District data file.

NOTES: Sample includes 8th-grade students from one First Things First (FTF) middle school and fourteen comparison schools. The sample consists of students for whom administrative records exist between the 1998-1999 and 2003-2004 academic years.

Each bar represents the "deviation from baseline," or the difference between the baseline level (average across three pre-implementation years) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the baseline for the FTF school and the deviation from the baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics and prior achievement.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in the percentage of students passing the state test.

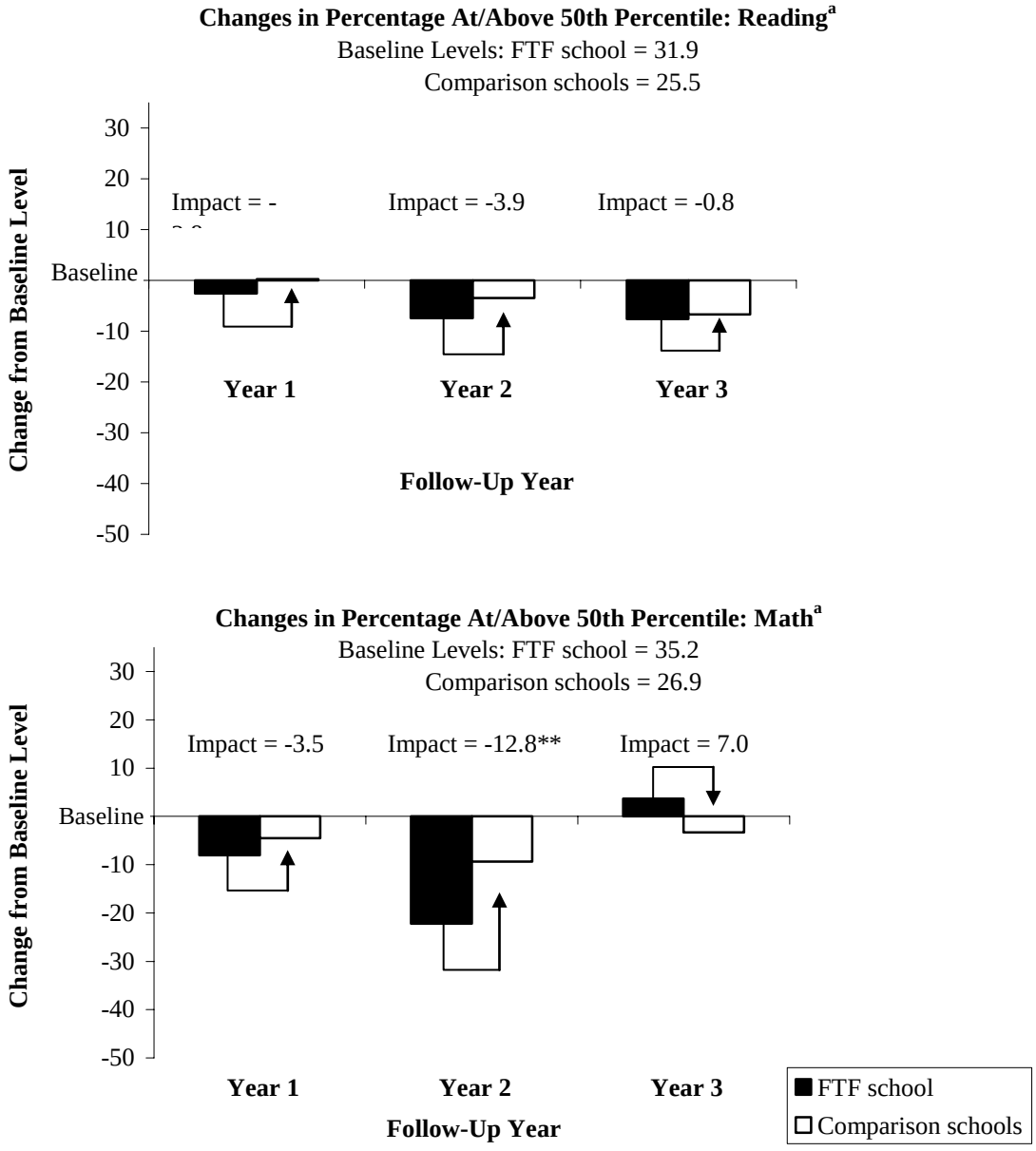
Figure 4.20 (found on page 123) presents estimates of impacts on state test scores. As noted above, these estimates do not adjust for the background characteristics of individual students. The findings for communications arts in the top panel of the figure indicate that the FTF school and comparison schools were well matched at baseline and that there was no consistent change in their relative performance during the follow-up period. Scores for the FTF school fluctuated erratically above and below its baseline mean, while those for the comparison schools hardly changed.

In contrast, the results for math provide some evidence that FTF might have improved student performance. Each follow-up year, the percentage of low-performing students from the FTF school declined by an increasing amount, while the corresponding percentage for comparison schools barely changed. Thus, the FTF school improved its relative performance by an amount that increased over time from –2.8 percentage points to –10.0 percentage points. This change measures the relative reduction in the percentage of students scoring in the state's bottom two performance categories. Unfortunately, because of the relatively low statistical precision for this analysis (reflecting that it was based on only one FTF school and that there was considerable year-to-year fluctuation in school-level test results), the impact estimate was not

The First Things First Evaluation

Figure 4.15

Changes from Baseline Levels in the Percentage of 8th-Graders Scoring At/Above the 50th Percentile on the SAT-9 in Reading and Math for the 2001 Cohort Middle School (School S): Houston, Texas



(continued)

Figure 4.15 (continued)

SOURCE: MDRC calculations from individual student records from the Houston Independent School District data file.

NOTES: Sample includes 8th-grade students from one First Things First (FTF) middle school and fourteen comparison schools. The sample consists of students for whom administrative records exist between the 1998-1999 and 2003-2004 academic years.

Each bar represents the "deviation from baseline," or the difference between the baseline level (average across three pre-implementation years) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the baseline for the FTF school and the deviation from the baseline for comparison schools.

Estimates are regression -adjusted for students' background characteristics and prior achievement.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in the percentage of students scoring at or above the 50th percentile on a nationally normed test.

statistically significant. Thus, it cannot distinguish with confidence a true impact from random estimation and sampling error.[28]

Attendance, Dropout, and Graduation

Figure 4.21 (found on page 125) presents estimates of the impacts of FTF on three more student outcomes, based on school-level data reported by districts to the state each year. The top panel of the figure reports findings for high school attendance.[29] Baseline attendance rates were somewhat higher for the FTF school (90.4 percent) than for the comparison schools (84.5 percent). And subsequent changes after the launch of FTF were minimal for both groups. Hence, there was no consistent change in their relative performance on this outcome.

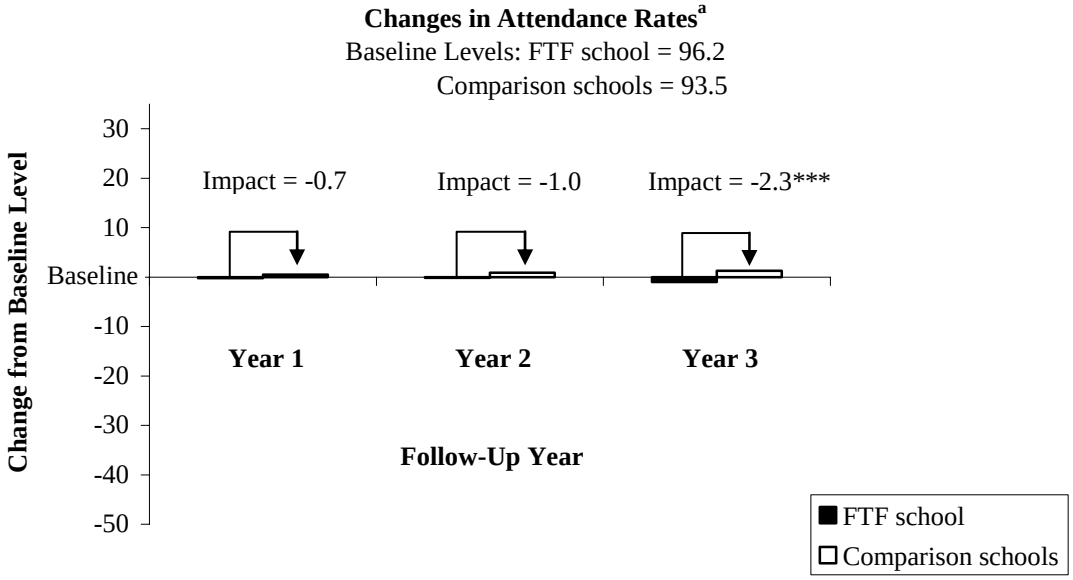
[28]The minimum detectable effect for this estimate is roughly 17 percentage points, which means that the evaluation design only has a good chance of identifying true impacts that are at least this large.

[29]In Missouri, the attendance rate for all grades in a school combined is computed as the average daily number of students attending throughout the academic year divided by the total January enrollment, multiplied by 100. This equals the total number of hours of student attendance divided by the sum of the total number hours of student attendance and the total number of hours of student absences.

The First Things First Evaluation

Figure 4.16

Changes from Baseline Levels in Middle School Attendance Rates for the 2001 Cohort Middle School (School S): Houston, Texas



SOURCE: MDRC calculations from individual student records from the Houston Independent School District data file.

NOTES: Sample includes 8th-grade students from one First Things First (FTF) middle school and fourteen comparison schools. The sample consists of students for whom administrative records exist between the 1998-1999 and 2003-2004 academic years.

Each bar represents the "deviation from baseline," or the difference between the baseline level (average across three pre-implementation years) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the baseline for the FTF school and the deviation from the baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics and prior achievement.

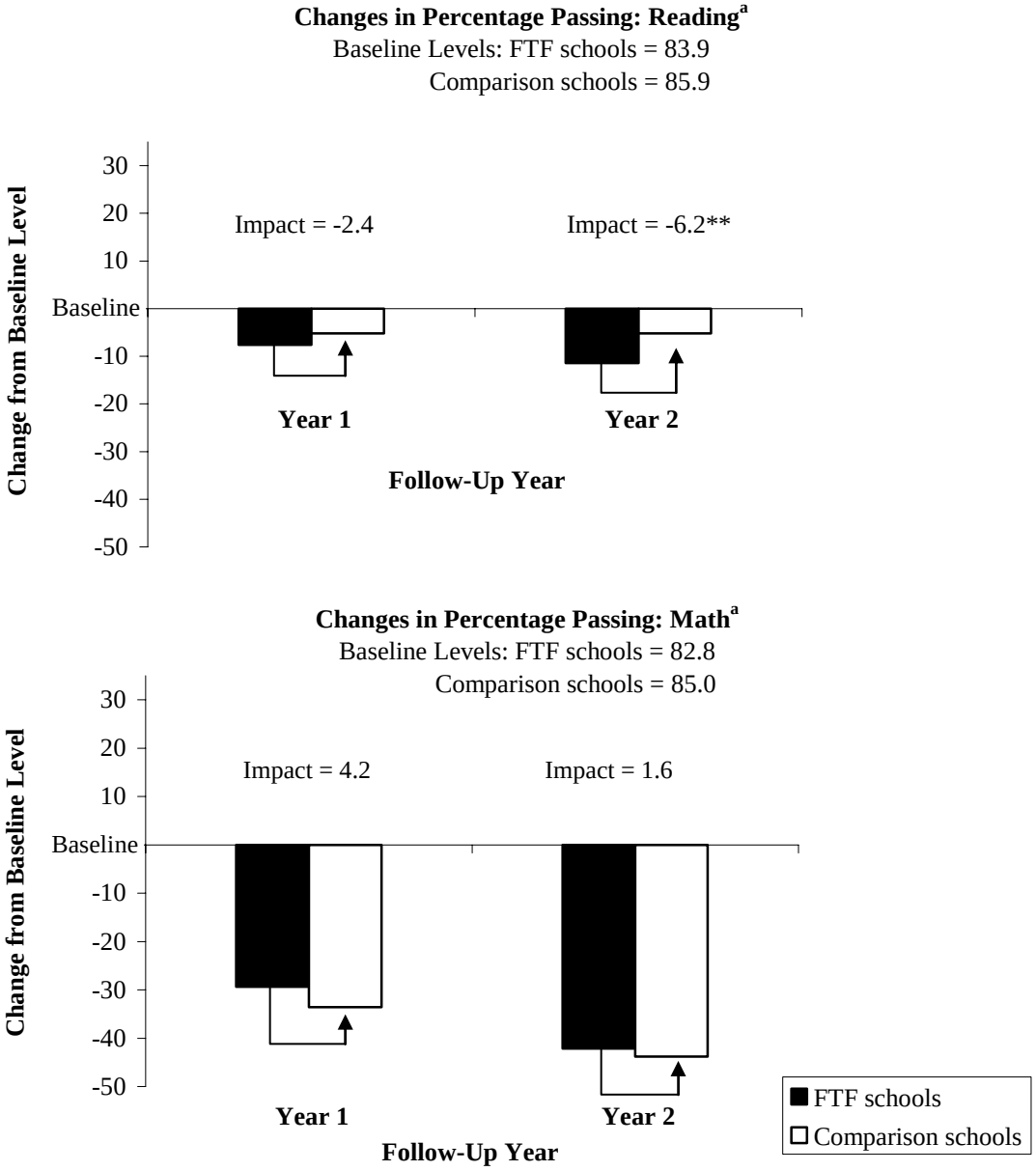
A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in student attendance rates.

The First Things First Evaluation

Figure 4.17

Changes from Baseline Levels in the Percentage of 8th-Graders Passing the TAAS/TAKS in Reading and Math for the 2002 Cohort Middle Schools (Schools U, V, and T): Houston, Texas



(continued)

Figure 4.17 (continued)

SOURCE: MDRC calculations from individual student records from the Houston Independent School District data file.

NOTES: Sample includes 8th-grade students from three clusters. Each cluster consists of a First Things First (FTF) middle school matched with a group of between three and fifteen non-FTF schools. The sample consists of students for whom administrative records exist between the 1999-2000 and 2003-2004 academic years.

Each bar represents the "deviation from baseline," or the difference between the baseline level (average across three pre-implementation years) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the baseline for FTF schools and the deviation from the baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics and prior achievement.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in the percentage of students passing the state test.

The second panel of the figure is for dropout rates.[30] Given the way Missouri calculates dropout rates, the baseline dropout rate for the FTF school is 5.3 percent and for the comparison schools is 7.1 percent. Subsequent rates for the FTF school and its comparison schools suggest small and inconsistent fluctuations that produced no lasting change in their relative performance.

The bottom panel of Figure 4.21 presents graduation rates. In Missouri, the graduation rate for a school is based on the ninth-grade class that entered four years previously.[31] The baseline graduation rate for the FTF school (73.0 percent) was higher than that for comparison schools (62.5 percent), but there was no consistent pattern in their subsequent changes.

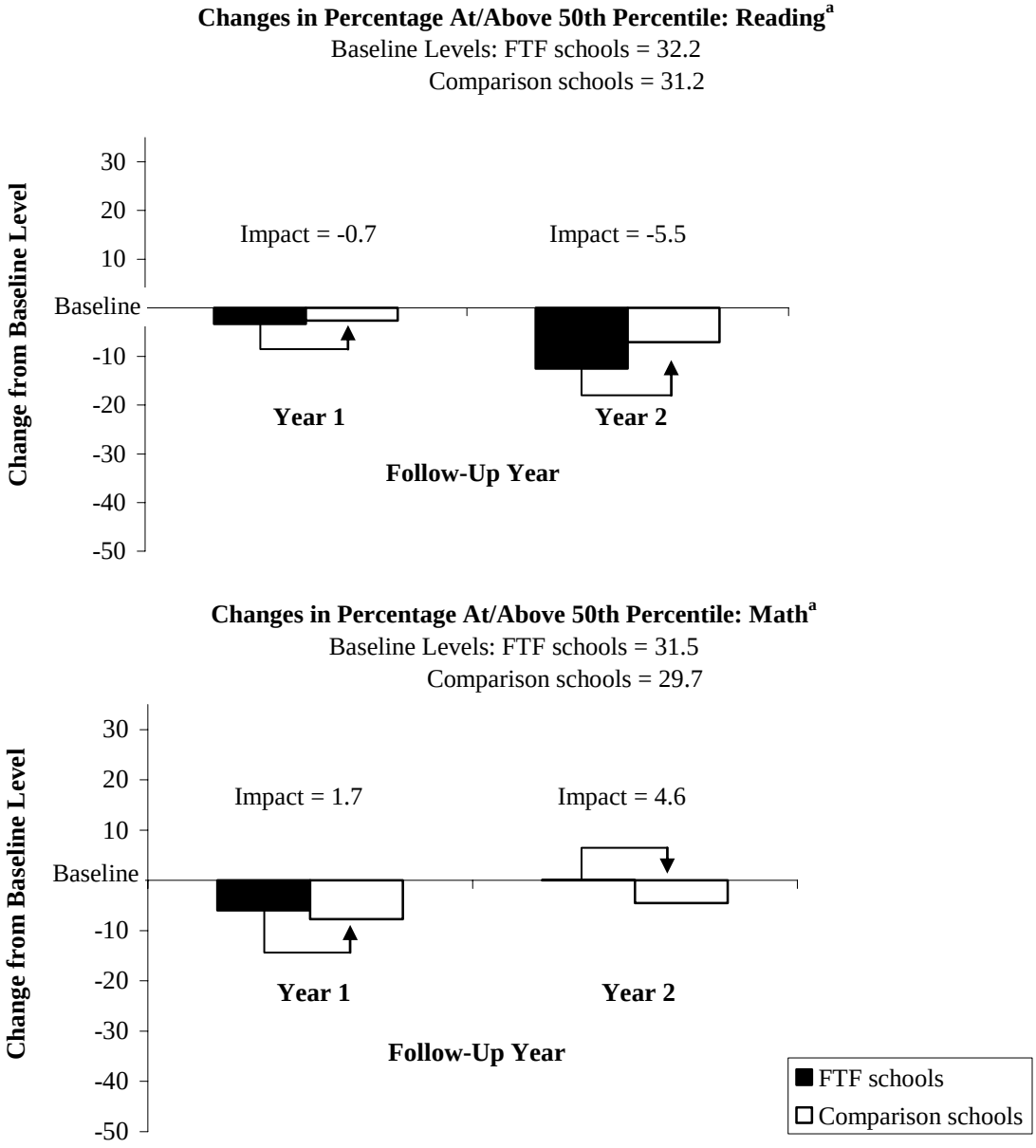
[30]The dropout rate for all grades in a school is computed each year as the total number of dropouts recorded by the school divided by its average enrollment. Average enrollment is computed as the mean of: (1) the total fall enrollment and (2) the fall enrollment plus transfers in during the year minus transfers out and minus dropouts.

[31]Specifically, it is computed as the total number of graduates in the current year (as of June 30) divided by the sum of the number of graduates plus the number of twelfth-graders who dropped out in the current year plus the number of eleventh-graders who dropped out in the preceding year plus the number of tenth-graders who dropped out in the second preceding year plus the number of ninth-graders who dropped out in the third preceding year.

The First Things First Evaluation

Figure 4.18

Changes from Baseline Levels in the Percentage of 8th-Graders Scoring At/Above the 50th Percentile on the SAT-9 in Reading and Math for the 2002 Cohort Middle Schools (Schools U, V, and T): Houston, Texas



(continued)

Figure 4.18 (continued)

SOURCE: MDRC calculations from individual student records from the Houston Independent School District data file.

NOTES: Sample includes 8th-grade students from three clusters. Each cluster consists of a First Things First (FTF) middle school matched with a group of between three and fifteen non-FTF schools. The sample consists of students for whom administrative records exist between the 1999-2000 and 2003-2004 academic years.

Each bar represents the "deviation from baseline," or the difference between the baseline level (average across three pre-implementation years) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the baseline for FTF schools and the deviation from the baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics and prior achievement.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in the percentage of students scoring at or above the 50th percentile on a nationally normed test.

Middle School Results

Achievement

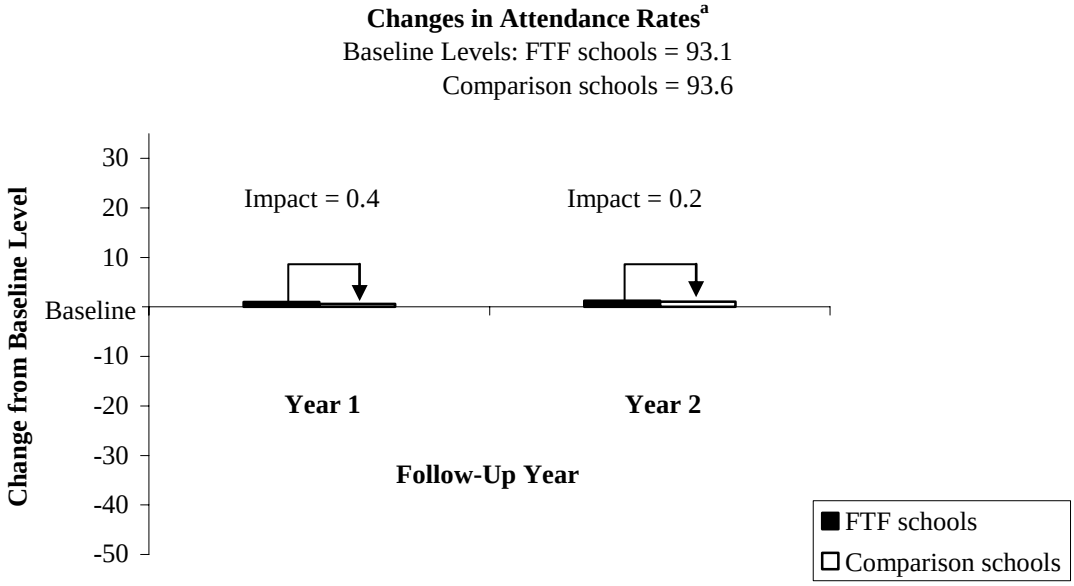
The MAP is administered each year in communications arts to seventh-graders and in math to eighth-graders. Its scores are reported by the State of Missouri in the five performance categories that are used for high schools. Table 4.6 (found on page 124) summarizes these findings for the composite FTF middle school, its comparison schools, and the state as a whole in terms of the percentage of students scoring in the bottom two state categories.

These findings indicate that, during the baseline period, students at the FTF school scored less well in communication arts than did their comparison school counterparts, although both groups of schools performed much less well than the state as a whole. During the follow-up period, there was no consistent change in the relative performance of the FTF school. For math, however, the story is somewhat different. There was a very close baseline match between the FTF school and its comparison schools (which, once again, performed well below schools from across the state). But during the follow-up period, the percentage of low-performing students at the FTF school declined, whereas the percentage for the comparison schools did not. Thus, the relative performance of the FTF improved somewhat.

The First Things First Evaluation

Figure 4.19

Changes from Baseline Levels in Middle School Attendance Rates for the 2002 Cohort Middle Schools (Schools U, V, and T): Houston, Texas



SOURCE: MDRC calculations from individual student records from the Houston Independent School District data file.

NOTES: Sample includes 6th- to 8th-grade students from three clusters. Each cluster consists of a First Things First (FTF) middle school matched with a group of between three and fifteen non-FTF schools. The sample consists of students for whom administrative records exist between the 1999-2000 and 2003-2004 academic years.

Each bar represents the "deviation from baseline," or the difference between the baseline level (average across three pre-implementation years) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the baseline for FTF schools and the deviation from the baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics and prior achievement.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in student attendance rates.

Table 4.5

High School MAP Test Scores for First Things First School, Comparison Schools, and the State: Riverview Gardens, Missouri

	Percentage in Bottom 2 Performance Categories					
	Baseline Years			Follow-Up Years		
	Spring 1999	Spring 2000	Spring 2001	Spring 2002	Spring 2003	Spring 2004
11th-grade communication arts						
FTF school	63.8	66.2	56.4	67.8	53.9	62.8
Comparison schools	64.8	64.7	64.4	64.7	63.5	63.8
State	43.2	44.0	38.9	39.1	40.7	39.7
10th-grade math						
FTF school	94.0	93.0	90.9	89.2	82.6	84.0
Comparison schools	87.9	86.6	87.6	86.8	84.9	88.7
State	65.0	64.3	61.3	62.7	59.6	57.0

SOURCE: MDRC calculations from school-level records of state data.

Figure 4.22 (found on page 127) presents estimates of the impacts of FTF on state test scores for middle school students. The top panel in the figure — for communication arts — suggests no consistent pattern in the changes that occurred in the relative performance of the FTF school. In contrast, the bottom panel — for math — presents a consistent pattern of relative improvement in all three follow-up years. However, these estimates are not statistically significant (because of the limited statistical precision of the evaluation design for this site) and thus are only suggestive.[32]

Attendance

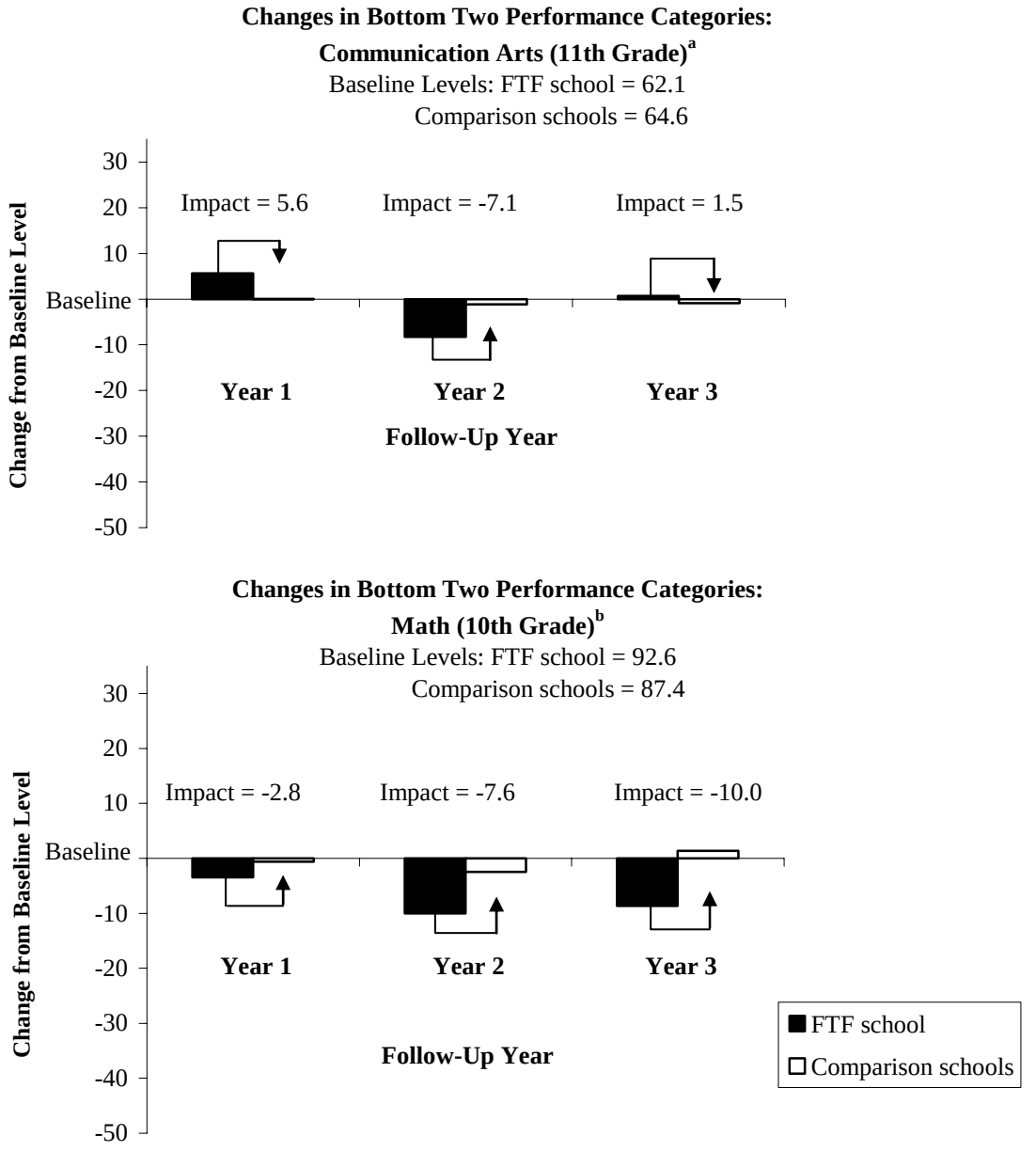
Lastly, Figure 4.23 (found on page 129) presents estimates of the impact of FTF on student attendance rates, which were relatively high for the FTF school and its comparison schools during the baseline period (91.6 percent and 88.6 percent, respectively) and changed very little at any school during the follow-up period.

[32]The minimum detectable effect for math was a change of roughly 24 percentage points. Thus only an impact of this magnitude or larger has a good chance (80 percent statistical power) of being detected, if it exists.

The First Things First Evaluation

Figure 4.20

Changes from Baseline Levels in the Percentage of High School Students Scoring in the Bottom Two Categories of the MAP State Test: Riverview Gardens, Missouri



(continued)

Figure 4.20 (continued)

SOURCE: MDRC calculations from school-level records of state data.

NOTES: Sample includes one First Things First (FTF) high school and eight comparison schools.

Each bar represents the "deviation from baseline," or the difference between the baseline average (three pre-implementation years) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the baseline for the FTF school and the deviation from the baseline for comparison schools.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is a decrease from baseline, which represents a decrease in the percentage of students scoring in the state's bottom two proficiency categories.

[b]The desired change in this measure is a decrease from baseline, which represents a decrease in the percentage of students scoring in the state's bottom two proficiency categories.

The First Things First Evaluation

Table 4.6

Middle School MAP Test Scores for First Things First Schools, Comparison Schools, and the State: Riverview Gardens, Missouri

	Percentage in Bottom 2 Performance Categories					
	Baseline Years			Follow-Up Years		
	Spring 1999	Spring 2000	Spring 2001	Spring 2002	Spring 2003	Spring 2004
7th-grade communication arts						
FTF schools	62.8	61.0	56.0	61.1	66.2	63.8
Comparison schools	73.8	71.0	67.7	68.3	70.6	74.7
State	41.2	41.4	37.4	38.5	38.2	38.8
8th-grade math						
FTF schools	90.9	89.7	84.4	82.4	79.0	80.5
Comparison schools	92.8	87.3	84.2	86.7	88.3	87.3
State	62.5	60.0	56.6	56.5	53.5	53.3

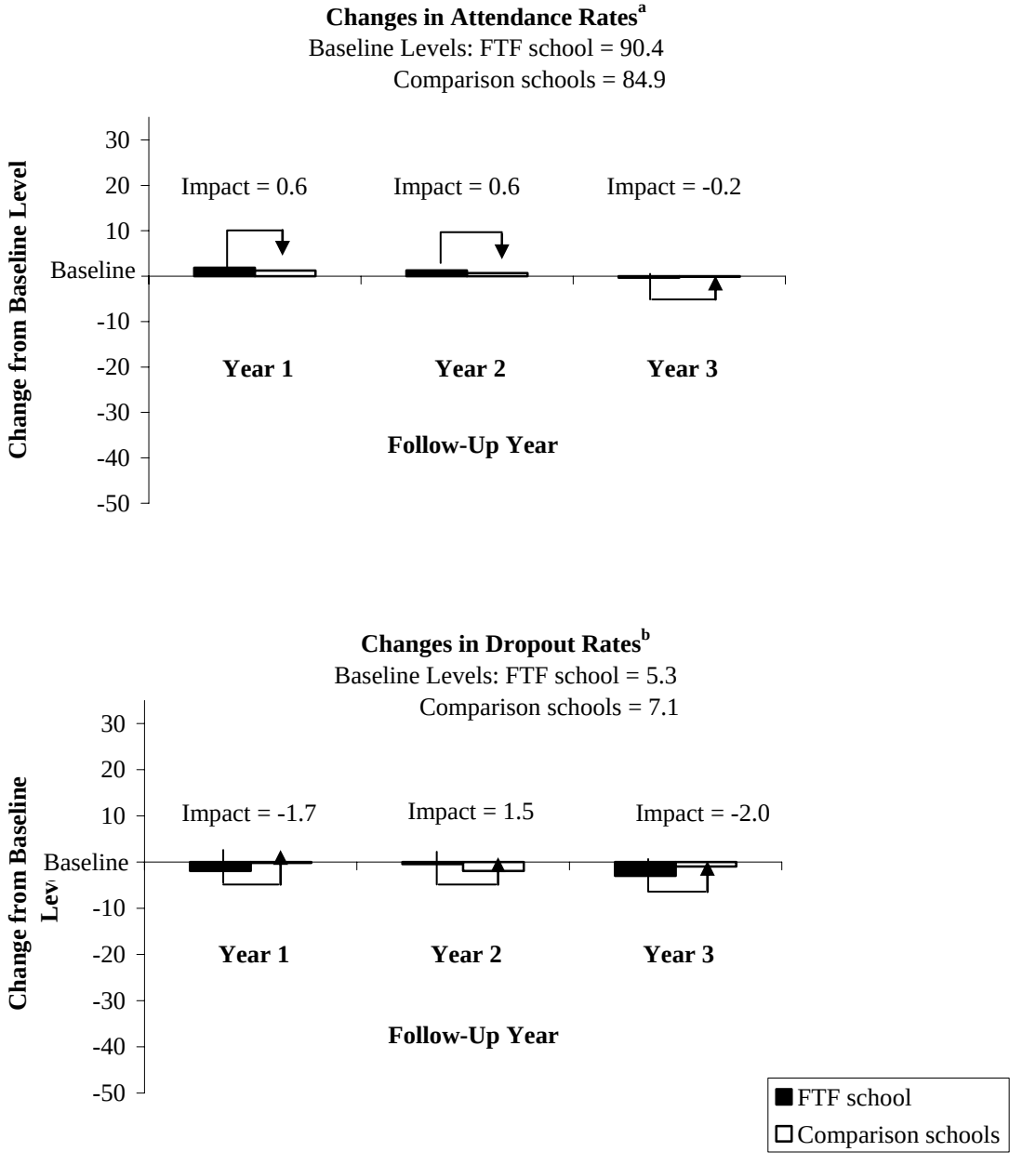
SOURCE: MDRC calculations from school-level records of state data.

NOTE: FTF schools include two First Things First (FTF) middle schools that are treated as one "composite" school in the analysis.

The First Things First Evaluation

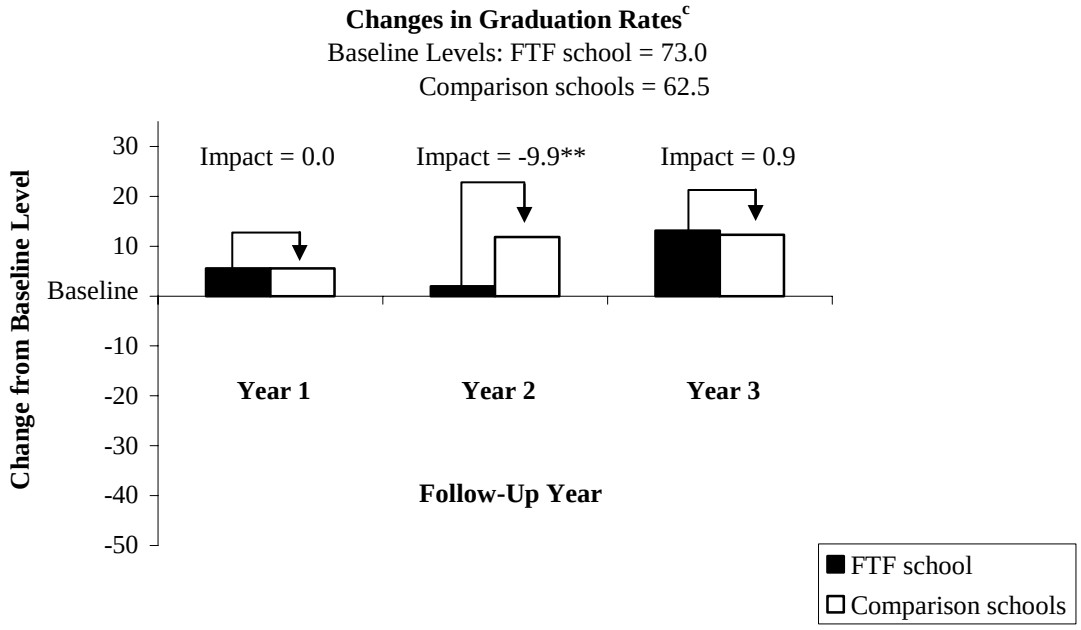
Figure 4.21

Changes from Baseline Levels in High School Attendance, Dropout, and Graduation Rates: Riverview Gardens, Missouri



(continued)

Figure 4.21 (continued)



SOURCE: MDRC calculations from school-level records of state data.

NOTES: Sample includes one First Things First (FTF) high school and eight comparison schools.

Each bar represents the "deviation from baseline," or the difference between the baseline average (three pre-implementation years) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the baseline for the FTF school and the deviation from the baseline for comparison schools.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in school-level attendance rates.

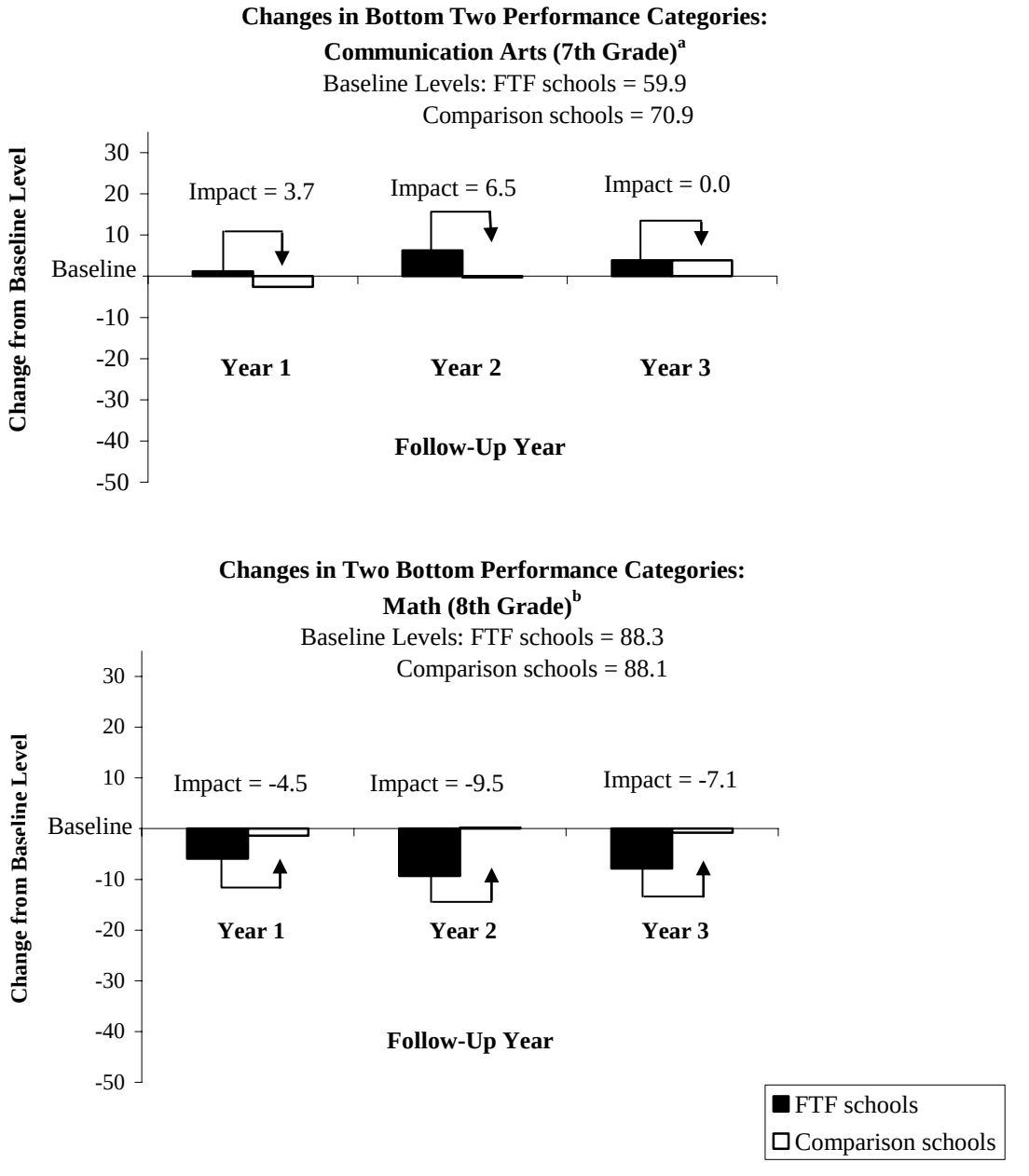
[b]The desired change in this measure is a decrease from baseline, which represents a decrease in school-level dropout rates.

[c]The desired change in this measure is an increase from baseline, which represents an increase in school-level graduation rates.

The First Things First Evaluation

Figure 4.22

Changes from Baseline Levels in the Percentage of Middle School Students Scoring in the Bottom Two Categories of the MAP State Test: Riverview Gardens, Missouri



(continued)

Figure 4.22 (continued)

SOURCE: MDRC calculations from school-level records of state data.

NOTES: Sample includes two "composite" First Things First (FTF) middle schools and twelve comparison schools.

Each bar represents the "deviation from baseline," or the difference between the baseline average (three pre-implementation years) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the baseline for FTF schools and the deviation from the baseline for comparison schools.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is a decrease from baseline, which represents a decrease in the percentage of students scoring in the state's bottom two proficiency categories.

[b]The desired change in this measure is a decrease from baseline, which represents a decrease in the percentage of students scoring in the state's bottom two proficiency categories.

FTF in the Delta Region of Mississippi

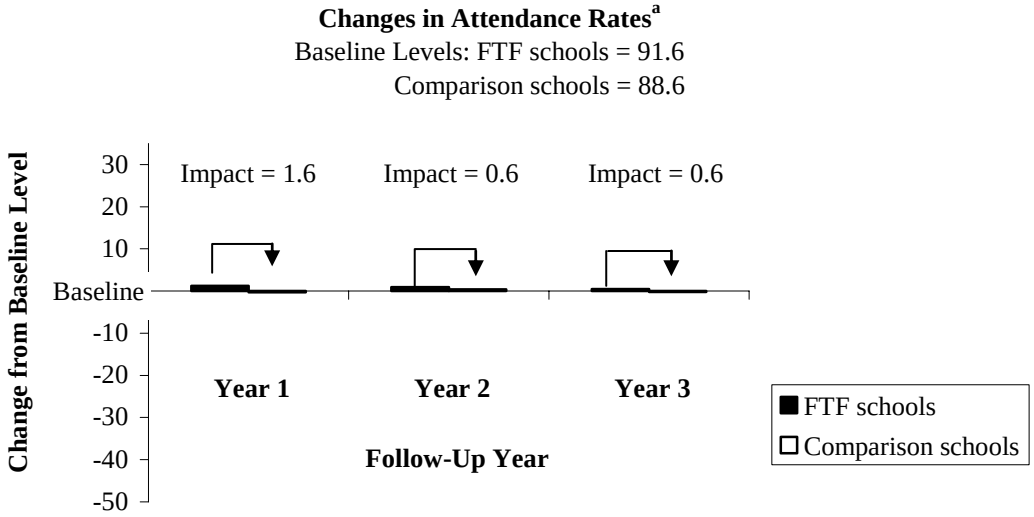
FTF was implemented at two high schools in the Delta Region of Mississippi. Comparison schools were selected (for both FTF schools) from other high schools in the Delta Region plus high schools with large enrollments (for School I) from elsewhere in the state. The impact analysis (presented in Figure 4.1 and repeated below) uses school-level standardized test data,[33] which are available for only three years, due to the newness of the high-stakes Mississippi state test. Spring 2002 data (for the first year of the new test) are used as a quasi-baseline, and spring 2003 and 2004 data are used for follow-up years. (Appendix D shows findings for the individual schools.)

[33]Efforts were made to acquire school-level data on other student outcomes for these sites, but such data have not been obtained.

The First Things First Evaluation

Figure 4.23

Changes from Baseline Levels in Middle School Attendance Rates: Riverview Gardens, Missouri



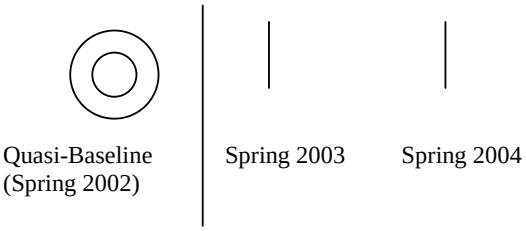
SOURCE: MDRC calculations from school-level records of state data.

NOTES: Sample includes two "composite" First Things First (FTF) middle schools and twelve comparison schools.

Each bar represents the "deviation from baseline," or the difference between the baseline average (three pre-implementation years) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the baseline for FTF schools and the deviation from the baseline for comparison schools.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in school-level attendance rates.



Results were obtained for student performance in English II (there is no test for English I) and algebra on the state test, known as the Subject Area Testing Program (SATP). The SATP English II test is administered each year to tenth-graders, and the algebra test is administered each year to ninth-graders.

The First Things First Evaluation

Table 4.7

High School SATP Test Scores for First Things First Schools, Comparison Schools, and the State: Delta Region of Mississippi

	Percentage Passing		
	Quasi-Baseline Year	Follow-Up Years	
	Spring 2002	Spring 2003	Spring 2004
10th-grade English II			
FTF schools	52.0	71.9	78.5
Comparison schools	55.7	62.9	72.9
State	67.5	76.8	82.0
9th-grade algebra			
FTF school (School I)	61.1	79.6	71.6
Comparison schools	49.7	52.4	75.8
FTF school (School J)	94.0	83.3	97.8
Comparison schools	87.2	86.5	94.3
State	82.1	83.8	91.6

SOURCE: MDRC calculations from school-level records of state data.

Table 4.7 summarizes the passing rates obtained on these tests by the FTF schools, their comparison schools, and the State of Mississippi. Findings for English II are presented for the

two high schools and for their comparison schools combined.[34] These rates are fairly similar for both groups of schools during the quasi-baseline year (spring 2002), with both groups performing less well than the state as a whole. Subsequent passing rates increased for all three groups of schools, but they did so more rapidly for the FTF schools than for their comparison groups.

Findings for algebra are presented separately for each FTF school because the two had vastly different passing rates during the quasi-baseline year. Passing rates for School I (and its comparison schools) were well below the state average and rose rapidly during the next two years. However, there was no consistent pattern in School I's relative performance. Passing rates for School J (and its comparison schools) were so high (94.0 percent and 87.2 percent, respectively, which were well above the state average) during the quasi-baseline year that there was very little room for improvement. Thus, although passing rates by 2004 had reached 97.8 percent and 94.3 percent for School J and its comparison schools, respectively, there was no real change in their relative performance.

Results from a statistical analysis of these data indicate that:

- There is some evidence that FTF improved the performance of high school students on the English II test, but this finding is not statistically significant.

Figure 4.24 illustrates the findings for English II. The two dark bars indicate that passing rates for the FTF schools increased by 19.9 percentage points in 2003 and by 26.5 percentage points in 2004. The two light bars indicate that corresponding increases were 7.0 and 18.1 for the comparison schools. The differences between these improvements for the FTF schools and their comparison schools (12.9 percentage points in 2003 and 8.4 percentage points in 2004) indicate the impact of FTF. Thus, it appears that the reform may have improved relative performance in English II. However, because of the very low statistical precision for the evaluation design at the Mississippi sites, these estimates are not statistically significant and thus contain considerable uncertainty.[35]

Figure 4.25 illustrates the findings for algebra, separately for Schools I and J. Findings for School I (in the top panel of the figure) show erratic and large changes in passing rates during each of the two follow-up years, with School I coming out ahead in 2003 and its comparison schools coming out ahead in 2004. Findings for School J (in the bottom panel) also show an erratic pattern of changes. However, given the limited room for improvement, these changes are much smaller than those in the top panel.

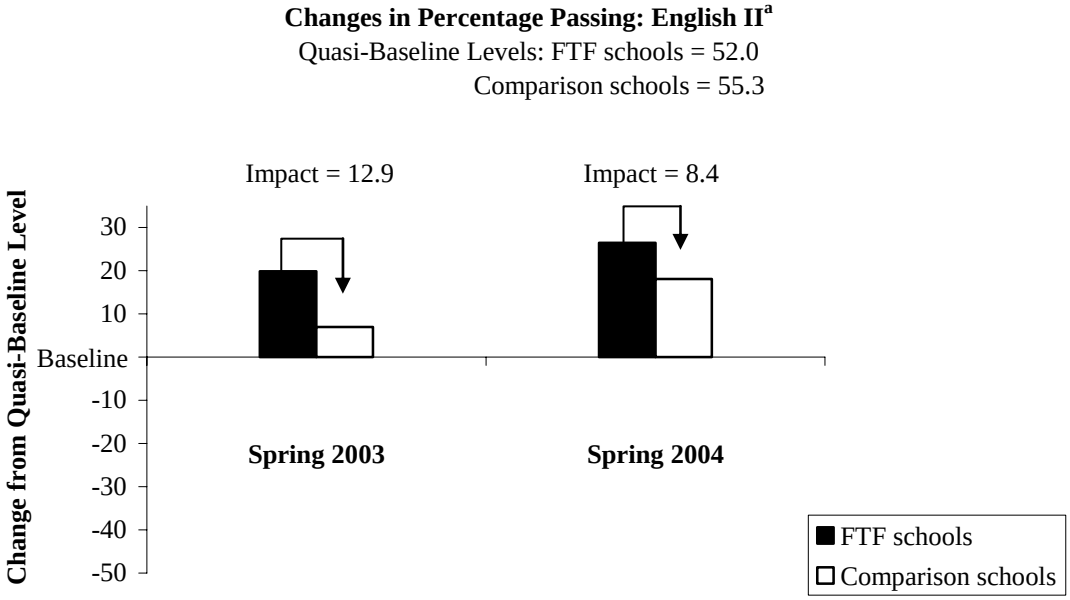
[34]To maximize the statistical power of the analysis for English II, the FTF schools were combined.

[35]The minimum detectable effect for English II was a change of roughly 27 percentage points. Thus, only an impact of this magnitude or larger has a good chance (80 percent statistical power) of being detected, if it exists.

The First Things First Evaluation

Figure 4.24

Changes from Quasi-Baseline Levels in the Percentage of 10th-Graders Passing English II: Delta Region of Mississippi



SOURCE: MDRC calculations from school-level records of state data.

NOTES: Sample includes two First Things First (FTF) high schools and six to ten comparison schools.

Each bar represents the "deviation from quasi-baseline," or the difference between the quasi-baseline level (average in spring 2002) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the quasi-baseline for FTF schools and the deviation from the quasi-baseline for comparison schools.

A two-tailed t-test was applied to differences in deviations from quasi-baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in the percentage of students passing the state English II test.

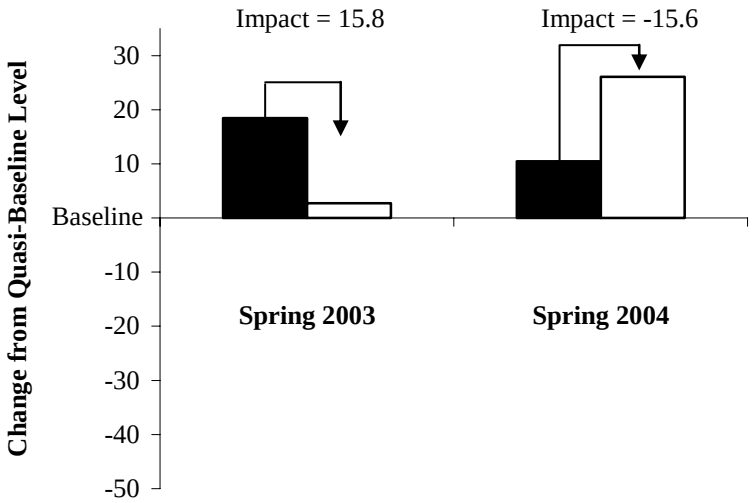
The First Things First Evaluation

Figure 4.25

Changes from Quasi-Baseline Levels in the Percentage of 9th-Graders Passing Algebra: Delta Region of Mississippi

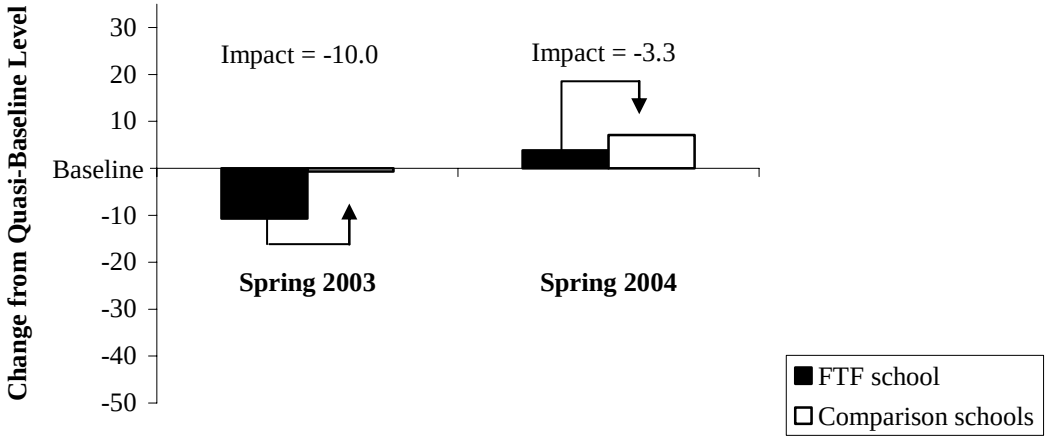
Changes in Percentage Passing for School I: Algebra[a]

Quasi-Baseline Levels: FTF school = 61.1
Comparison schools = 49.7



Changes in Percentage Passing for School J: Algebra[a]

Quasi-Baseline Levels: FTF school = 94.0
Comparison schools = 87.2



(continued)

Figure 4.25 (continued)

SOURCE: MDRC calculations from school-level records of state data.

NOTES: Sample includes two First Things First high schools and four to ten comparison schools. Schools I and J are displayed separately because their algebra pass rates are vastly different and averaging them together would provide a misleading summary of the FTF schools' pass rates over time.

Each bar represents the "deviation from quasi-baseline," or the difference between the quasi-baseline level (average in spring 2002) and the average for the given follow-up year. The "impact" was calculated as the difference between the deviation from the quasi-baseline for FTF schools and the deviation from the quasi-baseline for comparison schools.

A two-tailed t-test was applied to differences in deviations from quasi-baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired change in this measure is an increase from baseline, which represents an increase in the percentage of students passing the state algebra test.

Conclusions

The findings reported in this chapter support the conclusion that FTF markedly increased academic achievement in reading and math and improved other outcomes for high school students and middle school students at its home site, Kansas City, Kansas. The findings also indicate that the school reform initiative increased academic achievement in reading (which is statistically significant) and math (which is not statistically significant) at one of the Houston expansion-site high schools. In addition, there is suggestive evidence (which is not statistically significant) that the reform may have increased high school academic achievement in one subject at Riverview Gardens and one subject at the Mississippi sites. For the most part, however, there is not yet evidence that the expansion sites experienced the pronounced and pervasive impacts that were produced by the reform at its original site. Reflections about the likely reasons for this difference are the subject of Chapter 5.

Chapter 5

Reflections and Lessons

This report tells a complex story about a complex initiative. The implementation findings indicate that mounting First Things First (FTF) is hard work; doing it well requires commitment, persistence, and effort. The impact findings reflect these implementation challenges and accomplishments and present a picture of mixed success. The positive effects of the intervention in its home district of Kansas City, Kansas, were large, statistically significant, pervasive, and sustained. It is impossible to say whether, or how quickly, the schools in other districts to which the initiative subsequently expanded might be expected to register equally compelling effects; although they have not done so at this relatively early point, one high school, School E, has made an impressive start, and other schools have registered some positive but not statistically significant results. In this regard, it is notable that — in a review of the effectiveness of comprehensive school reforms (CSRs) in improving student achievement — the authors concluded that schools implementing CSR models for five years or more had stronger effects than those with briefer periods of implementation.[1]

This chapter reflects on the reasons for the differences in findings between Kansas City and the expansion sites, while acknowledging that many factors may have entered into play. Some of these factors may have to do with limitations of the research design, as noted in Chapter 4. Other factors may not have been measured in this study.[2] What follows represents the investigators' best hypotheses about why the impact findings from Kansas City and the expansion sites differ at this point; other explanations can undoubtedly be adduced.

FTF's implementation in Kansas City, Kansas, points to four conditions that were *sufficient* to produce meaningful impacts on a wide array of outcomes across twelve secondary schools serving disadvantaged populations:

---

[1]See Borman, Hewes, Overman, and Brown, 2004.

[2]For example, in a climate where all schools are under pressure to improve outcomes, some comparison schools may have introduced reforms of their own, a possibility that the evaluation did not have the resources to investigate in depth. In fact, the Houston Independent School District undertook a major high school reform initiative in the second year of FTF operations there. The initiative called for some of the same design elements as were present in FTF — for example, small learning communities (SLCs) and adult advocacy for students. By all accounts, however, actual implementation of the initiative was slow in getting off the ground. Similarly, another reform initiative was introduced into many Mississippi high schools after FTF was implemented in Greenville and Shaw. If comparison schools for the expansion sites were improving their outcomes more than were the comparison schools for Kansas City, estimates of FTF's effects in the expansion sites might be smaller than those for the original site.

1. A districtwide focus, with the district's staying the course for many years in its provision of pressure and supports for the FTF changes
2. Extended follow-up of schools that had been operating FTF — or had been functioning in an environment shaped by FTF — for at least five years
3. A reform model that balanced more personalized learning environments with a comprehensive and intensive approach to improving instruction that emphasized student engagement and curricular alignment and rigor
4. Intensive and responsive technical assistance from providers who assisted the reform and propelled it forward and who were willing to make midcourse corrections where needed

But are all these conditions both *sufficient* and *necessary* to produce significant improvements in student outcomes? The implementation and impact analyses at the expansion sites only partly address this critical policy issue. For example, the experience of School E suggests that district support may not be required if truly extraordinary school-level leadership is in place to drive the reform forward, but it would be unwise to generalize from this single example. The four conditions — and the special circumstances contributing to School E's success — are discussed in the remainder of this chapter.

District Support

From the outset, the Kansas City, Kansas, school district viewed FTF as its major school reform initiative, and key district leaders took thorough ownership of the reform. As detailed in Chapter 2, this meant that central office policies were consistently modified or developed to support successful program implementation. For example, the district assigned curriculum specialists to positions as School Improvement Facilitators (SIFs) to assist principals and teachers in implementing FTF; it established weekly early release time for instructional improvement; and it created a new capacity within the central office to analyze data related to the initiative's intended outcomes.

The central office leadership not only provided support to the Kansas City schools but also exerted pressure on the schools to operate in conformity with FTF principles — particularly as implementation of these principles began to show early signs of improving school climate and outcomes. It reorganized the organizational hierarchy to create two Executive Directors of Instruction, each responsible for overseeing two high schools and their feeder middle and elementary schools. This created a clear line of accountability and responsibility for principals and SIFs. The two executive directors spent a considerable amount of time in the schools under their

supervision, conferring with school administrators, visiting classrooms, and acting in other ways to identify issues, propose solutions, and generally monitor goings-on at the school level.

This combination of supports and pressures remained in place throughout the research period. The very fact that Kansas City has stayed with FTF for so long makes it atypical. In many large urban districts, the story has been one of inconstancy and flux, with each new superintendent bringing in a new reform initiative — and with the job tenure of superintendents being brief. Still, the presence of FTF in Kansas City for what is now nearly a decade shows that securing the requisite stability and commitment is possible.

In contrast, at the scaling-up, or expansion-site, schools, the school districts did not provide similar support and oversight to FTF. Fewer central office staff paid attention to the initiative, and they did so with less consistency over time, than did their counterparts in Kansas City. Indeed, for a period in Riverview Gardens, Missouri, and in one subdistrict in Houston, new administrators turned away from FTF, and implementation largely ground to a halt in the schools under their jurisdiction. In the Houston Independent School District, where FTF was mounted in only 3 of 28 comprehensive and magnet high schools and only 4 of 32 middle schools, the school district's inconsistent attention to the FTF schools was predictable. In both Greenville and Shaw, Mississippi, FTF was the district's chosen vehicle for high school reform, and it had the ongoing support of high-level district leadership. But in Greenville, the superintendent's endorsement was not always backed up by the actions of key central office staff, and in Shaw, the central office staff was simply too small to provide consistent oversight and assistance.

Extended Follow-Up

By the spring of 2002, significant impacts on math scores were first evident in Kansas City, Kansas, and, the next year, there were significant impacts on reading scores as well. At the earlier point, the first group of Kansas City, Kansas, schools to mount the initiative had been involved either in planning for or implementing FTF for five years, and the last clusters of schools had been similarly engaged in planning and implementation for three years.

Unfortunately, the impact evaluation that was conducted in Kansas City, Kansas, provides little guidance on just how long a period is needed to produce program impacts. That some Kansas City schools had been formally operating as FTF schools for only two to three years when they registered impacts does not mean that schools in other districts could necessarily achieve similar impacts after operating FTF for two or three years. The Kansas City district's focus on such themes as personalization, literacy, active engagement strategies, and standards-based instruction extended to *all* schools well before the last schools were officially phased in. With this phase-in came the formation of small learning communities (SLCs) and additional technical assistance from the Institute for Research and Reform in Education (IRRE). In other

respects, however, these schools were already operating parts of the intervention before they formally joined the initiative.

The Kansas City, Kansas, example is consistent with the results of the Borman study cited above, which showed that comprehensive school reforms that had been in place for at least five years yielded stronger effects than those in operation for shorter periods. But this does not mean that it will necessarily take five years of sustained district support for FTF before significant impacts will be apparent. The institution of a new state test in Kansas in 2001 established the "baseline" for the Kansas City impact analysis in this report, and when some of the schools in the first cluster of Kansas City schools registered impacts in 2002, they had been planning for or operating FTF for five years. It is possible that impacts would have been registered at an earlier point in the schools' FTF trajectories, had the analysis been able to reach backward in time to measure such impacts.[3] The results for School E, discussed below, indicate that, with very strong implementation of the intervention, impacts may be achieved at a much earlier point, but it would be unwise to generalize from a single example.

Balancing Personalization and Instructional Improvement

SLCs and the Family Advocate System are key elements of FTF aimed at promoting a more personalized environment in large, impersonal schools. In both Kansas City, Kansas, and the expansion sites, FTF was accompanied by a marked increase in students' feelings of being liked and supported by their teachers. This is an important element in the FTF theory of change described in Chapter 1, because, as the theory goes, students who feel that their teachers care about them will work harder on their schoolwork, and teachers who care about their students will work harder to make their classes challenging and engaging.

By all accounts, the SLCs enabled teachers and students to develop closer relationships with each other. Kansas City — where the Family Advocate System was not part of the original program model and was phased in only gradually over time — provides a sort of natural experiment. The fact that schools in Kansas City produced substantial impacts on attendance and achievement before family advocacy was fully in place strongly suggests that that component of FTF may be less essential than SLCs and instructional improvement for producing effects on these outcomes. On the other hand, it may also be that the impacts produced by the Kansas City schools would have been even larger had the Family Advocate System been operating in all schools. And the potential benefits of a well-functioning advocacy system are unarguable: en-

[3]Some Kansas City, Kansas, schools registered significant early impacts on attendance, which is often viewed as an early sign that students are more involved in learning and thus more likely to exhibit higher test scores and, ultimately, to graduate. None of the expansion sites has yet had statistically significant effects on this outcome.

suring that every student has an adult in the school to whom he or she can turn and who is responsible for monitoring and assisting the student's progress, creating a more positive relationship between the school and the family, and working with parents to promote their children's academic success.

At the expansion sites, an increase in students' feelings of support from their teachers was not consistently accompanied by greater reported academic engagement, as measured on the scales presented in Chapter 3, nor by positive and statistically significant impacts on achievement or other outcomes. It may well be that increases in engagement — and, concomitantly, in test scores — are largely the result of instruction that is more engaging, more demanding, and more closely aligned with the curriculum and standards underlying the state tests. These are issues that have become the primary focus of IRRE and its partner schools now that the structural dimensions of FTF are in place at the expansion sites.

Instructional improvement has been a key element of FTF since the initiative's earliest days in Kansas City, Kansas, and district and school leaders in both Kansas City and the expansion sites have recognized its importance. As Table 2.1 indicates, during FTF's first year of implementation, the Kansas City district leadership announced a districtwide focus on improved literacy as part of FTF; the approach included such components as literacy-centered professional development activities. In 1999-2000, the district introduced a standards-based curriculum into the schools and issued an instructional guide delineating the characteristics of high-quality teaching and learning. And in 2000-2001, administrators began systematic visits to classes to watch for engaging and standards-based instruction. Thus, by the time the last two clusters of schools began to operate FTF, a full instructional improvement plan stressing high standards, literacy, and the use of engaging pedagogical strategies was in place in the Kansas City, Kansas, district. Although the result of all these efforts was not high-quality teaching in every classroom, over time, instruction throughout the system came to be more engaging, and students came to see it as more demanding as well.

In contrast, early instructional improvement efforts in the scaling-up sites were much less systematic. The professional development activities that IRRE first organized at the expansion sites — instruction first in read-alouds and then in cooperative learning strategies — supplied an approach to developing literacy across the disciplines and techniques for engaging students in learning. But the training placed relatively little emphasis on embedding these strategies in particular content areas (like math or science). Not until the third year of the demonstration did IRRE begin deploying the elements of its current comprehensive approach that centers on standards-based instruction in content areas and more effective use of teachers' common planning time.

The FTF experience suggests that striking a balance between the creation of personalized environments and an emphasis on instructional improvement is not easy. The expansion

schools were able to put SLCs in place quickly, during the demonstration's planning year and first year of implementation. But until the final year of the demonstration, conversations in the SLCs centered on individual students who presented behavioral or learning problems or on SLC events — field trips, student recognition ceremonies, celebrations, and so on. Such conversations helped to build team spirit but had little to do with instructional practice. Not until the 2003-2004 academic year did SLCs routinely turn the spotlight on the classroom (even then, however, this did not happen in all schools or all SLCs) — a development that reflects IRRE's own evolution and its adoption of a more thorough and more systematic approach to instructional improvement.

It seems likely that if SLCs have a role in enhancing student achievement, it is because they serve as a setting in which discussions about individual students are interwoven with discussions of pedagogy and curriculum. At what point in their development SLCs should begin to address instruction is an open question. In a conversation with researchers, IRRE staff members noted that it had been a mistake to introduce instructional improvement activities to teachers during the FTF planning year. At that point, they noted, teachers were too preoccupied with "relationship" questions — Which other teachers will be in my SLC? Which students will I have? How will we all get along? — to pay adequate attention to efforts to improve teaching. IRRE has modified its approach, directing its planning-year instructional improvement efforts primarily toward a select group of instructional and content-area leaders (although the faculty as a whole also receives an orientation to the concepts of engagement, rigor, and alignment with state and local standards during the planning year).

Still, changing instructional practice has proved to be an especially challenging aspect of FTF — as it is likely to be of any school reform initiative — and it is precisely for this reason that discourse about teaching and learning needs to start early and continue regularly. Perhaps it is unrealistic for SLCs to serve as a major venue for discussions of instructional improvement during the planning year, but it seems critical to get to this point as quickly as possible. In accelerating SLC formation in the second group of expansion-site schools (as described in Chapter 2), IRRE took a major step in this direction. And in designing and disseminating procedures and forms for SLC members and administrators to use in describing and discussing instruction, it took another such step.

Intensive Technical Assistance

IRRE's president, James Connell, was on the scene in Kansas City, Kansas, through the planning period and early implementation years of FTF in the district. Visiting the district approximately every six weeks and available by phone between visits, he provided ongoing support and advice to the superintendent, other district administrators, principals, School Improve-

ment Facilitators (SIFs), and teachers. He exerted pressure as well, urging the district to fulfill its stated commitment to move the reform forward.

The addition of so many expansion sites at once stretched IRRE's capacity, as did the organization's decision to bring FTF to secondary schools in two more large urban districts that were not part of the Scaling-Up First Things First Demonstration. As a consequence, staff at some of the expansion sites felt that while they had received considerable attention during the planning year and the first implementation year, IRRE's support waned thereafter.

Over the course of the demonstration, IRRE produced a number of documents (for example, a guide to family advocacy and a set of tools for measuring engagement, alignment, and rigor in instruction) that could serve as "self-help" instruments for school and district personnel. While not a substitute for in-person technical assistance, the documents crystallized IRRE's vision of what schools needed to accomplish and the steps that school staff members could take to get there.

School E

School E, an expansion-site high school that registered notable increases in student achievement, lacked strong district support. What it did have was a thoughtful, dedicated, and smart principal who, more than any other administrator at the expansion schools, turned regularly to IRRE for encouragement, assistance, and a sounding board. The school also benefited from the presence of an SIF who was trusted by the principal, respected by the faculty, and seen as knowledgeable about instruction.

Among the scaling-up sites, School E was most like the Kansas City, Kansas, schools, with high-quality FTF implementation and a distinctwide focus on instructional improvement. (As Table 2.3 shows, School E shared with one middle school the distinction of having the highest implementation rating.) Its SLCs were entrusted with making important decisions; the Family Advocate System was successfully launched; and teacher-student relationships were strong.

From the outset, the principal of School E recognized that changing teaching practices presented the biggest implementation hurdle. With moral support from his immediate supervisor but little concrete assistance from district administrators, he undertook a number of steps to make instruction at the school more challenging and engaging. He provided the school's assistant principals — whose role had previously centered on student discipline — with training so that they could serve as instructional leaders. Serving as SLC administrators, the assistant principals attended SLC meetings regularly and also monitored instruction in the SLC teachers' classrooms. The principal, who was also something of a "data wonk," worked with IRRE to figure out a way to disaggregate data collected by the school district so that teachers in each

SLC could see how their students were faring in terms of test scores and other outcomes and could adjust their instruction accordingly.

Each SLC in School E was expected to devote one of the two 90-minute SLC meetings per week to discussions of instructional improvement — a standard now for all FTF schools — with the choice of which particular area of instruction to focus on being left to SLC members and administrators. SLCs also drew up instructional improvement calendars. Twice a month, SLC meetings focused on peer observations or examination of student work. In department meetings, faculty members also worked on aligning curriculum with standards and developing common assessments and grading rubrics. According to field research reports, teachers initially felt somewhat threatened working with peers in this way, but over time they came to feel that they were learning a lot from it. In this regard, it is notable that School E was the only school to register statistically significant increases in teachers' feelings of both support and engagement. In short, the experience of School E suggests that where discourse about instruction is thoroughly integrated into the life of the school and into teachers' consciousness, it can have a rich payoff in terms of enhanced student achievement.

It is, of course, possible that, with its energetic, committed principal, School E would have improved with or without FTF. But FTF was the initiative to which the principal devoted his energy and commitment. In any event, the experiences of FTF in Kansas City, Kansas, and in the expansion sites suggest that school reform is too important — and too difficult — to depend on an exemplary principal. The success of First Things First in Kansas City points to the critical role that districts can play in providing a unified message, a supportive context, and the shared expectations that help all educators to keep their eyes on the prize: better teacher-student relationships and improved teaching and learning in the classroom.

Appendix A

Measuring the Implementation of First Things First

A key objective of the First Things First (FTF) evaluation was to develop quantitative measures of the extent of implementation of key components of the initiative, and of the initiative as a whole, at individual schools and for groups of schools.[1] In order to do this, it was necessary to identify key dimensions related to the successful operation of small learning communities (SLCs), the Family Advocate System, and instructional improvement and then to develop a method for assigning numerical ratings to these dimensions.

Identifying Structural and Functional Dimensions of Implementation

The left-hand column of Appendix Table A.1 lists the key components and dimensions of implementation that are measured in this study, and the middle column presents a more complete definition of each dimension. By design, these dimensions relate to the seven "critical features" that figure prominently in early descriptions of the initiative by the Institute for Research and Reform in Education (IRRE) and that are therefore of substantive importance to the initiative.[2] The dimensions are both structural and functional in nature — that is, they relate both to changes in formal configurations (for example, scheduling, student assignment patterns) and to the way that teachers and students behave within these configurations.

Deriving Ratings of Implementation of the Dimensions

The ratings of implementation are based mainly on reports prepared by the field researchers in Houston and Riverview Gardens midway through the 2003-2004 school year — that is, two and a half years after program startup at the 2001 cohort schools and one and a half years after startup at the 2002 cohort schools. In these reports, the researchers were asked to address questions specifically related to the dimensions shown in Appendix Table A.1 and to describe the progress that each of the schools they were studying had made in implementing these.

Two MDRC central office researchers subsequently read the field research reports independently, assigning a quantitative rating to each dimension of implementation. Each dimension was rated on a scale of 1 to 4, where 1 indicated no implementation of the dimension,

[1]Variation in implementation also occurred among the small learning communities (SLCs) within an individual school. It would have been prohibitively costly to collect information on all the SLCs, however. Moreover, since program impacts could not be evaluated at the SLC level, it would not have been possible to relate variation in SLC implementation of FTF to variation in program impacts.

[2]Potential dimensions were eliminated from consideration if it was known in advance that there was little variation on these dimensions among schools. For example, it was known that virtually all high school SLCs (except for special "catch-up" academies for ninth-graders and for communities geared toward English language learners) covered all four years of school and that all were thematic (at least in name).

Appendix Table A.1

Structural and Functional Dimensions of Implementation to Be Measured

Key Component and Dimension	Definition of Dimension	Criteria for Judging Extent of Implementation
Small learning communities (SLCs)		
Structural dimensions		
Purity	Students take their core-subject classes from teachers in their SLCs, and teachers primarily teach students who are in their SLCs.	Maximum rating of 4 given when percentage of language arts and math classes composed exclusively of SLC students (as indicated by sample of class rosters) is 100 percent.
Adequacy of common planning time for teachers	Teachers have at least 180 minutes a week of common planning time for meetings.	Maximum rating of 4 given when schedule allows at least 180 hours a week of common planning time.
Functional dimensions		
Personalization	Teachers and students in SLCs have developed ties within and across groups.	Maximum rating of 4 given when field research reports indicate that SLCs have distinct identities, there are close ties between teachers and students and among teachers within SLCs.
Decision-making	SLC staff have a role in making decisions about such issues as scheduling, hiring, staffing, and expenditure of funds.	Maximum rating of 4 given when field research reports indicate that SLCs have role in decisions about scheduling, school structure, discipline, hiring, and budget.
Accountability	Teachers feel responsible for academic outcomes of students within their SLCs; they hold each other accountable and assist each other to achieve better outcomes; and they review student outcomes and develop plans and timetables for improving these outcomes.	Maximum rating of 4 given when field research reports indicate that SLCs have developed specific statistical/numerical goals for student improvement, review progress against standards, teachers hold peers accountable and confront each other about instructional and other concerns.

(continued)

Appendix Table A.1 (continued)

Key Component and Dimension	Definition of Dimension	Criteria for Judging Extent of Implementation
Family Advocate System		
Structural dimensions		
Universal assignments	All students have family advocates.	Maximum rating of 4 given when field research reports indicate that all students have family advocates.
Functional dimensions		
Close relationships between advocates and students	Advocates and students are close to each other and believe the relationship is helpful.	Maximum rating of 4 given when field research reports indicate that meetings with parents are individualized, face-to-face (not group meetings or phone conferences), that teachers know students well and maintain records on the students for whom they serve as advocates.
Meaningful Family Advocate Period activities	Teachers use the period to meet with their students and/or discuss important topics with them.	Maximum rating of 4 given when field research reports indicate that frequency and duration of period are adequate to allow for meaningful activities; a variety of meaningful/constructive/ activities (such as discussions of current events, activities centered on conflict resolution or building self-esteem) take place during the period; teachers express positive opinions about the effectiveness of the period reported.
Instructional Improvement		
Structural dimensions		
Extended instructional time in language arts and math	Language arts and math classes are longer and/or meet more frequently than classes in other subjects.	Maximum rating of 4 given when field research reports that students enroll in supplementary classes or spend more class time in existing language arts and math classes.
Reduced student-teacher ratios	Students receive instruction in language arts and math in classes where the student-teacher ratio is 15:1 (or lower than before FTF was implemented).	Maximum rating of 4 given when field research reports indicate that a student-teacher ratio of 15:1 or lower exists in language arts and math classes throughout all grades and SLCs.

(continued)

Appendix Table A.1 (continued)

Key Component and Dimension	Definition of Dimension	Criteria for Judging Extent of Implementation
Functional dimensions		
Alignment of curriculum with state/local standards	Teachers align what they teach with what is prescribed under state and local standards, or are making significant progress toward that goal.	Maximum rating of 4 given when field research reports indicate that teachers within a subject area routinely discuss alignment of curriculum with state and local standards, teachers report planning lessons with standards in mind.
Active learning	Teachers routinely make use of active learning strategies in their lessons.	Maximum rating of 4 given when field research reports indicate consistent use of Kagan or other cooperative learning strategies, SLC meetings include discussion of Kagan or other cooperative learning strategies, and teachers report planning lessons to incorporate Kagan strategies.
High, clear academic standards	Teachers meet to discuss what constitutes high-quality student work, have established (or are moving toward) common grading standards, and hold high expectations for students. Students believe that teachers have high expectations of them and that standards are clear.	Maximum rating of 4 given when field research reports indicate that SLC and subject-area teachers have adopted common student evaluation standards, teachers express high expectations for students, teachers regularly discuss student work to determine level, clarity of assignments.
Theme-related instruction	Teachers develop special thematic units or relate regular instruction to the theme of their SLC.	Maximum rating of 4 given when field research reports indicate that many SLCs have put in place theme-based/interdisciplinary units, teachers infuse the SLC theme into core-subject classes, SLCs conduct theme-related extracurricular activities (such as Career Days, mentoring by adults working in the area suggested by the theme).
Development of teacher professional learning community	Teachers meet to discuss professional development needs, instructional techniques, curriculum, and so on, and see their colleagues as resources for improving their practice.	Maximum rating of 4 given when field research reports indicate that teachers within their SLCs regularly conduct peer observations, discuss best practices and student work, participate actively in in-service training and other professional development activities, and seek each others' counsel about instruction.

SOURCE: MDRC analysis of field research reports.

2 indicated that implementation had begun but was relatively undeveloped, 3 indicated good implementation but with room for growth, and 4 indicated that implementation had reached a high level. Because a 4-point scale seemed insufficiently nuanced, midpoints (1.5, 2.5, and 3.5) were added, creating a 7-point scale. The right-hand column of Appendix Table A.1 shows the criteria that the researchers considered in rating the implementation of each dimension. The two researchers then discussed their separate ratings. These were often identical, and where they differed, it was rarely by more than half a point and easy to achieve a consensus rating.

For the Mississippi sites, the general approach was similar, but since the on-site researcher's position ended at the conclusion of the 2002-2003 school year, the data source was different. The research director examined the interviews completed with school officials in Greenville and derived ratings accordingly. The interviews with administrators in Shaw yielded less complete information (in part because the Shaw principal had been ill for much of the year). A Mississippi-based technical assistance provider who had worked in Shaw was asked to rate implementation at that site.[3] Unlike the other ratings, those for the Mississippi schools thus reflect the state of implementation by the end of the 2003-2004 academic year rather than midway through it.

After the MDRC researchers had compiled the ratings, they asked IRRE staff members whether the resulting rank order of schools corresponded with their own experience. IRRE reviewers reported that the lineup of school ratings accurately reflected variation in the extent of implementation among the schools. This independent corroboration of the field research ratings by the party in the best position to make such an assessment provides considerable reassurance about the underlying validity of the ratings.

At the same time, the reader is cautioned that the ratings are more useful in a relative than in an absolute sense and that small differences in ratings may have little or no meaning. It would be difficult to conclude, for example, that a school rated 2.9 had implemented FTF far more completely than one rated 2.8. On the other hand, a school rated 3.1 can be safely assumed to have implemented the initiative at a higher level than one rated 2.5. In general, the larger the differences and the more consistently they show up across groups of schools, the greater the confidence that should be placed in their meaningfulness.

Initially, the researchers planned to use teacher survey and student survey responses as well as the field research to create a rank order of schools, and considerable effort went into the creation of scales for measuring the dimensions in Appendix Table A.1. The lack of correspondence between school ratings derived from the field research and the teacher survey made the

[3]While there was some concern that the technical assistance provider would not be objective in her ratings, the ratings themselves suggested otherwise.

use of the teacher survey as a data source for this purpose highly suspect, however.[4] The student surveys were aligned somewhat more closely with the field research, but still imperfectly. In the end, the researchers decided to report aggregate findings from the teacher and staff surveys where appropriate but not to use these two data sources to develop summary ratings of implementation at individual schools.

[4]One problem was that there was little variation in teacher ratings of the extent to which implementation had progressed at their schools. From the teacher surveys, the researchers developed scales with high reliabilities (alphas of 0.7 or higher) to measure implementation along eight of the dimensions shown in Appendix Table A.1. They then averaged the ratings across these eight dimensions to derive teacher-survey-based implementation ratings for each school. The range of average scores was 2.9 to 3.1. (In contrast, the range of field research ratings on these same eight dimensions was from 2.3 to 3.5, and the ranking of schools along these eight dimensions accorded closely with the ranking of schools along all fifteen dimensions measured in the field research.)

Appendix B

Support and Engagement Among Teachers and Students

Results for Individual Schools

The methodology for measuring support and engagement among teachers and students in the First Things First (FTF) evaluation is described in Chapter 3. This appendix presents findings for each of the twelve expansion-site schools, indicated by letter.

Teachers' Feelings of Support

Appendix Table B.1 shows statistically significant increases in support among teachers at one high school and one middle school between the planning year and the end of the follow-up period.

Teachers' Feelings of Engagement

Appendix Table B.2 shows statistically significant increases in teacher engagement at three of the six high schools between the planning year and the end of the follow-up period. (It is notable that teachers at School E reported significant increases in both support and engagement in the third year of follow-up.) There were no changes in engagement registered by teachers at any of the middle schools.

Students' Feelings of Support

Appendix Table B.3 shows that students at three of the six high schools registered statistically significant increases in their feelings of support from their teachers between the planning year and the last year of follow-up. Among middle school students, the picture was more mixed: Students at two schools reported significant increases in support, while students at two other schools registered declines; before FTF was implemented, the latter two schools had been organized into student-teacher groupings that resembled the initiative's small learning communities (SLCs).

Students' Feelings of Engagement

Appendix Table B.4 indicates that scores for student engagement fell significantly at three of the six high schools between the planning year and the last year of follow-up. At three of the six middle schools, there were statistically significant increases in engagement; over the same period, two other middle schools experienced statistically significant decreases.

The First Things First Evaluation

Appendix Table B.1

Teachers' Average Scale Scores for Individual Schools: Support

Schools	Planning Year	Year 1	Year 2	Year 3	Effect Size Year 2	Effect Size Year 3
All high schools	2.74	2.72	2.77	NA	0.05	NA
School E	2.65	2.71	2.81	2.82	0.25	0.27 **
School F	2.74	2.62	2.73	NA	-0.02	NA
School G	2.66	2.65	2.80	NA	0.23	NA
School H	2.68	2.78	2.55	2.58	-0.21	-0.17
School I	2.82	2.78	2.87	2.59	0.09	-0.34 ***
School J	3.22	3.08	3.18	3.27	-0.10	0.12
All middle schools	2.80	2.74	2.78	NA	-0.04	NA
School S	3.00	2.88	2.85	2.94	-0.25	-0.10
School T	2.79	2.59	2.94	NA	0.26	NA
School U	2.71	2.72	2.34	NA	-0.65	NA
School V	2.72	2.86	2.96	NA	0.42 **	NA
School W	2.79	2.67	2.93	2.83	0.29	0.08
School X	2.65	2.54	2.58	2.69	-0.13	0.06

SOURCES: MDRC calculations based on 2001, 2002, 2003, and 2004 First Things First staff surveys.

NOTES: Scale scores range from 1 (the lowest possible outcome) to 4 (the highest possible outcome).

Statistical significance levels are indicated as *** = 1 percent; ** = 5 percent; * = 10 percent. Statistical significance is indicated for differences between the planning year and the second and third years of implementation.

The size of the sample used to measure changes in support among teachers ranged from 553 to 557 across all high schools and from 354 to 365 across all middle schools between the planning year and the third implementation year.

The size of the sample used to measure changes in support among teachers ranged from 316 to 364 across 2001 cohort high schools and from 147 to 164 across 2001 cohort middle schools between the planning year and the third implementation year.

The First Things First Evaluation

Appendix Table B.2

Teachers' Average Scale Scores for Individual Schools: Engagement

Schools	Planning Year	Year 1	Year 2	Year 3	Effect Size Year 2	Effect Size Year 3
All high schools	3.00	3.00	3.07	NA	0.15 ***	NA
School E	3.04	3.05	3.12	3.17	0.17	0.28 **
School F	3.01	2.95	3.08	NA	0.16	NA
School G	2.92	2.90	3.07	NA	0.29 **	NA
School H	2.95	3.09	3.00	3.10	0.12	0.37 **
School I	3.01	2.97	3.08	2.86	0.15	-0.23 **
School J	3.16	3.14	3.11	3.11	-0.14	-0.14
All middle schools	3.02	3.00	3.07	NA	0.12	NA
School S	3.11	2.98	3.20	3.15	0.22	0.12
School T	3.04	2.85	3.05	NA	0.02	NA
School U	2.96	3.08	3.00	NA	0.10	NA
School V	3.00	3.06	3.07	NA	0.14	NA
School W	2.98	2.98	3.02	3.04	0.10	0.14
School X	2.99	3.02	2.94	3.03	-0.12	0.10

SOURCES: MDRC calculations based on 2001, 2002, 2003, and 2004 First Things First staff surveys.

The size of the sample used to measure changes in support among teachers ranged from 553 to 557 across all high schools and from 354 to 365 across all middle schools between the planning year and the third implementation year.

The size of the sample used to measure changes in support among teachers ranged from 316 to 364 across 2001 cohort high schools and from 147 to 164 across 2001 cohort middle schools between the planning year and the third implementation year.

The First Things First Evaluation

Appendix Table B.3

Students' Average Scale Scores for Individual Schools: Support from Teachers

Schools	Planning Year	Year 1	Year 2	Year 3	Effect Size Year 2	Effect Size Year 3
All high schools	2.75	2.79	2.81	NA	0.10 ***	NA
School E	2.83	2.85	2.88	2.92	0.09 ***	0.16 ***
School F	2.69	2.69	2.69	NA	0.00	NA
School G	2.75	2.74	2.77	NA	0.05	NA
School H	2.71	2.85	2.87	2.85	0.28 ***	0.24 ***
School I	2.80	2.85	2.86	2.80	0.10 ***	0.00
School J	2.70	2.87	2.84	2.84	0.24 ***	0.24 ***
All middle schools	2.73	2.76	2.76	NA	0.05 **	NA
School S	2.76	2.78	2.81	2.90	0.08 **	0.22 ***
School T	2.69	2.74	2.66	NA	-0.05	NA
School U	2.52	2.67	2.75	NA	0.37 ***	NA
School V	2.72	2.72	2.71	NA	-0.02	NA
School W	2.88	2.84	2.77	2.82	-0.19 ***	-0.10 *
School X	2.99	2.95	2.91	2.88	-0.14	-0.19 **

SOURCES: MDRC calculations based on 2001, 2002, 2003, and 2004 First Things First student surveys.

NOTES: Scale scores range from 1 (the lowest possible outcome) to 4 (the highest possible outcome).

Statistical significance levels are indicated as *** = 1 percent; ** = 5 percent; * = 10 percent.

Statistical significance is indicated for differences between the planning year and the second and third years of implementation.

"Effect size" is a metric used to describe the magnitude of a difference. Effect sizes between 0 and 0.32 may be considered small.

The size of the sample used to measure changes in support from teachers among students ranged from 7,209 to 7,877 across all high schools and from 5,438 to 5,699 across all middle schools between the planning year and the second implementation year.

The size of the sample used to measure changes in support from teachers among students ranged from 4,535 to 4,615 across 2001 cohort high schools and from 2,322 to 2,472 across 2001 cohort middle schools between the planning year and the third implementation year.

The First Things First Evaluation

Appendix Table B.4

Students' Average Scale Scores for Individual Schools: Engagement

Schools	Planning Year	Year 1	Year 2	Year 3	Effect Size Year 2	Effect Size Year 3
All high schools	3.27	3.22	3.22	NA	-0.11 ***	NA
School E	3.30	3.22	3.23	3.27	-0.15 ***	-0.06 **
School F	3.12	3.14	3.12	NA	0.00	NA
School G	3.22	3.17	3.21	NA	-0.03	NA
School H	3.34	3.29	3.29	3.28	-0.11 ***	-0.14 ***
School I	3.38	3.29	3.29	3.20	-0.20 ***	-0.39 ***
School J	3.39	3.39	3.34	3.33	-0.12	-0.14
All middle schools	3.18	3.21	3.23	NA	0.10 ***	NA
School S	3.26	3.13	3.20	3.24	-0.12 ***	-0.04
School T	3.15	3.26	3.23	NA	0.17 ***	NA
School U	2.91	3.13	3.15	NA	0.41 ***	NA
School V	3.18	3.25	3.24	NA	0.13	NA
School W	3.33	3.32	3.32	3.38	-0.02	0.11 **
School X	3.37	3.32	3.34	3.28	-0.07	-0.22 ***

SOURCES: MDRC calculations based on 2001, 2002, 2003, and 2004 First Things First student surveys.

NOTES: Scale scores range from 1 (the lowest possible outcome) to 4 (the highest possible outcome).

Statistical significance levels are indicated as *** = 1 percent; ** = 5 percent; * = 10 percent.
Statistical significance is indicated for differences between the planning year and the second and third years of implementation.

"Effect size" is a metric used to describe the magnitude of a difference. Effect sizes between 0 and 0.32 may be considered small.

The size of the sample used to measure changes in engagement among students ranged from 7,209 to 7,877 across all high schools and from 5,438 to 5,699 across all middle schools between the planning year and the second implementation year.

The size of the sample used to measure changes in engagement among students ranged from 4,535 to 4,615 across 2001 cohort high schools and from 2,322 to 2,472 across 2001 cohort middle schools between the planning year and the third implementation year.

Appendix C

Estimating the Impacts of First Things First

This appendix describes how the impacts of First Things First (FTF) were estimated based on the evaluation design for each site. These designs, which are described in Chapter 4, are repeated for readers' reference in Figure C.1.[1] First, the appendix describes the outcome measures used for the impact analysis at each site. Then it describes how comparison schools were selected to help gauge what changes in outcomes would have occurred in the absence of the reform initiative. Finally, it presents the statistical models used to estimate impacts for each site.

Measures of Student Outcomes

Achievement

Table C.1 lists the student outcome measures that were used to evaluate FTF. Most important among these measures are those for student achievement. In each site, the most salient such measures — which are given the most attention by local officials and are used most often to guide educational decisions — are based on state tests. Each site has a state test in (1) mathematics or algebra and (2) reading or communications arts or English that focuses on a single grade in high school and a single grade in middle school.[2] The current versions of these tests are referred to as the Kansas State Assessment Test (KSAT), Texas Assessment of Knowledge and Skills (TAKS), Missouri Assessment Program (MAP), and Mississippi Subject Area Testing Program (SATP). These tests represent recent major changes from their predecessors and have high stakes for schools. Thus, considerable effort has gone into aligning local curricula to these tests and preparing students for them. Consequently, the tests provide the best indication of how well schools are doing with respect to the educational goals and objectives put forth by their states and against which their performance is judged.

In the two largest sites — Kansas City, Kansas, and Houston — an additional nationally normed test has been administered for many years that, in theory, could provide an alternative basis for measuring the effects of FTF on student achievement. In Kansas City, Kansas, the Metropolitan Achievement Test (MAT-7) in reading and mathematics has been administered districtwide since 1995. However, for the past few years, little attention has been paid to this test

[1]This information is presented as Figure 4.1 in Chapter 4.

[2]Other subjects were covered by state tests for some sites. But only their mathematics/algebra and reading/communications arts/English components are included in the present analysis because they are the most comparable across sites and they represent the central academic foci of FTF.

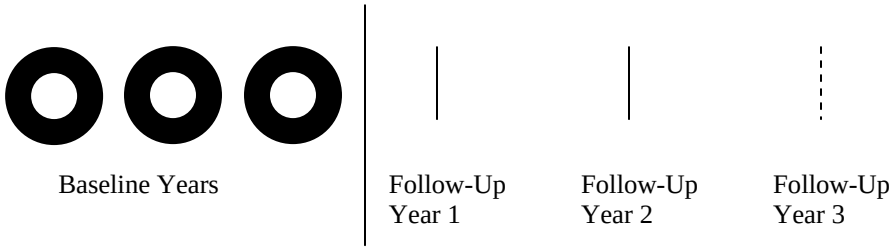
The First Things First Evaluation

Appendix Figure C.1

Design Diagrams for the Impact Analysis

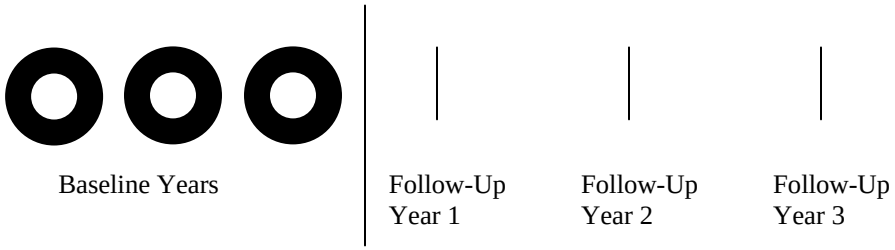
A. Houston, Texas

Student-level data, regression-adjusted for demographics and pretest (3 FTF high schools and 10 to 11 comparison schools; 4 FTF middle schools and 3 to 15 comparison schools)



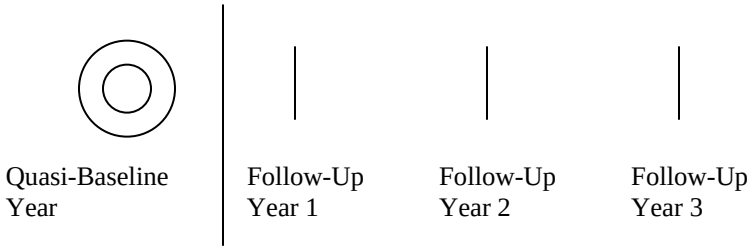
B. Riverview Gardens, Missouri

Aggregate-level data, no regression adjustments (1 FTF high school and 8 comparison schools; 1 composite FTF middle school [Central and East combined] and 12 comparison schools)



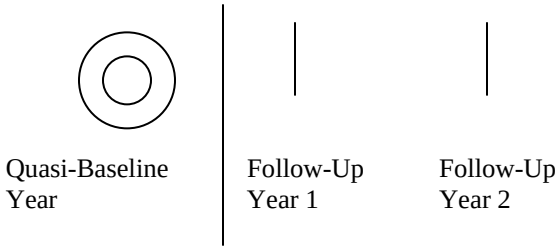
C. Kansas City, Kansas

Student-level data, regression-adjusted for demographics (4 FTF high schools and 7 comparison schools; 8 FTF middle schools and 9 comparison schools)



D. Shaw and Greenville, Mississippi

Aggregate-level data, no regression adjustments (2 FTF high schools and 4 to 10 comparison schools)



The First Things First Evaluation

Appendix Table C.1

Outcome Measures for the First Things First Impact Analysis

School District	State Test	Nationally Normed Test	Attendance Rate	Persistence Rate	Dropout Rate	Graduation Rate
Houston, Texas[a]						
	Texas Assessment of Academic Skills (TAAS)/Texas Assessment of Knowledge and Skills (TAKS): percentage passing	Stanford Achievement Test (SAT-9): percentage scoring at/below 25th percentile and percentage scoring at/above 50th percentile				
High school	10th-grade reading and math	10th-grade reading and math	school level	9th-grade		
Middle school	8th-grade reading and math	8th-grade reading and math	school level			
Riverview Gardens, Missouri[b]						
	Missouri Assessment Program (MAP): percentage in bottom 2 performance categories and percentage in top 2 performance categories					
High school	11th-grade communication arts and 10th-grade math		school level		school level	school level
Middle school	7th-grade communication arts and 8th-grade math		school level			
Kansas City, Kansas[c]						
	Kansas State Assessment Test: percentage in top 3 performance categories and percentage in bottom performance category					
High school	11th-grade reading and 10th-grade math		school level		school level	school level
Middle school	8th-grade reading and 7th-grade math		school level			
Greenville and Shaw, Mississippi						
	Subject Area Testing Program (SATP): mean score and percentage passing					
High school	10th-grade English II and 9th-grade Algebra					

(continued)

Appendix Table C.1 (continued)

NOTES:

[a]In Houston, the attendance rate is the average of the number of days a student attended school divided by the number of days the student was enrolled for each school. The persistence rate is the percentage of 9th-grade students in a given year who are still enrolled in a school in the Houston school district the following year, regardless of which school or grade they enter.

[b]In Missouri, the attendance rate is the "total hours of student attendance divided by the sum of total hours of student attendance and total hours of absence." The dropout rate is the "number of high school dropouts divided by the total of September enrollment plus transfers in, minus transfers out, minus dropouts, added to total September enrollment, then divided by two (2)." The graduation rate is the "quotient of the number of graduates in the current year as of June thirtieth divided by the sum of the number of graduates in the current year as of June thirtieth plus the number of twelfth-graders who dropped out in the current year plus the number of eleventh-graders who dropped out in the preceding year plus the number of tenth-graders who dropped out in the second preceding year plus the number of ninth-graders who dropped out in the third preceding year." Website: http://dese.mo.gov/schoollaw/rulesregs/50340200.html#2B.

[c]In Kansas City, Kansas, the attendance rate is the daily attendance divided by the average daily enrollment. Dropout and graduation rates are calculated on cohorts beginning with 9th-grade enrollment. Students who move into a school are added to the cohort, even if it is late in 12th grade. Kids who move away (and are verified) are removed from the calculations. The denominator for both rates is composed of the number of 12th-graders plus all those who dropped along the way. (More precisely, it would be the beginning cohort, plus all who moved in, minus all who moved away.) Thus, the dropout rate is the cumulative number of dropouts in a given cohort, divided by the denominator defined above. However, if dropped students enroll back in school by Oct. 4 of the following year, they are removed from the dropout stats (until or unless they drop again). The graduation rate is the number of graduates in a given cohort, divided by the denominator defined above. If students do not graduate by end of year but complete requirements during that summer, they are added to the graduates.

as the newly revised state test has come to monopolize student assessment.[3] Thus, results on this test have become increasingly irrelevant to the operation of schools in the district. In addition, the MAT-7 was not administered statewide across Kansas and thus is not available for comparison schools. Therefore, it could not be used to measure the impacts of FTF.

Since 1998, the Houston Integrated School District has administered the Stanford Achievement Test (SAT-9) in grades 1 to 11. Hence, this test provides consistently measured scores over time for FTF schools and comparison schools. The SAT-9 was instituted to provide an alternative to the current state test at the time, the Texas Assessment of Academic Skills (TAAS), which had come under heavy criticism for being "too easy." Thus, serious attention was paid to the SAT-9 throughout the Houston district for a number of years. However, with the onset of a new, more difficult, and more highly regarded state test in the past several years, the SAT-9 (most recently changed to the SAT-10) has begun to recede into the background of student assessment in Houston. Thus, while still probably a valid measure of student achievement, the relevance of the SAT-9/10 for local decision-making has declined considerably, and it might not be a very sensitive measure of progress toward the specific educational goals and standards that are guiding the district at this time.

Attendance

Attendance rates are available in three of the five districts. However, the rate is calculated a bit differently in each district. In Houston, where individual student-level data are available, attendance rates for each school are calculated for each student by dividing his or her total number of days present by the total number of days enrolled. Then the individual student rates are averaged for all students at each school. In Kansas and Missouri, data are available only at the school level. In Kansas, attendance is measured for each school as the total number of days of student attendance reported for all grades in a given year divided by the total number of days of recorded enrollment (multiplied by 100). In Missouri, attendance rates for each school are computed as the average daily number of students attending throughout the academic year divided by the total January enrollment (multiplied by 100).

[3]The MAT-7 was administered initially each spring in grades 1 though 11. This was done in response to a state mandate that all school districts in Missouri "triangulate" their measures of student performance based on three testing regimes: (1) local assessments customized to district standards and benchmarks, (2) a state test aligned with state standards and benchmarks, and (3) a nationally normed test chosen from a list of acceptable alternatives. During the past five years, the new state test took precedence over all others, and the MAT-7 was deemphasized substantially. It was shifted from a spring administration to a fall administration, in order to "get it out of the way" of the state test. It was reduced in scope to grades 5, 8, and 11. And many have argued that it is an unnecessary burden that should be eliminated. Thus, district officials, school principals, and teachers now pay little attention to the test or its results.

Persistence

Persistence rates are calculated in Houston because student-level data allow tracking of students from one year to the next. Persistence rates among ninth-graders, then, are the percentage of ninth-grade students in a given year who are enrolled in school in the Houston school district at any time during the following year (taking into account students who were recorded as having transferred to another district).

Dropouts

School dropout rates are available for schools in Kansas and Missouri.[4] In Kansas, dropout rates are defined cumulatively for each entering cohort of ninth-grade students as they proceed (or not) through high school during the next four years. Students who move into a school, even if it is late in twelfth grade, are added to the cohort. Students who move away (and are verified) are removed from the calculations. Thus, the dropout rate is calculated by taking the cumulative number of dropouts (over the course of four years) in a given cohort and dividing it by the beginning ninth-grade cohort, plus all who moved in, minus all who moved away.[5] In Missouri, the dropout rate is computed for each year for each school as the total number of dropouts recorded by the school divided by its average enrollment. Average enrollment is computed as the mean of (1) the September enrollment and (2) the September enrollment plus transfers in during the year, minus transfers out, and minus dropouts.

Graduation

Graduation rates are available for schools in Kansas and Mississippi.[6] In Kansas, graduation rates, like dropout rates, are defined for a cohort of ninth-graders. The graduation rate is calculated as the number of graduates in the cohort divided by the beginning ninth-grade cohort, plus all who moved in, minus all who moved away. Students who do not graduate by the end of the year but who complete requirements during that summer are added to the graduates. The Missouri schools' graduation rate is calculated similarly. Specifically, it is computed as the total number of graduates in the current year (as of June 30) divided by the sum of the number of graduates in the current year as of June 30 plus the number of twelfth-graders who dropped out in the current year plus the number of eleventh-graders who dropped out in the preceding year plus the number of tenth-graders who dropped out in the second preceding year plus the number of ninth-graders who dropped out in the third preceding year.

---

[4]Dropout data are not reliable in Houston and, therefore, were not used.

[5]If dropped students reenroll in school by October 4 of the following year, they are removed from the numerator of the dropout statistic (until or unless they drop out again).

[6]Graduation rates were not calculated for Houston because reliable dropout data are not available, and dropouts are used in calculating graduation rates.

Process and Criteria for Selecting Comparison Schools

Table C.2 summarizes the process and criteria used to select comparison schools for each site in the present evaluation. The first step in the process (for all sites but Houston) was to define a pool of candidate schools based on their enrollment size and racial/ethnic composition during the baseline or quasi-baseline period. Once this pool was identified, those schools with baseline test scores that were markedly different from the FTF schools were excluded. At this point, additional information on such factors as student mobility or receipt of free/reduced-price lunch was used to exclude schools that differed substantially from the FTF schools. Lastly, data on enrollment and racial/ethnic composition for the entire analysis period (baseline or quasi-baseline years plus follow-up years) were examined to exclude schools that experienced dramatic changes that might signify structural shifts such as redistricting.

Before the impacts of FTF were estimated, the selection of comparison schools was conducted, assessed, and revised (when necessary).

Houston

The Houston Independent School District implemented FTF in three high schools and four middle schools, with one high school and one middle school launching the intervention a year before the others. A separate group of comparison schools was created for each FTF school, although in most cases these groups overlapped substantially. Because the FTF schools in Houston are heavily Hispanic and very low-performing, whereas most other low-performing schools are heavily African-American, it was not possible to match closely on race/ethnicity *and* performance. Faced with this tradeoff, precedence was given to student performance, because the best predictor of future performance is past performance.

An important complication arose for one FTF high school that was redistricted during the follow-up period, so that many students who would have attended the school under its earlier configuration subsequently went elsewhere. Before selecting a comparison group for this school, it was necessary to identify a student catchment area that was defined consistently over time. This eliminated shifts in the composition of the student population that were produced by redistricting. To make this adjustment required obtaining geocoded information about the past and present boundaries of the school's district plus data on students' addresses. With this information, it was possible to omit students from earlier years who lived in the part of the catchment area that was removed in later years by redistricting.

The comparison group for each FTF school comprised all comprehensive and magnet non-FTF schools whose mean baseline scores on the SAT-9 for reading and math combined were within 0.25 standard deviation of the mean for the FTF school. Candidate schools meeting

The First Things First Evaluation

Appendix Table C.2

Selection of Comparison Schools

Houston

- Selected for each FTF school from within the district
- Within 0.25 standard deviation on mean total baseline SAT-9 score in reading and math
- Did not experience a radical change in enrollment or student composition during baseline and follow-up periods

Riverview Gardens

High schools

- Selected from other districts in the state
- Mean enrollment of 500 or more students during baseline period
- At least 60 percent African-American students, on average, during baseline period
- Mean scores on state tests in both math and communications arts in bottom two (out of five) performance categories during baseline period
- Did not experience unusually high student mobility or radical change in enrollment or student composition during baseline and follow-up periods

Middle schools

- Selected from other districts in the state
- At least 70 percent African-American students during baseline period
- Did not experience unusually high student mobility or radical change in enrollment or student composition during baseline and follow-up periods

Kansas City, Kansas

High schools

- Selected from nonrural parts of other metropolitan and micropolitan areas in the state
- Enrollment of 500 or more students during quasi-baseline year
- At least 35 percent minority students during quasi-baseline year
- Mean percentage of students scoring in top three (out of five) categories on state tests in reading and math no more than 50 percent
- Did not experience a radical change in enrollment or student composition during quasi-baseline and follow-up periods

Middle schools

- Selected from nonrural parts of other metropolitan and micropolitan areas in the state
- Enrollment of 350 or more students during quasi-baseline year
- At least 50 percent minority students during quasi-baseline year
- Mean percentage of students scoring in top three (out of five) categories on state tests in reading and math was 50 percent or less
- Did not experience a radical change in enrollment or student composition during quasi-baseline and follow-up periods

Mississippi Delta

- Selected from the Mississippi Delta for School J plus other large schools in the state (with enrollment of 900 or more students in the quasi-baseline year) for School I
- At least 85 percent African-American during quasi-baseline year
- At least 80 percent of students receiving free/reduced-price lunch during quasi-baseline year

this criterion were checked to see whether they had experienced a dramatic shift in their enrollment or student composition. Those that did were omitted from the analysis. This produced comparison groups for each FTF school ranging from ten to eleven high schools and from three to fifteen middle schools.

A further step was taken for the high schools. For these schools, a second comparison group was constructed for the analysis of impacts on state test scores (the TAAS/TAKS). This was done because matching on SAT-9 baseline scores produced comparison groups whose baseline scores on the state test were very different from those for the FTF schools. These alternative comparison groups were created by selecting all candidate schools whose average baseline scores on the state test were within 0.25 standard deviation of the FTF school.[7] Alternative estimates of impacts on the state test were obtained using each comparison group.

Riverview Gardens

Riverview Gardens is an independent school district located on the urban fringe of St. Louis, Missouri, with one high school and two middle schools. During its baseline period, from 1998-1999 to 2000-2001, Riverview Gardens High School had an average enrollment of roughly 1,600 students in grades 9 through 12. About 88 percent of these students were African-American; 44 percent received subsidized lunches; and their annual mobility rate was 31 percent. Aggregate data from the Missouri Department of Elementary and Secondary Education were used to choose comparison schools.[8]

Given the large size of the FTF high school and its predominantly African-American student population, comparison schools in Missouri were chosen from among all public high schools that had grades 9 through 12 and that had a mean enrollment of 500 or more students from 1998-1999 to 1999-2000 — at least 60 percent of whom were African-American. These criteria identified a pool of eleven candidate schools. Student scores on the state's tests in mathematics and communication arts were then used to narrow the pool. Missouri classifies its state test scores from lowest to highest in five levels of performance (step 1, progressing, nearing proficiency, proficient, and advanced). The baseline average math score for Riverview Gardens High School was in the lowest performance level, and its average communication arts score was in the second-lowest performance level. Candidate comparison schools therefore remained in contention if their average baseline scores for both subjects were in either of the bottom two performance levels.

[7]For one high school, a criterion of 0.25 standard deviation produced only one comparison school, so the criterion for that school was expanded to 0.33 standard deviation.

[8]Web site: http://www.dese.state.mo.us/schooldata/ftpdata.html.

Two of the remaining candidate schools were eliminated because their annual rates of student mobility were over 200 percent. Lastly, data for the baseline and follow-up years were examined to identify any dramatic enrollment or demographic shifts that occurred, which might signify redistricting. The result of this process was a high school comparison group that included eight schools from St. Louis City, the urban fringe of St. Louis, and Kansas City, Missouri.

The two FTF middle schools in Riverview Gardens were created during the 2000-2001 school year from a single middle school that had existed for many years. In the analysis, these schools are treated as a single "composite" school to maintain a student catchment area (and thus a student population) that was defined consistently over time. Aggregate data from the state were used to select comparison schools. However, because enrollment is not a meaningful consideration for a composite school, it was not used as a criterion for selecting comparison schools. Thus, the primary criterion was whether candidate schools had 70 percent or more African-American students during the baseline period. From the pool of schools identified using this criterion, several were dropped because they had annual rates of student mobility over 70 percent, which was more than twice the rate for Riverview Gardens. This yielded seventeen candidate schools, two of which were subsequently eliminated because their 2001 state test scores were not available from the data source used. Lastly, data for the baseline and follow-up periods were examined to identify dramatic shifts in student enrollment or racial/ethnic composition that might have occurred. This eliminated three more schools, which left twelve schools in the middle school comparison group.

Kansas City, Kansas

Kansas City, Kansas, has four FTF high schools and eight FTF middle schools. These schools have a much higher percentage of minority students and substantially lower scores on state tests than do schools in other parts of the state. Thus, it was very difficult to create a comparison group of closely matched schools. Because of this, no attempt was made to match individual FTF schools. Instead, a single comparison group was constructed for all FTF high schools, and another comparison group was constructed for all FTF middle schools. To identify each comparison group, the following criteria were applied to all high schools or middle schools in metropolitan and micropolitan areas in Kansas, based on their individual student data for the quasi-baseline year 2000-2001.

Selection criteria for high schools were:

1. Total enrollment of 500 students or more[9]

[9]Total enrollment was estimated by doubling the sum of the number of students tested in each grade, because the test was administered only to two of four grades.

2. Minority student population of 35 percent or more[10]

3. Average math and reading proficiency levels of 50 percent or less on the state test[11]

4. Not located in a rural area

Selection criteria for middle schools were:

1. Total enrollment of 350 students or more

2. Minority student population of 50 percent or more

3. Average math and reading proficiency levels of 50 percent or less on the state test

4. Not located in a rural area

From the pool of candidate schools that met these criteria, those that experienced dramatic shifts in size or racial/ethnic composition during the analysis period were dropped. The resulting comparison groups contained seven high schools and nine middle schools.

Delta Region of Mississippi

There are two FTF high schools and no FTF middle schools from the Delta Region of Mississippi. The two high schools are from different school districts and vary dramatically in size. Thus, a separate comparison group was constructed for each.

Comparison schools were chosen based on their demographic characteristics and test scores for the 2001-2002 school year, which is the first year of available data for the revised Mississippi state test and, thus, is the quasi-baseline year for the analysis. Selection criteria were as follows:

1. Schools were located in the Delta Region.

2. Student enrollment was 85 percent or more African-American.

3. At least 80 percent of students received free/reduced-price lunches.

[10]Percentage minority was calculated by summing the number of students tested who were black or Hispanic and dividing by the total number of students tested.

[11]Overall proficiency was computed by averaging the percentage proficient in math and the percentage proficient in reading. Proficiency was defined as scoring in one of the state's top three (out of five) performance levels.

Because these criteria identified very few schools that were comparable in size to the larger FTF school, the pool was broadened to include all high schools in the state with enrollments over 900 students. The combined selection process identified a total of 25 candidate comparison schools, which were split into two subgroups: schools with fewer than 500 students (for the smaller FTF high school) and schools with 500 students or more (for the larger FTF school). State test scores for algebra and English II were then examined to eliminate candidate schools with exceptionally high or low scores ("outliers"). Seven schools from the two subgroups were dropped at this stage.

The smaller high school had a very large difference between its performance in algebra and English II on the state test. About 94 percent of its students passed algebra during the quasi-baseline year, whereas only 47 percent passed English II. This made it impossible to find comparison schools that were similar in terms of performance in both subjects. Therefore, an alternative comparison group was created for estimating impacts on algebra performance. The final comparison group for the larger FTF school contained ten schools, and the two comparison groups for the smaller FTF school contained four schools and six schools.

Statistical Models for Estimating Impacts

This section describes — site by site — the statistical models used to estimate the impacts of FTF on student outcomes. Because the central outcome of interest is student achievement, as measured by performance on standardized tests, the appendix focuses on the statistical models used to estimate impacts on those outcomes. Similar models were used to estimate impacts on other outcomes.

As described in Chapter 4, the basic logic of the impact analysis is the same for all sites. The impact of FTF on a student outcome for a given follow-up year is estimated as the *difference* between the FTF school and its comparison group in their changes from the baseline (or quasi-baseline) period to the follow-up year. The models used to estimate this "difference in changes" for each site are as follows.

Houston

Figure C.1 illustrates that, in Houston, there are three FTF high schools and four FTF middle schools, with individual student outcome and background data for three baseline years and two or three follow-up years plus corresponding data for comparison schools. The Houston impact analysis was based mainly on state test scores and SAT-9 scores in reading and math for tenth-graders (for high school) and eighth-graders (for middle school). Data on individual students' background characteristics were used to control for possible compositional shifts over time in the student population. These characteristics include each student's test scores three years ear-

lier (a pretest) plus a series of demographic (binary) indicators: male, African-American, Hispanic, overage for grade, and special education status. Missing data on background characteristics were imputed so that students were not omitted if these data were missing.

Equations C.1 and C.2 represent the two-level hierarchical model used to estimate the impacts of FTF on student test scores for a single FTF school in Houston. These estimates are based on student outcome data for a given test in a given subject for a given grade (for example, the tenth-grade state reading scores) during the baseline and follow-up years for the FTF school and its comparison schools plus data on student background characteristics. Note that the subscripts k and K in the model, which serve to distinguish different schools in the sample from each other, are not related to the letters used in the text of this report to identify specific schools.

Level 1: Students

$$Y_{ijk} = \Pi_{0\,jk} + \sum_{M} \Pi_{M} X_{Mijk} + \varepsilon_{ijk} \qquad \text{(C.1)}$$

Level 2: Cohorts

$$\Pi_{0\,jk} = \sum_{K} \beta_{K} S_{Kjk} + \sum_{N} \delta_{N} F_{Njk} + \sum_{N} \gamma_{N} F_{Njk} P_{jk} + e_{jk} \qquad \text{(C.2)}$$

Where

Y_{ijk} = the outcome for student i in cohort j from school k

X_{Mijk} = background characteristic M for student i in cohort j from school k

ε_{ijk} = a random error term for student i in cohort j from school k

S_{Kjk} = one for all cohorts from school K and zero for all others

F_{Njk} = one for all cohorts representing follow-up year N and zero for all others

P_{jk} = one for all cohorts from the FTF school and zero for all others

Level 1 of the model specifies that the outcome, Y_{ijk}, for a given student from a given school in a given year depends on his or her background characteristics, X_{Mijk}, plus a random error, ε_{ijk}, which is independently and identically distributed. For simplicity, the relationship between each background characteristic and the student outcome (coefficient, Π_M) is assumed not to vary across time or schools.

Level 2 of the model specifies that the "regression-adjusted" mean outcome, Π_{0jk}, for the cohort of students from a particular school in a particular year depends on the school in-

volved, S_{Kjk}, whether the cohort is for a given follow-up year, F_{Njk}, whether the cohort is from the FTF school, and a random error, e_{jk}, which is independently and identically distributed.

The coefficient, B_k, in this model represents the regression-adjusted mean outcome for the baseline period at school k; the coefficient, δ_N, represents the mean change from the baseline mean for comparison schools in follow-up year N, and the coefficient, γ_N, represents the difference between the changes for the FTF school and those for its comparison schools in follow-up year N. These latter coefficients are estimates of the impacts of FTF on the student outcome.

For discrete outcome measures — such as whether students scored in a given performance category — a zero/one indicator variable was used for Y_{ijk} and a multilevel linear probability model (a regression model with a zero/one dependent variable) was used to estimate the impacts of FTF on the percentage of students who scored in the indicated category. More complex models, based on multilevel logistic regression, were deemed not necessary because almost all the student outcomes involved were, on average, far from extreme values (they were not near zero or one). In addition, all the independent variables in the model were simple categorical indicators. Hence, their coefficients in a linear probability model represent straightforward conditional differences in proportions. For both these reasons, the results from linear probability models (which were used) are almost identical to those from more complex logistic regression models (which were not used). To confirm this conclusion, the results from the two types of models were compared for a few student outcomes.

Impact findings obtained by estimating Equations C.1 and C.2 for each FTF school and its comparison schools were averaged across all FTF high schools and all FTF middle schools in Houston, to summarize their findings. The standard error of the average impact estimator was computed for a "fixed-effect" because this average was used to infer findings for the specific group of FTF schools, not to generalize findings to a broader population of schools. Because the impact estimator for each FTF school was independent of those for the others, the variance of the mean impact estimator was computed as the mean of the variances of the impact estimators for individual FTF schools divided by the number of schools involved. The standard error of the mean impact estimator was then obtained as the square root of its variance.[12]

Riverview Gardens

Figure C.1 indicates that, in Riverview Gardens, there is one FTF high school and one composite FTF middle school, with aggregate school-level data on student outcomes for three baseline years and three follow-up years plus corresponding data for comparison schools. Be-

[12]The standard error of the average impact estimator was adjusted for the fact that some comparison schools were used to estimate impacts for more than one FTF school.

cause individual student data are not available, it was not possible to control statistically for compositional shifts in the student population. However, an examination of aggregate student characteristics confirmed that such shifts did not occur.

The following multiple regression model was used to estimate impacts on student outcomes separately for the FTF high school and the composite FTF middle school.

$$\bar{Y}_{jk} = \sum_{K} \beta_K S_{Kjk} + \sum_{N} \delta_N F_{Njk} + \sum_{N} \gamma_N F_{Njk} P_{jk} + e_{jk} \qquad \text{(C.3)}$$

Where

$\bar{Y}_{jk}$ = the mean value of the outcome for student cohort j from school k

S_{Kjk} = one for all cohorts from school K and zero for all others

F_{Njk} = one for all cohorts representing follow-up year N and zero for all others

P_{jk} = one for all cohorts from the FTF school and zero for all others

The independent variables in the model are defined in the same way as those in Level 2 of the model for Houston. Hence, the Riverview Gardens model is the same as that for the second level of the Houston model, except that its dependent variable is the observed unconditional mean score, $\bar{Y}_{jk}$, for each cohort instead of its inferred conditional mean score, Π_{jk}. Consequently, the only effective difference between the Riverview Gardens model, which uses aggregate data for schools, and the Houston model, which uses individual data for students, is that the former does not control statistically for shifts over time in observed student characteristics. If no large shifts occur, which was the case for Riverview Gardens, there is no appreciable difference in the two models' abilities to produce valid and reliable impact estimates.

Kansas City, Kansas

Figure C.1 indicates that there are four FTF high schools and eight FTF middle schools in Kansas City, Kansas, with a single set of comparison high schools and a single set of comparison middle schools. The main impact analysis for the site was based on students' performance on the newly revised state test in reading and math. Because this test was implemented for the first time in 2001, which is after FTF had begun at all FTF schools at the site, 2001 was used as a quasi-baseline year, and subsequent years make up the follow-up period. Individual data on student outcomes plus the following background characteristics were used for the analysis: limited English proficiency, socioeconomic status, minority, and gender. These data provided input to the following two-level hierarchical model for estimating impacts.

Level 1: Students

$$Y_{ijk} = \Pi_{0jk} + \sum_{M} \Pi_{M} X_{Mijk} + \varepsilon_{ijk} \quad \text{(C.4)}$$

Level 2: Cohorts

$$\Pi_{0jk} = \sum_{K} \beta_{K} S_{Kjk} + \sum_{N} \delta_{N} F_{Njk} + \sum_{N} \gamma_{N} F_{Njk} P_{jk} + e_{jk} \quad \text{(C.5)}$$

Where

Y_{ijk} = the outcome for student i in cohort j from school k

X_{Mijk} = background characteristic M for student i in cohort j from school k

ε_{ijk} = a random error term for student i in cohort j from school k

S_{Kjk} = one for all cohorts from school K and zero for all others

F_{Njk} = one for all cohorts representing follow-up year N and zero for all others

P_{jk} = one for all cohorts from FTF schools and zero for all others

This model has the same structure as the one used for Houston. However, there are three differences in its application to the two sites. In Kansas City, Kansas, impacts were estimated for all schools together using one comparison group, whereas, in Houston, impacts were estimated for each school separately using different but overlapping comparison groups. Hence, no separate step was required in Kansas City, Kansas, to obtain the average impact for FTF schools. It was obtained directly from estimates of γ_N in the model. Second, it was not possible to use data on pretests for individual students in Kansas City, Kansas, because local confidentiality requirements precluded access to the individual identifiers needed to link students' test scores over time. Thus, only student demographics were used as background characteristics. Third, impact estimates for Kansas City, Kansas, were based on schools' performance relative to a single post-intervention quasi-baseline year, whereas those for Houston were relative to schools' performance during a multiyear pre-intervention baseline period.

As described in Chapter 4, the impacts of FTF on outcomes for Kansas City, Kansas — other than state test scores (rates of student attendance, dropout, and graduation) — were based on a somewhat different evaluation design, with corresponding differences in its statistical model. Because consistent data on these outcomes are available for a period of time that was longer than the period for the new state test, it was possible to define an earlier point of reference, or benchmark, for measuring change over time. For these analyses, then, the average out-

come for 1997-1998, 1998-1999, and 1999-2000 (which were pre-intervention years for some schools and both pre- and post-intervention years for others) was used as a benchmark, and subsequent years make up the follow-up period. Data for these outcomes are available only at the aggregate school level. Thus, a multiple regression model like Equation C.3 was used to estimate impacts for these outcomes.

The Delta Region of Mississippi

Figure C.1 indicates that there are two FTF high schools and no FTF middle schools from the Mississippi Delta, with aggregate school-level data for a quasi-baseline year and two follow-up years. Given the substantial differences between the two FTF high schools, a separate comparison group was constructed for each.

The sole outcome measures available for the impact analysis at this site are student scores on state tests in algebra and English II. These tests were substantially revised in 2002, after FTF was launched. Hence, 2002 is used as a quasi-baseline year, and the next two years are used as a follow-up period. From aggregate school-level data on student outcomes, the impacts of FTF were estimated using the following regression model:

$$\bar{Y}_{jk} = \sum_{K} \beta_K S_{Kjk} + \sum_{N} \delta_N F_{Njk} + \sum_{N} \gamma_N F_{Njk} P_{jk} + e_{jk} \qquad \text{(C.6)}$$

Where

$\bar{Y}_{jk}$ = the mean value of the outcome for student cohort j from school k

S_{Kjk} = one for all cohorts from school K and zero for all others

F_{Njk} = one for all cohorts representing follow-up year N and zero for all others

P_{jk} = one for all cohorts from the FTF school and zero for all others

This model is the same as that used for Riverview Gardens, although its application differs in that the Mississippi Delta has a single post-intervention quasi-baseline year, whereas Riverview Gardens has three pre-intervention baseline years.

Appendix D

Supplementary Tables for Chapter 4

The First Things First Evaluation

Appendix Table D.1

Estimated Impact of First Things First on 11th-Grade State Reading Tests: Kansas City, Kansas

	Quasi-Baseline Level (2001)		Estimated Impact by Follow-Up Year[b]		
	FTF[a]	Comparison	Spring 2002	Spring 2003	Spring 2004
			Impact on Percentage Proficient[c]		
All First Things First high schools	**17.4**	**40.6**	**6.9**	**10.2 ****	**11.1 ****
School A	19.6	40.6	7.2	17.7 **	5.7
School B	12.9	40.6	9.5	7.1	17.0 **
School C	20.8	40.6	8.0	13.6 *	20.7 **
School D	26.1	40.6	-3.2	-3.0	-4.1
			Impact on Percentage Unsatisfactory[d]		
All First Things First high schools	**50.0**	**25.8**	**-5.4**	**-11.1 ****	**-15.5 *****
School A	55.5	25.8	-13.5 **	-25.8 ***	-16.9 **
School B	55.2	25.8	-7.7	-11.7 *	-17.5 **
School C	39.5	25.8	-0.8	-5.9	-22.4 ***
School D	42.4	25.8	5.4	3.8	-1.6

(continued)

SOURCE: MDRC calculations from individual student records from a statewide data file.

NOTES: Sample includes 11th-grade students from four First Things First (FTF) high schools and seven comparison schools. Students in the sample consist of test-takers for whom administrative records exist between the 2000-2001 and 2003-2004 academic years.

The "impact" was calculated as the difference between the deviation from the quasi-baseline for FTF schools and the deviation from the quasi-baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics.

Appendix Table D.1 (continued)

A two-tailed t-test was applied to differences in deviations from quasi-baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The quasi-baseline year is the third year of implementation for School A, the second year of implementation for School B, and the first year of implementation for Schools C and D.

[b]For impacts at the school level, statistical significance is presented; however, these results are much less reliable than pooled impacts.

[c]The desired impact for this measure is positive.

[d]The desired impact for this measure is negative.

The First Things First Evaluation

Appendix Table D.2

Estimated Impact of First Things First on 10th-Grade State Math Tests: Kansas City, Kansas

	Quasi-Baseline Level (2001)		Estimated Impact by Follow-Up Year[b]		
	FTF[a]	Comparison	Spring 2002	Spring 2003	Spring 2004
			Impact on Percentage Proficient[c]		
All First Things First high schools	**6.4**	**22.4**	**1.2**	**3.4**	**-4.4 ***
School A	10.1	22.4	-0.1	1.2	-4.5
School B	3.9	22.4	4.2	2.7	-3.7
School C	8.7	22.4	3.7	11.7 **	-1.9
School D	9.0	22.4	-3.9	-2.3	-9.3 **
			Impact on Percentage Unsatisfactory[d]		
All First Things First high schools	**72.6**	**45.4**	**-10.8 *** **	**-6.7 ** **	**-5.2**
School A	71.0	45.4	-8.5	-4.4	-8.8
School B	72.2	45.4	-12.0 **	-0.8	-5.5
School C	66.9	45.4	-13.2 **	-19.1 ***	-3.0
School D	74.7	45.4	-8.8	-2.3	-3.3

(continued)

SOURCE: MDRC calculations from individual student records from a statewide data file.

NOTES: Sample includes 10th-grade students from four First Things First (FTF) high schools and seven comparison schools. Students in the sample consist of test-takers for whom administrative records exist between the 2000-2001 and 2003-2004 academic years.

The "impact" was calculated as the difference between the deviation from the quasi-baseline for FTF schools and the deviation from the quasi-baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics.

Appendix Table D.2 (continued)

A two-tailed t-test was applied to differences in deviations from quasi-baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The quasi-baseline year is the third year of implementation for School A, the second year of implementation for School B, and the first year of implementation for Schools C and D.

[b]For impacts at the school level, statistical significance is presented; however, these results are much less reliable than pooled impacts.

[c]The desired impact for this measure is positive.

[d]The desired impact for this measure is negative.

The First Things First Evaluation

Appendix Table D.3

Estimated Impact of First Things First on High School Attendance, Dropout, and Graduation Rates: Kansas City, Kansas

	Quasi-Baseline Level		Estimated Impact by Follow-Up Year[b]			
	FTF[a]	Comparison	2000-2001	2001-2002	2002-2003	2003-2004
			Impact on Attendance Rate[c] (%)			
All First Things First high schools	**80.9**	**87.5**	**1.6**	**8.6 *** **	**6.3 *** **	**2.0**
School A	74.7	87.5	1.2	11.7 ***	10.7 ***	7.1 **
School B	84.9	87.5	2.1	6.5 **	4.5	0.2
School C	85.3	87.5	-1.7	6.2 **	2.4	-2.5
School D	78.9	87.5	5.0	10.0 ***	7.7 **	3.3
			Impact on Dropout Rate[d] (%)			
All First Things First high schools	**10.6**	**7.8**	**-2.6**	**-6.3 *** **	**-4.0 * **	**-4.3 * **
School A	7.2	7.8	0.4	-2.6	-1.3	0.0
School B	10.9	7.8	-7.5 ***	-6.6 **	-2.7	-2.0
School C	17.9	7.8	-0.1	-11.0 ***	-9.9 ***	-14.2 ***
School D	6.3	7.8	-3.1	-4.9 *	-2.0	-0.9
			Impact on Graduation Rate[e] (%)			
All First Things First high schools	**54.9**	**68.6**	**10.6 ** **	**12.3 ** **	**14.9 *** **	**15.7 *** **
School A	44.0	68.6	29.0 ***	15.4 **	26.1 ***	7.5
School B	49.5	68.6	20.7 ***	28.8 ***	31.8 ***	26.1 ***
School C	48.8	68.6	-6.7	1.7	-5.0	20.6 ***
School D	77.2	68.6	-0.5	3.1	6.5	8.6

(continued)

SOURCE: MDRC calculations from school-level records of state data.

Appendix Table D.3 (continued)

NOTES: Sample includes four First Things First (FTF) high schools and seven comparison schools.

The "impact" was calculated as the difference between the deviation from the quasi-baseline for FTF schools and the deviation from the quasi-baseline for comparison schools.

A two-tailed t-test was applied to differences in deviations from quasi-baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The quasi-baseline years (three prior school years) are pre- and post-implementation years for Schools A and B, and the pre-implementation years for Schools C and D.

[b]For impacts at the school level, statistical significance is presented; however, these results are much less reliable than pooled impacts.

[c]The desired impact for this measure is positive.

[d]The desired impact for this measure is negative.

[e]The desired impact for this measure is positive.

The First Things First Evaluation

Appendix Table D.4

Estimated Impact of First Things First on 8th-Grade State Reading Tests: Kansas City, Kansas

	Quasi-Baseline Level (2001)		Estimated Impact by Follow-Up Year[b]		
	FTF[a]	Comparison	Spring 2002	Spring 2003	Spring 2004
			Impact on Percentage Proficient[c]		
All First Things First middle schools	**27.6**	**43.7**	**3.0**	**23.1 *** **	**13.7 *** **
School K	29.5	43.7	3.0	38.1 ***	28.3 ***
School L	22.2	43.7	7.0	28.1 **	6.9
School M	28.1	43.7	10.9	14.7	8.3
School N	26.2	43.7	6.6	28.0 **	4.6
School O	31.9	43.7	-3.3	11.9	31.3 ***
School P	30.5	43.7	1.1	17.5	3.7
School Q	38.0	43.7	-15.6	14.2	8.7
School R	29.1	43.7	6.9	21.2 **	6.3
			Impact on Percentage Unsatisfactory[d]		
All First Things First middle schools	**38.3**	**25.5**	**-5.4**	**-22.3 *** **	**-13.6 *** **
School K	36.5	25.5	0.4	-32.5 ***	-20.8 **
School L	59.6	25.5	-25.5 **	-38.4 ***	-28.4 **
School M	35.1	25.5	-6.3	-12.5	-5.8
School N	37.4	25.5	-3.4	-27.1 **	-5.2
School O	35.1	25.5	-1.5	-13.0	-18.7 *
School P	29.2	25.5	-2.0	-18.1 *	-3.0
School Q	29.3	25.5	7.4	-15.6	-12.0
School R	35.1	25.5	-9.5	-15.6	-8.2

(continued)

SOURCE: MDRC calculations from individual student records from a statewide data file.

Appendix Table D.4 (continued)

NOTES: Sample includes 8th-grade students from eight First Things First (FTF) middle schools and nine comparison schools. Students in the sample consist of test-takers for whom administrative records exist between the 2000-2001 and 2003-2004 academic years.

The "impact" was calculated as the difference between the deviation from the quasi-baseline for FTF schools and the deviation from the quasi-baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics.

A two-tailed t-test was applied to differences in deviations from quasi-baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The quasi-baseline year is the third year of implementation for Schools K and L, the second year of implementation for Schools M and N, and the first year of implementation for Schools O, P, Q, and R.

[b]For impacts at the school level, statistical significance is presented; however, these results are much less reliable than pooled impacts.

[c]The desired impact for this measure is positive.

[d]The desired impact for this measure is negative.

The First Things First Evaluation

Appendix Table D.5

Estimated Impact of First Things First on 7th-Grade State Math Tests: Kansas City, Kansas

	Quasi-Baseline Level (2001)		Estimated Impact by Follow-Up Year[b]		
	FTF[a]	Comparison	Spring 2002	Spring 2003	Spring 2004
			Impact on Percentage Proficient[c]		
All First Things First middle schools	**14.4**	**29.5**	**5.0**	**11.0 *** **	**9.6 ** **
School K	15.7	29.5	8.2	12.6 *	9.7
School L	7.0	29.5	6.5	23.9 ***	8.7
School M	30.4	29.5	-5.2	-2.8	-7.2
School N	22.8	29.5	-4.8	3.5	9.5
School O	16.5	29.5	6.5	6.6	17.6 **
School P	13.5	29.5	6.5	17.8 **	4.8
School Q	9.8	29.5	16.1 **	21.0 ***	13.3 *
School R	5.9	29.5	7.1	7.2	21.3 ***
			Impact on Percentage Unsatisfactory[d]		
All First Things First middle schools	**62.1**	**43.6**	**-7.3 ***	**-13.1 *** **	**-9.0 ** **
School K	66.8	43.6	-18.0 **	-23.8 ***	-19.2 **
School L	74.4	43.6	-9.6	-25.0 ***	-10.1
School M	41.3	43.6	4.8	6.2	11.3
School N	54.0	43.6	0.9	-5.1	-5.8
School O	59.7	43.6	-9.3	-10.7	-16.8 **
School P	60.9	43.6	2.2	-5.8	1.7
School Q	66.8	43.6	-17.4 **	-25.4 ***	-9.4
School R	72.5	43.6	-11.8	-14.8 *	-23.1 ***

(continued)

SOURCE: MDRC calculations from individual student records from a statewide data file.

Appendix Table D.5 (continued)

NOTES: Sample includes 7th-grade students from eight First Things First (FTF) middle schools and nine comparison schools. Students in the sample consist of test-takers for whom administrative records exist between the 2000-2001 and 2003-2004 academic years.

The "impact" was calculated as the difference between the deviation from the quasi-baseline for FTF schools and the deviation from the quasi-baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics.

A two-tailed t-test was applied to differences in deviations from quasi-baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The quasi-baseline year is the third year of implementation for Schools K and L, the second year of implementation for Schools M and N, and the first year of implementation for Schools O, P, Q, and R.

[b]For impacts at the school level, statistical significance is presented; however, these results are much less reliable than pooled impacts.

[c]The desired impact for this measure is positive.

[d]The desired impact for this measure is negative.

The First Things First Evaluation

Appendix Table D.6

Estimated Impact of First Things First on Middle School Attendance Rates: Kansas City, Kansas

	Quasi-Baseline Level		Estimated Impact by Follow-Up Year[b]			
	FTF[a]	Comparison	2000-2001	2001-2002	2002-2003	2003-2004
			Impact on Attendance Rate[c] (%)			
All First Things First middle schools	**90.5**	**90.1**	**-1.0**	**2.4 ****	**3.3 *****	**1.9 ***
School K	90.4	90.1	-3.0	3.2	4.6 **	3.6
School L	89.7	90.1	-2.6	3.5	3.3	1.0
School M	92.0	90.1	-0.7	0.3	1.6	0.8
School N	91.2	90.1	-0.3	0.8	1.4	0.4
School O	89.7	90.1	4.0 **	5.2 **	5.4 ***	5.5 ***
School P	90.3	90.1	-2.9	-0.3	1.1	0.7
School Q	90.4	90.1	0.1	2.1	3.0	1.5
School R	90.1	90.1	-2.5	4.3 **	5.9 ***	2.0

SOURCE: MDRC calculations from school-level records of state data.

NOTES: Sample includes eight First Things First (FTF) middle schools and nine comparison schools.

The "impact" was calculated as the difference between the deviation from the quasi-baseline for FTF schools and the deviation from the quasi-baseline for comparison schools.

A two-tailed t-test was applied to differences in deviations from quasi-baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The quasi-baseline years (three prior school years) are pre- and post-implementation years for Schools K, L, M and N, and the pre-implementation years for Schools O, P, Q, and R.

[b]For impacts at the school level, statistical significance is presented; however, these results are much less reliable than pooled impacts.

[c]The desired impact for this measure is positive.

The First Things First Evaluation

Appendix Table D.7

Estimated Impact of First Things First on the Percentage of 10th-Grade Students Passing the TAAS/TAKS in Reading and Math: Houston, Texas

	Baseline Level		Estimated Impact by Follow-Up Year[b]		
	FTF[a]	Comparison	Year 1	Year 2	Year 3
			Impact on (TAAS/TAKS) Reading Pass Rates[d] (%)		
All First Things First high schools[c]	**70.9**	**75.8**	**-1.1**	**6.6**	
School E	59.5	72.6	4.9	12.5 ***	8.8 *
School F	77.2	78.1	-7.3	0.4	
School G	76.1	76.6	-0.9	6.8	
			Impact on (TAAS/TAKS) Math Pass Rates[d] (%)		
All First Things First high schools[c]	**75.4**	**74.5**	**-3.3**	**4.2**	
School E	66.1	71.9	-3.9	9.6	7.0
School F	83.7	76.2	-5.7	3.7	
School G	76.4	75.5	-0.4	-0.7	

(continued)

SOURCE: MDRC calculations from individual student records from the Houston Independent School District data file.

NOTES: Sample includes 10th-grade students from three clusters. Each cluster consists of a First Things First (FTF) high school matched with a group of between five and seven non-FTF schools. The sample consists of students for whom administrative records exist between the 1998-1999 and 2003-2004 academic years.

The "impact" was calculated as the difference between the deviation from the baseline for FTF schools and the deviation from the baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics and prior achievement.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

Appendix Table D.7 (continued)

[a]The baseline level is the average of three pre-implementation years, which are not the same calendar years for each school. For school E, baseline includes the 1998-1999, 1999-2000, and 2000-2001 school years. For the other schools, baseline includes the 1999-2000, 2000-2001, and 2001-2002 school years. This is why one school has three follow-up years and the others have two.

[b]For impacts at the school level, statistical significance is presented; however, these results are much less reliable than pooled impacts.

[c]Follow-up Years 1 and 2 in the "All First Things First high schools" row average together the first and second post-implementation years for each school, which are not the same calendar years for each school.

[d]The desired impact for this measure is positive.

Appendix Table D.8

Estimated Impact of First Things First on the Percentage of 10th-Graders Scoring At/Above the 50th Percentile and At/Below the 25th Percentile on the SAT-9 in Reading: Houston, Texas

	Baseline Level		Estimated Impact by Follow-Up Year[b]		
	FTF[a]	Comparison	Year 1	Year 2	Year 3
			Impact on SAT-9 (%): At/Above the 50th Percentile[d]		
All First Things First high schools[c]	**18.3**	**19.5**	**0.2**	**-5.2**	
School E	16.1	20.7	2.9	0.2	0.2
School F	14.6	17.9	-2.8	-7.7	
School G	24.2	20.0	0.4	-8.2	
			Impact on SAT-9 (%): At/Below the 25th Percentile[e]		
All First Things First high schools[c]	**61.7**	**56.4**	**-1.4**	**5.4**	
School E	66.3	55.4	-2.4	3.3	4.0
School F	65.0	58.2	1.0	6.5	
School G	53.8	55.5	-2.9	6.4	

(continued)

SOURCE: MDRC calculations from individual student records from the Houston Independent School District data file.

NOTES: Sample includes 10th-grade students from four clusters. Each cluster consists of a First Things First (FTF) high school matched with a group of between ten and eleven non-FTF schools. The sample consists of students for whom administrative records exist between the 1998-1999 and 2003-2004 academic years.

The "impact" was calculated as the difference between the deviation from the baseline for FTF schools and the deviation from the baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics and prior achievement.

Appendix Table D.8 (continued)

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The baseline level is the average of three pre-implementation years, which are not the same calendar years for each school. For school E, baseline includes the 1998-1999, 1999-2000, and 2000-2001 school years. For the other schools, baseline includes the 1999-2000, 2000-2001, and 2001-2002 school years. This is why one school has three follow-up years and the others have two.

[b]For impacts at the school level, statistical significance is presented; however, these results are much less reliable than pooled impacts.

[c]Follow-up Years 1 and 2 in the "All First Things First high schools" row average together the first and second post-implementation years for each school, which are not the same calendar years for each school.

[d]The desired impact for this measure is positive.

[e]The desired impact for this measure is negative.

The First Things First Evaluation

Appendix Table D.9

Estimated Impact of First Things First on the Percentage of 10th-Graders Scoring At/Above the 50th Percentile and At/Below the 25th Percentile on the SAT-9 in Math: Houston, Texas

	Baseline Level		Estimated Impact by Follow-Up Year[b]		
	FTF[a]	Comparison	Year 1	Year 2	Year 3
			Impact on SAT-9 (%): At/Above the 50th Percentile[d]		
All First Things First high schools[c]	**30.5**	**29.5**	**2.2**	**-2.5**	
School E	30.1	29.6	-1.2	-1.0	-3.0
School F	30.0	27.9	-1.1	-6.1	
School G	31.5	30.9	8.8	-0.5	
			Impact on SAT-9 (%): At/Below the 25th Percentile[e]		
All First Things First high schools[c]	**44.7**	**45.9**	**-4.1**	**3.1**	
School E	48.0	47.2	0.2	-2.5	4.1
School F	43.4	46.4	0.7	8.9	
School G	42.7	44.0	-13.2 **	2.8	

(continued)

SOURCE: MDRC calculations from individual student records from the Houston Independent School District data file.

NOTES: Sample includes 10th-grade students from four clusters. Each cluster consists of a First Things First (FTF) high school matched with a group of between ten and eleven non-FTF schools. The sample consists of students for whom administrative records exist between the 1998-1999 and 2003-2004 academic years.

The "impact" was calculated as the difference between the deviation from the baseline for FTF schools and the deviation from the baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics and prior achievement.

Appendix Table D.9 (continued)

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The baseline level is the average of three pre-implementation years, which are not the same calendar years for each school. For school E, baseline includes the 1998-1999, 1999-2000, and 2000-2001 school years. For the other schools, baseline includes the 1999-2000, 2000-2001, and 2001-2002 school years. This is why one school has three follow-up years and the others have two.

[b]For impacts at the school level, statistical significance is presented; however, these results are much less reliable than pooled impacts.

[c]Follow-up Years 1 and 2 in the "All First Things First high schools" row average together the first and second post-implementation years for each school, which are not the same calendar years for each school.

[d]The desired impact for this measure is positive.

[e]The desired impact for this measure is negative.

The First Things First Evaluation

Appendix Table D.10

Estimated Impact of First Things First on High School Attendance Rates and 9th-Grade Persistence Rates: Houston, Texas

	Baseline Level		Estimated Impact by Follow-Up Year[b]		
	FTF[a]	Comparison	Year 1	Year 2	Year 3
			Impact on Attendance Rates[e] (%)		
All First Things First high schools[c]	**89.6**	**90.2**	**0.0**	**0.2**	
School E	87.8	89.9	0.4	1.9	1.7
School F	89.5	89.9	0.1	0.5	
School G	91.5	90.7	-0.6	-1.8	
			Impact on Persistence Rates[f] (%)		
All First Things First high schools[d]	**73.7**	**77.2**	**-1.0**		
School E	71.4	77.8	0.2	1.9	
School F	77.0	76.5	-1.2		
School G	72.7	77.3	-1.9		

(continued)

SOURCE: MDRC calculations from individual student records from the Houston Independent School District data file.

NOTES: Sample includes 10th-grade students from four clusters. Each cluster consists of a First Things First (FTF) high school matched with a group of between ten and eleven non-FTF schools. The sample consists of students for whom administrative records exist between the 1998-1999 and 2003-2004 academic years.

The "impact" was calculated as the difference between the deviation from the baseline for FTF schools and the deviation from the baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics and prior achievement.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

Appendix Table D.10 (continued)

[a]The baseline level is the average of three pre-implementation years, which are not the same calendar years for each school. For school E, baseline includes the 1998-1999, 1999-2000, and 2000-2001 school years. For the other schools, baseline includes the 1999-2000, 2000-2001, and 2001-2002 school years.

[b]For impacts at the school level, statistical significance is presented; however, these results are much less reliable than pooled impacts.

[c]Follow-up Years 1 and 2 in the "All First Things First high schools" row average together the first and second post-implementation years for each school, which are not the same calendar years for each school.

[d]Follow-up Year 1 in the "All First Things First high schools" row averages together the first post-implementation year for each school, which is not the same calendar year for each school.

[e]The desired impact for this measure is positive.

[f]Persistence rates have one fewer follow-up year because the 2004-2005 academic year, which is not yet available, is needed to calculate the 2003-2004 persistence rate. The desired impact for this measure is positive.

The First Things First Evaluation

Appendix Table D.11

Estimated Impact of First Things First on the Percentage of 8th-Grade Students Passing the TAAS/TAKS in Reading and Math: Houston, Texas

	Baseline Level		Estimated Impact by Follow-Up Year[b]		
	FTF[a]	Comparison	Year 1	Year 2	Year 3
			Impact on (TAAS/TAKS) Reading Pass Rates[d] (%)		
All First Things First middle schools[c]	**82.2**	**84.2**	**-1.6**	**-5.1 ****	
School S	77.2	79.1	1.1	-1.6	-1.9
School U	80.1	83.5	0.1	-5.6	
School V	90.4	89.6	0.3	-4.5	
School T	81.3	84.6	-7.7	-8.6 *	
			Impact on (TAAS/TAKS) Math Pass Rates[d] (%)		
All First Things First middle schools[c]	**82.2**	**83.5**	**2.5**	**1.8**	
School S	80.4	79.0	-2.7	2.2	6.5
School U	82.3	83.4	4.2	3.3	
School V	87.7	86.2	4.1	6.6	
School T	78.3	85.5	4.3	-5.1	

(continued)

SOURCE: MDRC calculations from individual student records from the Houston Independent School District data file.

NOTES: Sample includes 8th-grade students from four clusters. Each cluster consists of a First Things First (FTF) middle school matched with a group of between three and fifteen non-FTF schools. The sample consists of students for whom administrative records exist between the 1998-1999 and 2003-2004 academic years.

The "impact" was calculated as the difference between the deviation from the baseline for FTF schools and the deviation from the baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics and prior achievement.

Appendix Table D.11 (continued)

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The baseline level is the average of three pre-implementation years, which are not the same calendar years for each school. For school S, baseline includes the 1998-1999, 1999-2000, and 2000-2001 school years. For the other schools, baseline includes the 1999-2000, 2000-2001, and 2001-2002 school years. This is why one school has three follow-up years and the others have two.

[b]For impacts at the school level, statistical significance is presented; however, these results are much less reliable than pooled impacts.

[c]Follow-up Years 1 and 2 in the "All First Things First middle schools" row average together the first and second post-implementation years for each school, which are not the same calendar years for each school.

[d]The desired impact for this measure is positive.

The First Things First Evaluation

Appendix Table D.12

Estimated Impact of First Things First on the Percentage of 8th-Graders Scoring At/Above the 50th Percentile and At/Below the 25th Percentile on the SAT-9 in Reading: Houston, Texas

	Baseline Level		Estimated Impact by Follow-Up Year[b]		
	FTF[a]	Comparison	Year 1	Year 2	Year 3
			Impact on SAT-9 (%): At/Above the 50th Percentile[d]		
All First Things First middle schools[c]	**32.1**	**29.7**	**-1.2**	**-5.1**	
School S	31.9	25.5	-2.8	-3.9	-0.8
School U	23.8	25.1	-4.2	-6.5	
School V	39.8	40.4	2.6	-5.4	
School T	32.9	27.9	-0.5	-4.5	
			Impact on SAT-9 (%): At/Below the 25th Percentile[e]		
All First Things First middle schools[c]	**40.1**	**42.2**	**3.2**	**8.0**	
School S	43.9	45.9	-2.7	4.6	-3.3
School U	45.2	47.1	5.7	12.4 *	
School V	32.0	31.8	0.9	1.1	
School T	39.6	44.1	8.7	13.8 *	

(continued)

SOURCE: MDRC calculations from individual student records from the Houston Independent School District data file.

NOTES: Sample includes 8th-grade students from four clusters. Each cluster consists of a First Things First (FTF) middle school matched with a group of between three and fifteen non-FTF schools. The sample consists of students for whom administrative records exist between the 1998-1999 and 2003-2004 academic years.

The "impact" was calculated as the difference between the deviation from the baseline for FTF schools and the deviation from the baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics and prior achievement.

Appendix Table D.12 (continued)

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The baseline level is the average of three pre-implementation years, which are not the same calendar years for each school. For school S, baseline includes the 1998-1999, 1999-2000, and 2000-2001 school years. For the other schools, baseline includes the 1999-2000, 2000-2001, and 2001-2002 school years. This is why one school has three follow-up years and the others have two.

[b]For impacts at the school level, statistical significance is presented; however, these results are much less reliable than pooled impacts.

[c]Follow-up Years 1 and 2 in the "All First Things First middle schools" row average together the first and second post-implementation years for each school, which are not the same calendar years for each school.

[d]The desired impact for this measure is positive.

[e]The desired impact for this measure is negative.

The First Things First Evaluation

Appendix Table D.13

Estimated Impact of First Things First on the Percentage of 8th-Graders Scoring At/Above the 50th Percentile and At/Below the 25th Percentile on the SAT-9 in Math: Houston, Texas

	Baseline Level		Estimated Impact by Follow-Up Year[b]		
	FTF[a]	Comparison	Year 1	Year 2	Year 3
			Impact on SAT-9 (%): At/Above the 50th Percentile[d]		
All First Things First middle schools[c]	**32.4**	**29.0**	**0.4**	**0.3**	
School S	35.2	26.9	-3.5	-12.8 **	7.0
School U	25.5	25.5	-5.4	-2.5	
School V	44.4	35.0	10.7	12.0 *	
School T	24.7	28.4	-0.1	4.5	
			Impact on SAT-9 (%): At/Below the 25th Percentile[e]		
All First Things First middle schools[c]	**40.8**	**43.9**	**0.5**	**3.7**	
School S	41.0	46.3	2.3	7.3	-1.8
School U	47.3	47.7	3.3	5.6	
School V	28.2	37.2	-5.1	-0.7	
School T	46.7	44.2	1.6	2.6	

(continued)

SOURCE: MDRC calculations from individual student records from the Houston Independent School District data file.

NOTES: Sample includes 8th-grade students from four clusters. Each cluster consists of a First Things First (FTF) middle school matched with a group of between three and fifteen non-FTF schools. The sample consists of students for whom administrative records exist between the 1998-1999 and 2003-2004 academic years.

The "impact" was calculated as the difference between the deviation from the baseline for FTF schools and the deviation from the baseline for comparison schools.

Appendix Table D.13 (continued)

Estimates are regression-adjusted for students' background characteristics and prior achievement.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The baseline level is the average of three pre-implementation years, which are not the same calendar years for each school. For school S, baseline includes the 1998-1999, 1999-2000, and 2000-2001 school years. For the other schools, baseline includes the 1999-2000, 2000-2001, and 2001-2002 school years. This is why one school has three follow-up years and the others have two.

[b]For impacts at the school level, statistical significance is presented; however, these results are much less reliable than pooled impacts.

[c]Follow-up Years 1 and 2 in the "All First Things First middle schools" row average together the first and second post-implementation years for each school, which are not the same calendar years for each school.

[d]The desired impact for this measure is positive.

[e]The desired impact for this measure is negative.

The First Things First Evaluation

Appendix Table D.14

Estimated Impact of First Things First on Middle School Attendance Rates: Houston, Texas

	Baseline Level		Estimated Impact by Follow-Up Year[b]		
	FTF[a]	Comparison	Year 1	Year 2	Year 3
			Impact on Attendance Rates[d] (%)		
All First Things First middle schools[c]	**93.8**	**93.6**	**0.1**	**-0.1**	
School S	96.2	93.5	-0.7	-1.0	-2.3 ***
School U	94.4	93.5	0.5	-1.1 *	
School V	93.1	93.5	0.9 *	1.2 **	
School T	91.7	93.8	-0.3	0.4	

(continued)

SOURCE: MDRC calculations from individual student records from the Houston Independent School District data file.

NOTES: Sample includes 6th- to 8th-grade students from four clusters. Each cluster consists of a First Things First (FTF) middle school matched with a group of between three and fifteen non-FTF schools. The sample consists of students for whom administrative records exist between the 1998-1999 and 2003-2004 academic years.

The "impact" was calculated as the difference between the deviation from the baseline for FTF schools and the deviation from the baseline for comparison schools.

Estimates are regression-adjusted for students' background characteristics and prior achievement.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The baseline level is the average of three pre-implementation years, which are not the same calendar years for each school. For school S, baseline includes the 1998-1999, 1999-2000, and 2000-2001 school years. For the other schools, baseline includes the 1999-2000, 2000-2001, and 2001-2002 school years. This is why one school has three follow-up years and the others have two.

[b]For impacts at the school level, statistical significance is presented; however, these results are much less reliable than pooled impacts.

Appendix Table D.14 (continued)

[c]Follow-up Years 1 and 2 in the "All First Things First middle schools" row average together the first and second post-implementation years for each school, which are not the same calendar years for each school.

[d]The desired impact for this measure is positive.

The First Things First Evaluation

Appendix Table D.15

Estimated Impact of First Things First on High School State Test Scores, Attendance Rates, Dropout Rates, and Graduation Rates: Riverview Gardens, Missouri

	Baseline Level		Estimated Impact by Follow-Up Year		
	FTF	Comparison	Year 1	Year 2	Year 3
11th-grade state communication arts test					
Percentage in bottom 2 categories[a]	62.1	64.6	5.6	-7.1	1.5
Percentage in top 2 categories[b]	5.6	7.9	-1.9	4.1	0.0
10th-grade state math test					
Percentage in bottom 2 categories[a]	92.6	87.4	-2.8	-7.6	-10.0
Percentage in top 2 categories[b]	0.4	1.8	0.3	0.9	1.1
Schoolwide average attendance rate[c] (%)	90.4	84.9	0.6	0.6	-0.2
Schoolwide average dropout rate[d] (%)	5.3	7.1	-1.7	1.5	-2.0
Schoolwide average graduation rate[e] (%)	73.0	62.5	0.0	-9.9 **	0.9

(continued)

SOURCE: MDRC calculations from school-level records of state data.

NOTES: Sample includes one First Things First (FTF) high school and eight comparison schools.

The "impact" was calculated as the difference between the deviation from the baseline for FTF schools and the deviation from the baseline for comparison schools.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired impact for this measure is negative.

[b]The desired impact for this measure is positive.

Appendix Table D.15 (continued)

[c]The desired impact for this measure is positive.

[d]The desired impact for this measure is negative.

[e]The desired impact for this measure is positive.

The First Things First Evaluation

Appendix Table D.16

Estimated Impact of First Things First on Middle School State Test Scores and Attendance Rates: Riverview Gardens, Missouri

	Baseline Level		Estimated Impact by Follow-Up Year		
	FTF	Comparison	Year 1	Year 2	Year 3
7th-grade state communication arts test					
Percentage in bottom 2 categories[a]	59.9	70.9	3.7	6.5	0.0
Percentage in top 2 categories[b]	11.2	8.5	-2.1	1.5	2.4
8th-grade state math test					
Percentage in bottom 2 categories[a]	88.3	88.1	-4.5	-9.5	-7.1
Percentage in top 2 categories[b]	1.1	2.4	1.7	3.1	4.1
Schoolwide average attendance rate[c] (%)	91.6	88.6	1.5	0.6	0.6

SOURCE: MDRC calculations from school-level records of state data.

NOTES: Sample includes two "composite" First Things First middle schools and twelve comparison schools.

The "impact" was calculated as the difference between the deviation from the baseline for FTF schools and the deviation from the baseline for comparison schools.

A two-tailed t-test was applied to differences in deviations from baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The desired impact for this measure is negative.

[b]The desired impact for this measure is positive.

[c]The desired impact for this measure is positive.

The First Things First Evaluation

Appendix Table D.17

Estimated Impact of First Things First on 10th-Grade State Test Scores in English II: Delta Region of Mississippi

	Quasi-Baseline Level (Spring 2002)		Estimated Impact by Follow-up Year[a]	
	FTF	Comparison	Spring 2003	Spring 2004
			Impact on English II Mean Score[b]	
All First Things First high schools	**302.4**	**304.7**	**8.4**	**8.5**
School I	306.4	306.1	6.3	3.3
School J	298.4	303.3	10.5	13.7
			Impact on English II Pass Rate[c] (%)	
All First Things First high schools	**52.0**	**55.3**	**12.9**	**8.4**
School I	56.5	57.1	10.4	8.2
School J	47.4	53.5	15.4	8.6

SOURCE: MDRC calculations from school-level records of state data.

NOTES: Sample includes two First Things First (FTF) high schools and six to ten comparison schools.

The "impact" was calculated as the difference between the deviation from the quasi-baseline for FTF schools and the deviation from the quasi-baseline for comparison schools.

A two-tailed t-test was applied to differences in deviations from quasi-baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]For impacts at the school level, statistical significance is presented; however, these results are much less reliable than pooled impacts.

[b]The desired impact for this measure is positive.

[c]The desired impact for this measure is positive.

The First Things First Evaluation

Appendix Table D.18

Estimated Impact of First Things First on 9th-Grade State Test Scores in Algebra: Delta Region of Mississippi

	Quasi-Baseline Level		Estimated Impacts by Follow-up Year[a]	
	FTF	Comparison	Spring 2003	Spring 2004
			Impact on Algebra Mean Score[b]	
School I	309.9	302.2	13.9	-7.9
School J				
Initial comparison group	340.9	309.3	-15.4	-10.8
New comparison group	340.9	334.2	-9.3	-4.7
			Impact on Algebra Pass Rate[c] (%)	
School I	61.1	49.7	15.8	-15.6
School J				
Initial comparison group	94.0	62.4	-16.0	-11.8
New comparison group	94.0	87.2	-10.0	-3.3

(continued)

SOURCE: MDRC calculations from school-level records of state data.

NOTES: Sample includes two First Things First (FTF) high schools and four to ten comparison schools. Impacts for School J are displayed twice, using two different comparison groups. The initial comparison group was selected by taking the average of the English II and algebra scores in the quasi-baseline year and finding non-FTF schools with similar scores. However, School J's English II and algebra scores are so different that the average score did not pick up schools that were performing similarly on the algebra test. A second comparison group was then selected using only the algebra score, and this is the "New" comparison group.

The "impact" was calculated as the difference between the deviation from the quasi-baseline for FTF schools and the deviation from the quasi-baseline for comparison schools.

Appendix Table D.18 (continued)

A two-tailed t-test was applied to differences in deviations from quasi-baseline between FTF and comparison schools. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]For impacts at the school level, statistical significance is presented; however, these results are much less reliable than pooled impacts.

[b]The desired impact for this measure is positive.

[c]The desired impact for this measure is positive.

References

Anderson, Lorin, and David Krathwohl (eds.). 2001. *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* New York: Addison Wesley Longman.

Bloom, Benjamin, Max Engelhart, E. Furst, W. Hill, and David Krathwohl. 1956. *Taxonomy of Educational Objectives, Handbook I: Cognitive Domain.* New York: McKay.

Bloom, Howard S. 2003. "Using Short Interrupted Time-Series Analysis to Measure the Impacts of Whole School Reform: With Applications to a Study of Accelerated Schools." *Evaluation Review* 2 1: 3-49.

Borman, Geoffrey D., Gina M. Hewes, Laura T. Overman, and Shelly Brown. 2004. "Comprehensive School Reform and Achievement: A Meta-Analysis." In Christopher T. Cross (ed.), *Putting the Pieces Together.* Washington, DC: National Clearinghouse for Comprehensive School Reform.

Bridges, Lisa J. 2000. *Measurement Report for FTF Evaluations.* A Paper Prepared for MDRC. New York: MDRC.

Connell, James P. 2003. *Getting Off The Dime: First Steps Toward Implementing First Things First*. Philadelphia: Institute for Research and Reform in Education.

Connell, James P., and Julie Broom. 2004. *The Toughest Nut to Crack: First Things First's (FTF) Approach to Improving Teaching and Learning.* Philadelphia: Institute for Research and Reform in Education.

Connell, James P., and J. G. Wellborn. 1991. "Competence, Autonomy, and Relatedness: A Motivational Analysis of Self-System Process." In M. R. Gunnar and L. A. Sroufe (eds.), *Self-Processes and Development: The Minnesota Symposia on Child Development,* vol. 23. Hillsdale, NJ: Erlbaum.

Estacion Angela, Teresa McMahon, and Janet Quint. 2004. *Conducting Classroom Observations in First Things First Schools.* New York: MDRC.

Gambone, Michelle, Adena M. Klem, William P. Moore, and Jean A. Summers. 2002. *First Things First: Creating the Conditions and Capacity for Community-Wide Reform in an Urban School District.* Philadelphia: Gambone and Associates.

Gambone, Michelle, Adena M. Klem, Jean A. Summers, Theresa Akey, and Cynthia Sipe. 2004. *Turning the Tide: The Achievements of the First Things First Education Reform in the Kansas City, Kansas, Public School District.* Philadelphia: Youth Development Strategies, Inc.

Institute for Research and Reform in Education (IRRE). 2002. *A Guide for Family Advocates*. Philadelphia: Institute for Research and Reform in Education.

Institute for Research and Reform in Education (IRRE). 2004. *A Guide to the Family Advocate System.* Philadelphia: Institute for Research and Reform in Education.

Klem, Adena M., Laura Levin, Susan Bloom, and James P. Connell. 2003. *First Things First's Family Advocate System: Building Relationships to Improve Student Success.* Philadelphia: Institute for Research and Reform in Education.

Quint, Janet C. 2002. *Scaling Up First Things First: Site Selection and the Planning Year*. New York: MDRC.

Quint, Janet C., D. Crystal Byndloss, and Bernice Melamud. 2003. *Scaling Up First Things First: Findings from the First Implementation Year.* New York: MDRC.

Skinner, E. A., M. J. Zimmer-Gembeck, and James P. Connell. 1998. "Individual Differences and the Development of Perceived Control." *Monographs of the Society for Research in Child Development* 63, 2-3. Chicago: University of Chicago Press.

PREVIOUS PUBLICATIONS ON SCALING UP FIRST THINGS FIRST

Conducting Classroom Observations in First Things First Schools
2004. Angela Estacion, Teresa McMahon, Janet Quint, with Bernice Melamud, LaFleur Stephens

First Things First
Findings from the First Implementation Year
2003. Janet C. Quint, D. Crystal Byndloss, with Bernice Melamud

Scaling Up First Things First
Site Selection and the Planning Year
2002. Janet Quint

First Things First
Creating the Conditions and Capacity for Community-Wide Reform in an Urban School District
2002. Prepared by Gambone & Associates

About MDRC

MDRC is a nonprofit, nonpartisan social policy research organization. We are dedicated to learning what works to improve the well-being of low-income people. Through our research and the active communication of our findings, we seek to enhance the effectiveness of social policies and programs. MDRC was founded in 1974 and is located in New York City and Oakland, California.

MDRC's current projects focus on welfare and economic security, education, and employment and community initiatives. Complementing our evaluations of a wide range of welfare reforms are new studies of supports for the working poor and emerging analyses of how programs affect children's development and their families' well-being. In the field of education, we are testing reforms aimed at improving the performance of public schools, especially in urban areas. Finally, our community projects are using innovative approaches to increase employment in low-income neighborhoods.

Our projects are a mix of demonstrations — field tests of promising program models — and evaluations of government and community initiatives, and we employ a wide range of methods to determine a program's effects, including large-scale studies, surveys, case studies, and ethnographies of individuals and families. We share the findings and lessons from our work — including best practices for program operators — with a broad audience within the policy and practitioner community, as well as the general public and the media.

Over the past quarter century, MDRC has worked in almost every state, all of the nation's largest cities, and Canada. We conduct our projects in partnership with state and local governments, the federal government, public school systems, community organizations, and numerous private philanthropies.