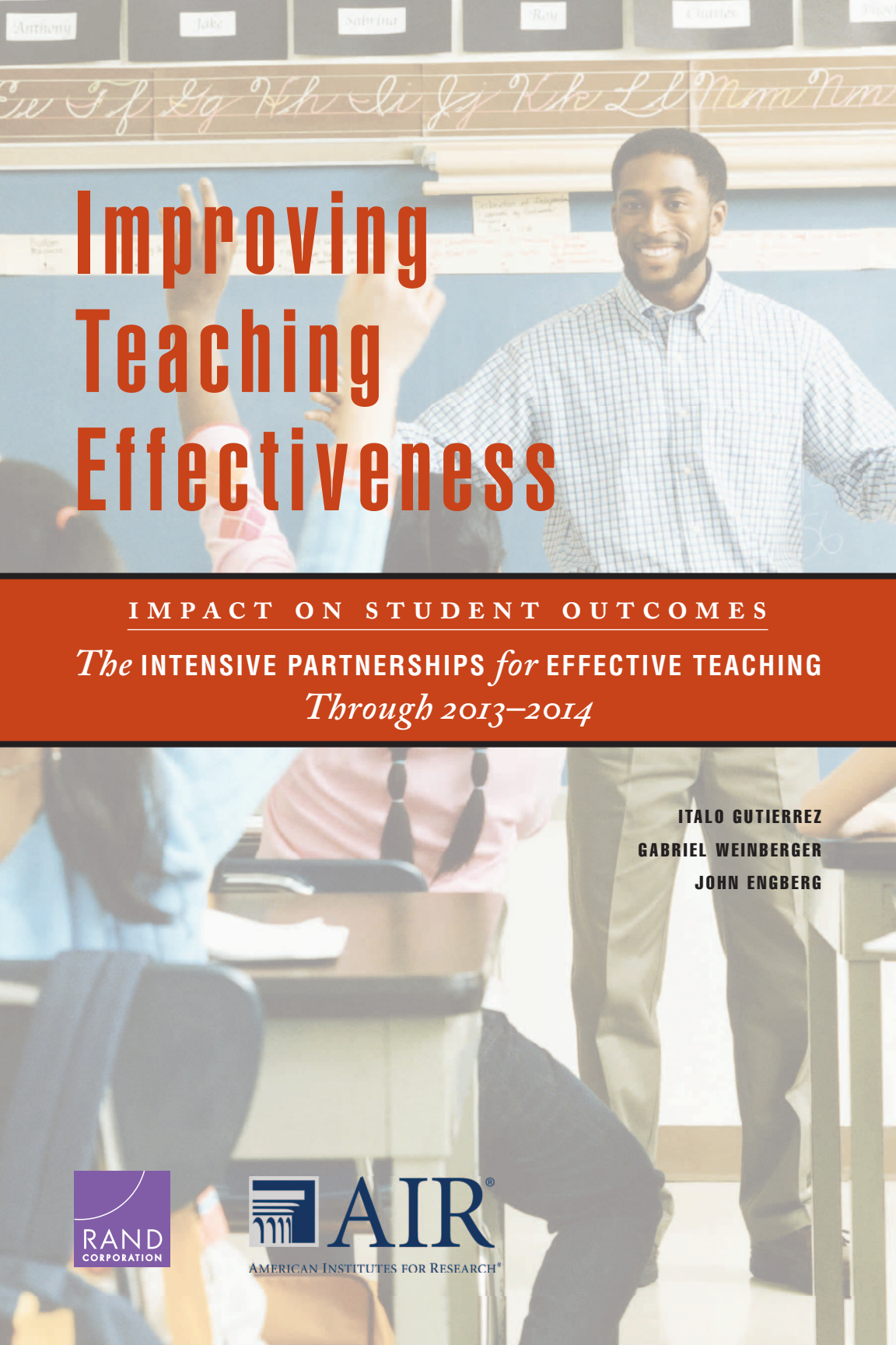


Improving Teaching Effectiveness

IMPACT ON STUDENT OUTCOMES

The INTENSIVE PARTNERSHIPS *for* EFFECTIVE TEACHING *Through 2013–2014*

ITALO GUTIERREZ
GABRIEL WEINBERGER
JOHN ENGBERG

RAND CORPORATION
AIR®
AMERICAN INSTITUTES FOR RESEARCH®

For more information on this publication, visit www.rand.org/t/RR1295z3-1

Library of Congress Cataloging-in-Publication Data is available for this publication.
ISBN: 978-0-8330-9552-7

Published by the RAND Corporation, Santa Monica, Calif.

Cover: Teacher Standing in Front of a Class of Raised Hands, Digital Vision.

Preface

The Bill & Melinda Gates Foundation launched the Intensive Partnerships for Effective Teaching in 2009–2010. After careful screening, the foundation identified seven Intensive Partnership sites—three school districts and four charter management organizations—to implement strategic human-capital reforms over a six-year period.[1] The grants to the districts will continue through the 2015–2016 school year. The foundation also selected the RAND Corporation and its partner, the American Institutes for Research, to evaluate the Intensive Partnership efforts. The RAND/American Institutes for Research team is conducting three interrelated studies examining the implementation of the reforms, the reforms' effect on student outcomes, and the extent to which the reforms are replicated in other districts.

The evaluation began in July 2010 and collected its first wave of data during the 2010–2011 school year; it will continue through the 2015–2016 school year and produce a final report in 2017. Until now, the evaluation produced primarily internal reports to the sites and the foundation. These documents have been useful for monitoring progress in implementing the reforms, identifying challenges and successful strategies to share among the sites, and making midcourse corrections. Now that the sites have implemented most of the elements of the Intensive Partnerships initiative and these elements have had an opportunity to influence teacher placement, teacher practices, and stu-

[1] We use the word *site* to describe the three school districts and the four charter management organizations that received funding from the foundation to implement the Intensive Partnerships initiative.

dent outcomes, it is appropriate to share information on their effects to date. Although these represent interim rather than final outcomes, if the reform is to succeed, we would expect to see some evidence of improved students' outcomes by now.

The present report focuses on the initiative's overall effect on student achievement. The report uses both a school-level analysis and a district-level analysis to compare student outcomes in the Intensive Partnership sites and comparable non–Intensive Partnership schools and districts in the same states. This report should be of interest to practitioners, leaders of education organizations that are funding or implementing human-resource reforms, and researchers interested in large-scale education reform.

Contents

Preface iii
Figures vii
Tables ix
Summary xi
Acknowledgments xvii
Abbreviations xix

CHAPTER ONE
Introduction 1

CHAPTER TWO
Implementation of the Intervention 7
Teacher Evaluation 8
Staffing 9
Professional Development 10
Compensation and Career Ladders 11
State Policy Changes 13

CHAPTER THREE
Data and Methods 17
Data and Outcomes 17
School-Level Difference-in-Differences Methodology 23
Synthetic-Control-Group Methodology 27

CHAPTER FOUR
Results 31
Hillsborough County Public Schools in Florida 31
Pittsburgh Public Schools, Pennsylvania 41
Memphis City Schools, Tennessee 49

CHAPTER FIVE
Putting the Estimates in Context 55
School-Level Interventions 55
District-Level Interventions 59

CHAPTER SIX
Summary and Conclusion 65

APPENDIXES
A. Estimation Methods 69
B. Results for Additional Outcomes 77
C. Specific Practices for Levers of Implementation 99

Bibliography 101

Figures

1.1. States That Require Teacher Evaluations to Include Student Achievement Measures 5
2.1. The Intensive Partnerships Initiative's Steps to Student Success 7
2.2. Proportion of the Teacher-Evaluation Lever Implemented, Spring 2010 to Spring 2014 9
2.3. Proportion of the Staffing Lever Implemented, Spring 2010 to Spring 2014 10
2.4. Proportion of the Professional-Development Lever Implemented, Spring 2010 to Spring 2014 11
2.5. Proportion of the Compensation and Career-Ladder Lever Implemented, Spring 2010 to Spring 2014 12
3.1. Graphical Depiction of Methodology for Computing Forecasts of Postintervention Trends 25
4.1. Average School-Level Test Scores on Grade 3 Through 8 Mathematics and Reading, Hillsborough County and Other Florida Districts 32
4.2. Estimates of the Intensive Partnerships Initiative's Effect on Grade 3 Through 8 Math, Hillsborough County, Florida 36
4.3. Estimates of the Intensive Partnerships Initiative's Effect on Grade 3 Through 8 Reading, Hillsborough County, Florida 37
4.4. Average School-Level Test Scores on High School Reading, Hillsborough County and All Other Florida Districts 39
4.5. Estimates of the Intensive Partnerships Initiative's Effect on High School Reading, Hillsborough County, Florida 40

4.6. Average School-Level Test Scores on Grade 3 Through 8 Mathematics and Reading, Pittsburgh and All Other Pennsylvania Districts 42
4.7. Estimates of the Intensive Partnerships Initiative's Effect on Grade 3 Through 8 Math, Pittsburgh, Pennsylvania 43
4.8. Estimates of the Intensive Partnerships Initiative's Effect on Grade 3 Through 8 Reading, Pittsburgh, Pennsylvania 44
4.9. Average School-Level Test Scores on High School Reading, Pittsburgh and All Other Pennsylvania Districts 47
4.10. Estimates of Effect of Intensive Partnerships Initiative on High School Reading, Pittsburgh 48
4.11. Average School-Level Test Scores on Grade 3 Through 8 Math, Memphis City Schools and All Other Tennessee Schools 50
4.12. Estimates of the Intensive Partnerships Initiative's Effect on Grade 3 Through 8 Math, Memphis, Tennessee 51
4.13. Estimates of the Intensive Partnerships Initiative's Effect on Grade 3 Through 8 Reading, Memphis, Tennessee 52
5.1. Effect Sizes for Reading Scores, Grades 3 Through 8 61
5.2. Effect Sizes for Math Scores, Grades 3 Through 8 62
5.3. Effect Sizes for Reading Scores, High School 63

Tables

2.1. Progress in States' Teacher Quality Policies 14
3.1. Summary of Data Elements 20
3.2. Mean Values for District Average Standardized Test Scores and Demographics 21
5.1. Average Annual Gains in Effect Size from Nationally Normed Tests 56
5.2. Average Effect Sizes in Educational Interventions 57
A.1. Synthetic-Control-Group Methodology: Number of Districts in the Construction of the Control Group and of the Placebo Distribution 75
B.1. Hillsborough County Public Schools Impact Estimates, by Grade, Subgroup, and Year 78
B.2. Pittsburgh Public Schools Impact Estimates, by Grade, Subgroup, and Year 88
B.3. Memphis City Schools Impact Estimates, by Grade, Subgroup, and Year 96

Summary

This interim report presents estimates of the overall effect that the Bill & Melinda Gates Foundation's Intensive Partnerships for Effective Teaching initiative has had on student outcomes through the 2013–2014 school year. The aim of the initiative is to encourage and support strategic human-capital reforms that are intended to improve the ways in which "teachers are recruited, evaluated, supported, retained, and rewarded" (Bill & Melinda Gates Foundation, 2011). The cornerstone of the reform is the development and implementation of teacher-evaluation systems that are based on student achievement growth; structured classroom observations by principals or trained peers; and other inputs, such as student or parent surveys. These evaluations are used to guide personnel practices in three broad areas—staffing, professional development, and compensation and career-ladder decisions—with the goal of giving every student access to highly effective teachers. Staffing practices include such activities as expedited recruiting and incentivizing effective teachers to work in high-need schools; professional-development practices include feedback, coaching, and mentoring related to teachers' identified strengths and weaknesses; and compensation practices include monetary rewards for effective teachers and incentives for teaching in high-need positions.

This initiative is being implemented in sites that the foundation chose, including three large urban districts and four charter management organizations (CMOs) that are a part of the College-Ready Promise. The districts are Hillsborough County Public Schools (HCPS)

in Florida, Memphis City Schools (MCS) in Tennessee,[2] and Pittsburgh Public Schools (PPS) in Pennsylvania. The CMOs are the Alliance College-Ready Public Schools, Aspire Public Schools, Green Dot Public Schools, and Partnerships to Uplift Communities Schools. All sites have implemented most of the elements of the initiative to some degree, although there is variation by site. Enough change has occurred that it is reasonable to test whether there is evidence of improved students' outcomes. This report does not include results for any of the CMOs because student achievement data for the 2013–2014 school year are not available in California, where most of these schools are located.

Analyses

The question our analysis is designed to answer is how much better students in the Intensive Partnership sites are doing than they would have done without the Intensive Partnership reforms. Our estimate of this impact is based primarily on a difference-in-differences (DiD) methodology. This method compares outcomes before and after the implementation of the Intensive Partnerships initiative between the schools in the Intensive Partnership sites and the rest of the schools in the same state. As a robustness check on the DiD methodology, we also conduct a synthetic-control-group (SCG) comparison for the Intensive Partnership sites. The SCG approach is conducted at the district level, and it involves combining data from other districts in the same state to construct an SCG that resembles the intervention district in terms of preintervention characteristics, i.e., they are similar before the start of the reform on key features. An advantage of the SCG method is that it allows for more-robust statistical inference in the case in which common shocks (i.e., unrelated reforms, such as an early-retirement package or changes in district leadership), affect all schools in the

[2] MCS has merged with Shelby County Schools, but our focus is on the schools that were formerly a part of MCS, so we continue to refer to them as MCS as a reminder of this focus.

Intensive Partnership site. However, this method has its own limitations as well. We present the results of the SCG method only as a way to check whether they are consistent with those from the DiD analysis.

To keep the discussion of the results to a manageable length, in the main text, we focus on the Intensive Partnerships initiative's effect on school-level average student scores (across grades) on state-administered standardized tests. In a supplemental spreadsheet described in Appendix B, we report additional results showing breakdowns by grade and population subgroup, as well as results on nontest outcomes, such as graduation rate.

Findings

Because of the way the sites measure student achievement, we estimated the effect of the Intensive Partnership reform in the lower grades (3 through 8) separately from the effect in high school. We found mixed but mostly insignificant effects of the initiative on student performance in the lower grades, with the exception of MCS, which fared significantly worse after the start of the initiative. However, impact estimates were increasing in 2013–2014 for all but one of the achievement outcomes in all grade levels in the sites. If the more-recent trends continue, the sites could observe significant positive impact in the next years. This would not be surprising because the sites needed several years to implement the broad set of reforms that the initiative promoted. It is important to note that the impact estimates using the DiD and the SCG methodologies are similar in most cases, a fact that lends credence to the robustness of the results and strengthens our confidence in the overall findings.

For example, focusing on the 2013–2014 school year (latest available data) and using the DiD methodology, we found that, on average, the schools of PPS experienced greater achievement gains in grade 3–8 mathematics (0.10 standard deviations) and reading (0.02 standard deviations) than comparable schools did in other Pennsylvania dis-

tricts.[3] Schools of HCPS experienced greater achievement gains in reading in the lower grades (3 through 8) (0.03 standard deviations) and performed similarly in mathematics to comparable schools in other Florida districts. However, the DiD estimates fell short of the conventional levels of statistical significance in most cases.[4] Similarly, using the DiD methods, we found that, for grades 3–8, schools in MCS experienced lower achievement gains in school year 2013–2014 in mathematics (–0.18 standard deviations) than comparable schools did in other Tennessee districts, although we observe a rebound in the recent years after a large dip in the first three years following the start of the initiative. We observe a similar dip and rebound in reading in the lower grades (3 through 8) in MCS but still find an average negative effect (–0.02 standard deviations) in school year 2013–2014 in comparison with similar schools in Tennessee. However, neither of these estimates is statistically significant at conventional levels.

We found evidence of negative effects of the Intensive Partnership intervention on high school student reading achievement. We found that the intervention is associated with a reduction in reading test scores of 0.08 standard deviations for high schools in PPS and in HCPS compared with other high schools in Pennsylvania and Florida, respectively.[5] We do not have estimates of the impact on high school mathematics because the change from end-of-grade to end-of-course tests made the postintervention samples incomparable to the preintervention samples.

---

[3] We express impact as fractions of a student-level statewide standard deviation of the relevant test score. In Chapter Five, we provide more detail on this measure. For example, an average year of school for students in grades 3 through 8 is equivalent to approximately one-third of a standard deviation for reading and one-half of a standard deviation for mathematics.

[4] In applied work, statistical significance is usually stated at the 95-percent confidence level.

[5] We do not estimate results for high school test scores for MCS because Tennessee administers only end-of-course exams for high school students. The sample of students taking these end-of-course exams is determined by many factors that we have not measured; therefore, the test results are not useful for school-level comparisons.

We also estimated the effects by subgroups of students, where these data were available. In particular, we estimated the initiative's impact on academic achievements for black students, Hispanic students, and economically disadvantaged students. In most cases, the effects for subgroups followed the same pattern as the overall effects. The only exception is that, in PPS, we found statistically significant improvements in reading in 2014 for black high school students and economically disadvantaged high school students (0.15 and 0.10 standard deviations, respectively).

To put the results in context, we compared the Intensive Partnership impact estimates for students' performance in math and reading with estimates of the average yearly academic growth reported elsewhere for these grades and with the impact of other large-scale interventions found in the literature. We found that the Intensive Partnership impact estimates for the last school year (2013–2014) are smaller than the normal yearly gains in achievement and in comparison with many other school-level interventions. However, the Intensive Partnership impact estimates for grade 3–8 reading in PPS and HCPS and for grade 3–8 math in PPS are in the same range as those estimates from other district-level interventions we identified. In addition, there is evidence of an upward trend in the estimates, which, if continued, would lead to a bigger impact in the coming years.

In addition to looking at district-level effects on standardized test scores, we estimated impacts on graduation and dropout rates in each of the sites. These results are also mixed. For instance, we obtained positive and statistically significant effects (i.e., increases) on high school graduation rates in PPS but negative and statistically significant effects (i.e., decreases) on graduation rates in MCS. Appendix B provides these estimates.

In conclusion, the evidence to date is mixed regarding the impact on student achievement of the broad set of reforms in teacher evaluation and workforce management embodied in the Intensive Partnership initiative. Because of the time needed to implement these reforms, it might still be too soon to draw definitive conclusions. For instance, we found an upward trajectory for most academic outcomes between the most-recent two school years, 2012–2013 and 2013–2014. This

suggests that the reforms might be on the way to having a positive impact after a few transition years during which the reforms produced no effects or even a negative impact. We continue to monitor these effects for the 2014–2015 and 2015–2016 school years and will report on our findings in a future report.

Acknowledgments

We are grateful to the large number of district administrative staff and community stakeholders who reviewed our tables and charts and helped us correctly analyze the Intensive Partnerships initiative's impact in each site. These individuals include Anna Brown and Ted Dwyer in Hillsborough County Public Schools; Bradley Leon and Jessica Lotz in Shelby County Schools; and Tara Tucci and Ashley Varrato in Pittsburgh Public Schools. In addition, our analyses used student achievement data from all the schools and districts in each state, and we received assistance in obtaining and understanding state testing data from John Nau and Milad Elhadri at the Pennsylvania Department of Education, Mary Batiwalla from the Tennessee Department of Education, and Sonja L. Bridges from the Florida Department of Education. We also appreciate the willingness of the foundation staff to engage with us; we are especially grateful to Sara Allan, Nate Brown, Sarah Buhayar, Steven Cantrell, Josh Edelman, Eli Pristoop, Irvin Scott, Greg Sommers, and Ky Vu for their advice and responsiveness throughout the project. We thank Kata Mihaly and Cathleen Stasz of RAND and Susanna Loeb of Stanford University for reviewing the document and providing constructive feedback. Finally, we acknowledge other members of the RAND team, including Gerald Hunter, Stephanie Lonsinger, Paco Martorell, Ethan Scherer, Brian M. Stecher, and Elizabeth D. Steiner, and of the American Institutes for Research team, especially Michael S. Garet.

Abbreviations

ASD	Achievement School District
CDDRE	Center for Data-Driven Reform in Education
CMO	charter management organization
CSR	Comprehensive School Reform
DAIT	District Assistance and Intervention Team
DiD	difference in differences
ELL	English-language learner
EOC	end of course
FCAT	Florida Comprehensive Assessment Test
FRPL	free and reduced-price lunch
HCPS	Hillsborough County Public Schools
MCS	Memphis City Schools
NCTQ	National Council on Teacher Quality
PPS	Pittsburgh Public Schools
PSSA	Pennsylvania System of School Assessment
RMSPE	root mean-squared prediction error
SCG	synthetic control group
SCS	Shelby County Schools

CHAPTER ONE

Introduction

This report presents findings from the evaluation by the RAND Corporation and the American Institutes for Research of the Bill & Melinda Gates Foundation's Intensive Partnerships for Effective Teaching initiative. The Intensive Partnerships initiative was implemented beginning in 2009–2010 in three districts and four charter management organizations (CMOs) that are a part of the College-Ready Promise. The districts are Hillsborough County Public Schools (HCPS) in Florida, Memphis City Schools (MCS) in Tennessee,[1] and Pittsburgh Public Schools (PPS) in Pennsylvania. The CMOs are the Alliance College-Ready Public Schools, Aspire Public Schools, Green Dot Public Schools, and Partnerships to Uplift Communities Schools.

The aim of the initiative is to encourage and support strategic human-capital reforms that are intended to improve the ways in which "teachers are recruited, evaluated, supported, retained, and rewarded" (Bill & Melinda Gates Foundation, 2011). The cornerstone of the reform is the development and implementation of teacher-evaluation systems that are based on student achievement growth; structured classroom observations by principals or trained peers; and other inputs, such as student or parent surveys. These evaluations are used to guide practices related to staffing, professional development, and compensation and career-ladder decisions, with the goal of giving every student

[1] MCS has merged with Shelby County Schools (SCS), but our focus is on the schools that were formerly a part of MCS, so we continue to refer to them as MCS as a reminder of this focus.

access to highly effective teachers. Chapter Two provides more details regarding the initiative.

Although it has taken multiple years to implement this broad set of reforms, enough change has occurred that it is reasonable to test whether there is evidence of improvements in overall teaching effectiveness, more-equitable distribution of effective teaching across schools and students, and better outcomes for all students. This report focuses on the latter. An accompanying report examines the level and distribution of effective teaching (Baird et al., 2016).

This report presents estimates of the reforms' impact on student outcomes (i.e., student achievement, graduation rates, and dropout rates) through the 2013–2014 school year for HCPS, MCS, and PPS. Specifically, we examine whether the initiative changed the performance of students in the Intensive Partnership sites relative to what would have happened had the initiative not been implemented.

This report does not include analyses for any of the CMOs that are participating in the Intensive Partnerships initiative because these sites are primarily in California, which is currently going through a transition period in its student assessment system. The state conducted a pilot of new Smarter Balanced tests in the 2013–2014 school year and did not publish the results. We do not think that it would be appropriate to present estimates for the CMOs based on data from a year prior. However, we plan to include the CMOs in our next report, which will be based on academic achievement data from the 2014–2015 school year in all Intensive Partnership states (including California).

We examine the Intensive Partnerships initiative's effects on student performance using two methods. First, we estimate a school-level difference-in-differences (DiD) regression that compares the average test scores of schools in the Intensive Partnership sites and the scores of schools in other districts in the state, accounting for demographics and prereform performance. Second, we estimate a district-level synthetic-control-group (SCG) model that compares the aggregate performance in the Intensive Partnership site and a weighted average of the perfor-

mances in other districts in the state.[2] We use these two approaches to estimate the impact of the initiative using data through the 2013–2014 school year. We report impacts for each of the five years after the 2008–2009 school year, which was the last preinitiative year.[3] The primary outcome measures are average scores on standardized state assessments.[4] In addition to state test scores, we report estimated effects for graduation rates and for different subgroups. Although we do not discuss these findings in detail in this report (to keep the text manageable), we include them in Appendix B.

There are two important caveats in interpreting the results presented in this report. The first one is that it might take longer than reported here for the initiative to affect student outcomes as the sites continue implementing the reforms designed to improve effective teaching. No available benchmarks from similar district-level interventions could help to determine when to expect the initiative's full impact. There is, however, some evidence on the relationship between years of implementation and effect size for school-level interventions. A meta-analysis study, reported in Borman et al., 2003, analyzed evidence on the effects of several Comprehensive School Reform (CSR) models, such as Direct Instruction, School Development Program, and Success for All. The authors found that their effect size was fairly similar in schools that had implemented these reforms for up to four years (on average, around 0.15 standard deviations). But the effects almost double for schools that had implemented these reforms for five or six

[2] The first method has the benefit of providing more power to detect effects but makes some assumptions on the distribution of districtwide shocks that might lead to bias. The second method makes fewer assumptions but has lower power. We found that the estimates using both methods are similar, a fact that favors the use of the DiD method for significance testing.

[3] Funding from the foundation began in the spring of 2010, so we consider 2008–2009 and earlier to be preintervention years. However, many aspects of the reforms did not begin to be implemented until the 2010–2011 school year or later.

[4] Average scale scores are a preferable outcome to other measures, such as the fraction of students meeting proficiency, because the fraction proficient will capture only effects on the test score distribution that result in students crossing the proficiency threshold, whereas averages reflect achievement by all students.

years (0.25 standard deviations) and increase to more than 2.5 times (0.39 standard deviations) for schools with seven years of implementation. Thus, it is possible that the results we present in this report constitute a lower bound of the effects of combining rigorous teacher evaluation with changes in workforce management practices. The effects might be larger in the future.

The second caveat in the interpretation of the results is that both methods used in this report provide estimates of the Intensive Partnerships initiative's impact in comparison with a counterfactual scenario in which the initiative did not occur. Some state-level policy changes during this period influenced both the Intensive Partnership sites and the other sites in the state, and our estimates of impact cannot reveal the initiative's effect had these changes not occurred. In particular, the Bill & Melinda Gates Foundation and other organizations have worked at the state and national levels to promote human-capital reforms in an effort to improve the quality of education and student outcomes. In fact, since the start of the Intensive Partnerships initiative, the vast majority of states have implemented legislation requiring student achievement to be incorporated into teacher evaluation (see Figure 1.1). Two of the Intensive Partnership sites, HCPS and SCS, are in states (Florida and Tennessee, respectively) that were early adopters of state-level policies and programs to reform teacher evaluation and use student achievement in teacher evaluations. The impact estimates presented in this report should be interpreted as the improvements in student performance that can be attributed to the Intensive Partnerships initiative *over and above* any improvement resulting from other state- or national-level policy changes.

The rest of the report is organized as follows: Chapter Two provides a brief description of the intervention and the progress of the implementation, in order to give more context to the findings. Chapter Three discusses the data and presents the DiD methodology and the SCG methodology for estimating the initiative's impact on students' outcomes. Chapter Four presents the impact estimates separately for math and reading, for lower grades (3 through 8) and for high school (grades 9 through 11) for each district. Chapter Five compares the Intensive Partnerships impact estimates and some other benchmarks,

Figure 1.1
States That Require Teacher Evaluations to Include Student Achievement Measures

SOURCES: Doherty and Jacobs, 2013; Center on Great Teachers and Leaders, undated; state legislative documents.
NOTE: Elementary and Secondary Education Act is Public Law 89-10, 1965.

RAND *RR1295/3-1-1.1*

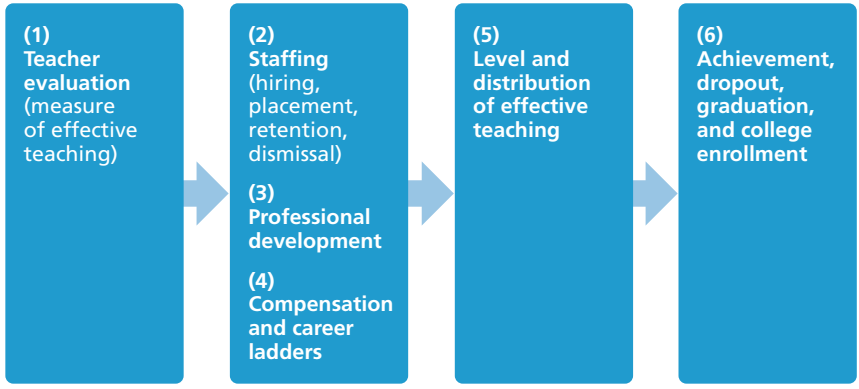
including the usual achievement gains expected in an school year and the effects from other past interventions; these findings provide some context for the size of the Intensive Partnerships initiative's estimated impacts. Chapter Six summarizes our findings and provides conclusions regarding the initiative's effectiveness in improving students' performance. We also provide three appendixes: Appendix A on estimation methods, Appendix B with results for additional outcomes, and Appendix C with specific practices evaluated for each component of implementation.

CHAPTER TWO

Implementation of the Intervention

The initiative's aim is to encourage and support strategic human-capital reforms that are intended to improve the ways in which "teachers are recruited, evaluated, supported, retained, and rewarded" (Bill & Melinda Gates Foundation, 2011). Figure 2.1, adapted from Stecher et al., 2016, illustrates the theory of action behind the Intensive Partnerships initiative. The cornerstone in the reform process is to adopt an improved teacher-evaluation system, including the adoption of a teacher-effectiveness measure (1). This measure is used in managing the teacher workforce over time, including decisions about (2) staffing, (3) professional development, and (4) compensation and specialized

Figure 2.1
The Intensive Partnerships Initiative's Steps to Student Success



SOURCE: Stecher et al., 2016.
RAND *RR1295/3-1-2.1*

positions (so-called career ladders). These decisions, in turn, should (5) increase the effectiveness of teaching and the distribution of teachers throughout schools. More-effective teaching should improve student performance, including (6) increased student achievement, lowered dropout rates, and increased graduation and college enrollment rates.

The Bill & Melinda Gates Foundation and the sites refer to elements 1 through 4 as policy levers, and this chapter offers a brief description of each of the levers. To provide some context for the findings, we also describe each site's progress in implementing the levers. We base this description mostly on Stecher et al., 2016, which provides a much more detailed account. Moreover, Stecher et al., 2016, Appendix B, lists and defines each of the specific practices that go into the broad levers described below, while Appendix D in that report describes whether (and to what extent) each site has implemented each specific practice.

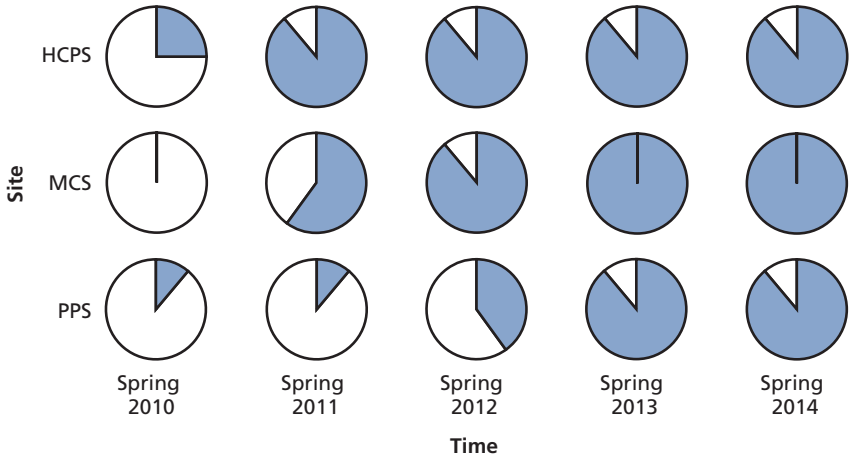
Most sites have taken multiple years to implement this broad set of reforms. Of necessity, changes to teacher-evaluation systems preceded changes to staffing, professional development, and compensation and career-ladder policies that rely on the teacher-evaluation measures. In the following sections, we briefly describe the implementation of the four broad levers.

Teacher Evaluation

The implementation of the teacher-evaluation lever covered eight specific practices, including classroom observations by principals and other observers, the use of student achievement growth in evaluations, and the use of parent surveys (see Appendix C for a complete list of practices related to teacher evaluation). Figure 2.2 shows the proportion of practices in the teacher-evaluation lever that each site implemented annually from the spring of 2010 to the spring of 2014.[1] Sites

[1] For a comprehensive description of which practices each site has implemented, see Stecher et al., 2016, Appendix D.

Figure 2.2
Proportion of the Teacher-Evaluation Lever Implemented, Spring 2010 to Spring 2014



SOURCE: Stecher et al., 2016.
RAND *RR1295/3-1-2.2*

did not necessarily plan, nor were they expected, to implement all of the practices included as part of this lever. Thus, we should not expect all sites to achieve 100-percent implementation (i.e., a fully colored circle in Figure 2.2). By the spring of 2012, the Intensive Partnership sites except PPS had implemented a majority of the practices. By the spring of 2014, all sites had implemented all of the teacher-evaluation practices that they intended to implement.

Staffing

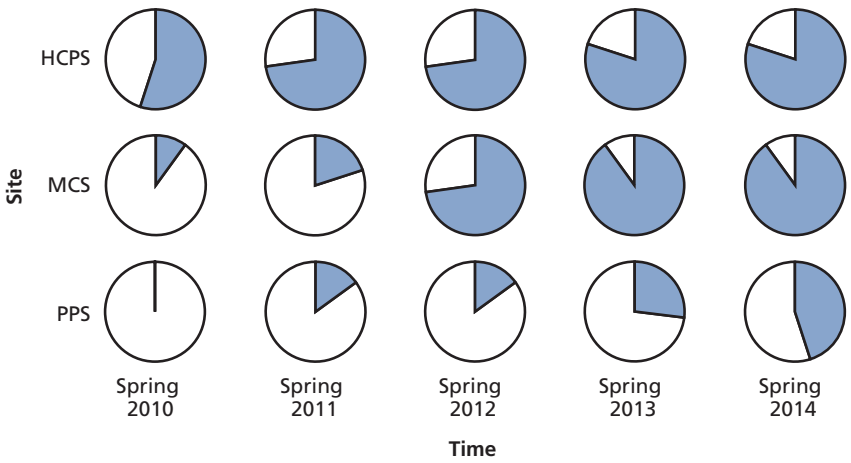
The sites made several changes in staffing practices, including expediting recruiting, training administrators to make good hiring decisions, offering incentives to work in high-need schools, and linking tenure and retention decisions to effectiveness ratings (see Appendix C for a complete list of practices related to staffing). The expedited recruiting, which can be accomplished early in the calendar year (e.g., February

or March), along with better recruiting and hiring methods will give the district an advantage over others competing for the same teachers. The specific changes varied from site to site, reflecting local laws and contractual agreements. Figure 2.3 shows the status of the staffing lever over time across Intensive Partnership sites. By 2012, HCPS and MCS had most practices in place. PPS has progressed more slowly in the implementation of the new staffing practices.

Professional Development

The professional-development lever includes using the evaluation data to identify teachers' individual development needs and then offering professional development, feedback, coaching, or mentoring targeted to those needs. The lever also includes supports for new teachers; supervisor oversight of teachers' participation in professional development; and an electronic system for data collection, which would record which teachers accessed what resources (see Appendix C for a complete list of

Figure 2.3
Proportion of the Staffing Lever Implemented, Spring 2010 to Spring 2014



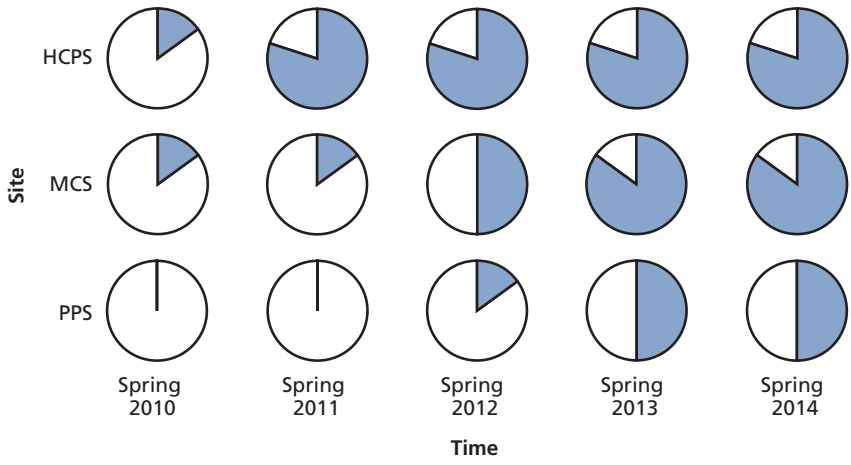
SOURCE: Stecher et al., 2016.
RAND *RR1295/3-1-2.3*

practices related to professional development). Figure 2.4 depicts the status of these practices over time across the Intensive Partnership sites. Sites did not begin to change policies related to professional development until they had their measure of effectiveness in place (roughly 2012), and they still have not implemented all possible policies. In particular, customizing professional development to address individual needs has proven to be logistically challenging.

Compensation and Career Ladders

Compensation and career-ladder policies are a key component of the Intensive Partnerships initiative. The compensation portion of the lever includes monetary rewards for effective teachers, as well as incentives for teaching in high-need positions. The lever also reflects whether the site bases some of teachers' salary on effective performance. The career-ladder portion of the lever includes creating specialized roles

Figure 2.4
Proportion of the Professional-Development Lever Implemented, Spring 2010 to Spring 2014

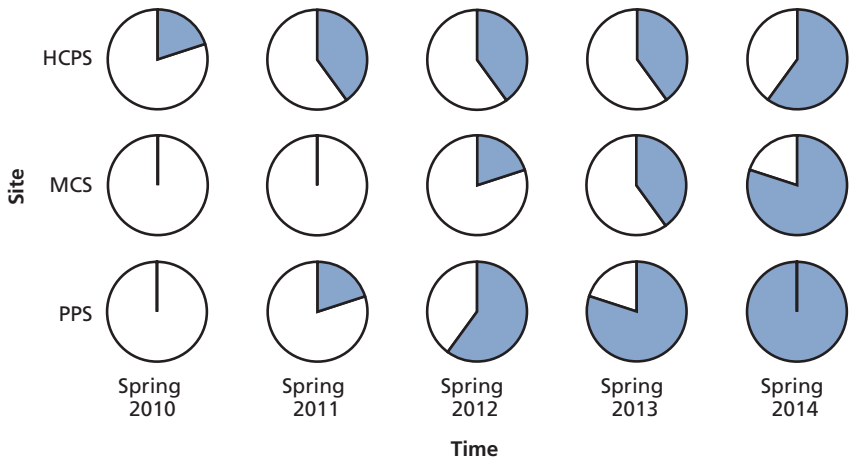


SOURCE: Stecher et al., 2016.

RAND*RR1295/3-1-2.4*

for teachers that offer rewards for taking on extra responsibilities and demonstrating greater leadership (see Appendix C for a complete list of practices related to compensation and career ladders). It is expected that these types of incentives will lead to better teaching, which should lead to better student outcomes. The implementation of these practices had to wait until the teaching-effectiveness measures were in place. It also involved changes in negotiated contracts. However, by the 2013–2014 school year, every site had adopted some form of effectiveness bonus. Also, every site had developed some form of career ladder in which effective teachers take on new roles, such as coaching or mentoring, and receive additional salary or stipend for these responsibilities. Figure 2.5 summarizes the implementation of the compensation and career-ladder lever over time across the Intensive Partnership sites.

Figure 2.5
Proportion of the Compensation and Career-Ladder Lever Implemented, Spring 2010 to Spring 2014



SOURCE: Stecher et al., 2016.

RAND *RR1295/3-1-2.5*

State Policy Changes

As the three districts have been making progress in implementing their reforms, their respective states have also been changing policy to require various reforms statewide. It is important to note these policy changes because they will change the behavior of comparison districts and schools that we use as a yardstick by which to measure an effect.

A complete listing of state policy changes is beyond the scope of this report, although a very comprehensive accounting is available from the National Council on Teacher Quality (NCTQ). In Table 2.1, we provide a summary of the three states' progress in each of the areas that they report that are related to the levers in the Intensive Partnerships initiative. Higher grades correspond to the state implementing policies that are similar to the reforms that the Intensive Partnership sites have undertaken. This information draws heavily on the state policy findings that NCTQ provides for each state on its website (NCTQ, 2015).

We see that Florida tends to have the highest policy grades over the entire period and Pennsylvania tends to have the lowest grades. More important than the average over the period, however, are the changes that occurred during the period that the initiative was in place in the Intensive Partnership sites. For example, during this period, Tennessee made substantial policy changes in the area of exiting ineffective teachers, including the elimination of tenure statewide. On the other hand, Pennsylvania's rating on similar policies remained virtually unchanged. To the extent that Tennessee's policy changes actually led to changes in school and school district practices, we expect the comparison group for MCS to be improving over time. Therefore, it is important to remember that our impact estimates capture changes over and above what would have happened in lieu of the Intensive Partnerships initiative. This means, on the one hand, that our estimates of impact are lower than they might be in the absence of these statewide changes. On the other hand, it also means that some of the improved outcomes in the Intensive Partnership sites might have occurred anyway because of state policy changes.

Averaging all the relevant policies, all three states had grades that improved from 2009 to 2015, with Pennsylvania showing less

Table 2.1
Progress in States' Teacher Quality Policies

Policy	Florida	Pennsylvania	Tennessee
Deliver well-prepared teachers			
2009	C	D+	B–
2011	B–	C	B–
2013	B+	C	B–
2015	B+	C–	C+
Expand the pool of teachers			
2009	B–	C–	C
2011	B–	C	C+
2013	B	C–	C+
2015	B–	C+	C
Identify effective teachers			
2009	C–	D	C
2011	B	D+	B
2013	B+	C	B+
2015	B+	C+	B
Retain effective teachers			
2009	C	D+	C
2011	B–	D+	C
2013	B+	D+	C+
2015	B	D	B–
Exit ineffective teachers			
2009	C	D–	F
2011	B+	F	C
2013	B–	D–	B–
2015	B+	D–	B+

SOURCE: NCTQ, 2015.

improvement than the other two states. Greatest gains were exhibited in policies associated with identifying effective teachers and exiting ineffective teachers, which are related to the Intensive Partnership teacher-evaluation and staffing levers, respectively.[2] All three states made improvements in policies related to identification of effective teachers. Tennessee made, by far, the largest improvement in policy related to exiting ineffective teachers, with Florida making moderate changes and Pennsylvania making little change. Florida made substantial improvements in delivering an effective teaching workforce, with the other two states making little improvement. Florida and Tennessee both made moderate improvements in policies related to retaining effective teachers, with Pennsylvania slightly reducing its policy effectiveness in this area. None of the states made much consistent change in policies related to expanding the teaching workforce. Although we do not know the extent to which any of these policy changes translated into changes in district practice, we would expect the comparison group for PPS to show the least change in practice and the comparison groups for HCPS and MCS to show more improvement in practice. To the extent that these improvements occurred, our impact estimates in HCPS and MCS are likely to be lower than they would have been in the absence of state actions.

---

[2] NCTQ advocates (and its ratings favor) using student achievement—that is, student growth or value-added data—as the preponderant criterion to evaluate teacher effectiveness.

CHAPTER THREE

Data and Methods

In this chapter, we first describe the outcomes on which we estimate the impact of the Intensive Partnerships initiative and the statewide school-level data that we use to measure these outcomes. The rest of the chapter describes the primary and secondary methods that we use to estimate the size of the impact and to calculate the precision of the estimates.

Data and Outcomes

In this report, we use school-level data on student achievement, graduation rates, and dropout rates from Florida, Pennsylvania, and Tennessee to estimate the impact of the Intensive Partnerships initiative.[1] The schools in HCPS, PPS, and MCS form the treatment group, while we use the remainder of schools in the three respective states as the comparison group. All analyses use only publicly available aggregate data. We obtained the data from state department of education websites or by making requests for such data to the state departments of education.[2] We do not include charter schools in the treatment group for any of the districts.

Changes in the composition of MCS in the 2013–2014 school year introduced complications. The most significant change was that

[1] We have not been able to obtain college-going rates.

[2] Data that had to be requested directly from the state departments of education included those for Pennsylvania and Tennessee.

MCS merged with SCS just prior to the 2013–2014 school year.[3] This means that, in 2013–2014, a single district included both the original schools from Memphis that were part of the Intensive Partnerships initiative and all the schools that used to be part of SCS, which did not receive the intervention.[4] If we were to use all 2013–2014 SCS schools in the analysis, a significant portion of the schools in our treatment group would not actually have received the intervention, causing significant measurement error. Therefore, before conducting the analysis, we removed all schools in the merged SCS that used to be part of SCS rather than MCS; we excluded these schools from the analysis for all years.

Another challenge to the MCS analysis is that some schools from MCS have been transferred into a new state organization called the Achievement School District (ASD), which either directly operates the schools or transfers their operation to other groups, including CMOs.[5] These schools were subject to the intervention up until they were transferred to the ASD but not after the transfer. To address the issue of partial exposure to the Intensive Partnerships initiative, we exclude ASD schools from the comparison group in all years and include schools that were originally from MCS in the analysis up to the year they transferred to the ASD.[6]

The main outcome of interest is the school-level average of student scale scores on the state assessments. Because the tests used in high school differ from those used in grades 3 through 8, we report results separately for grades 3 through 8 as a group and for high schools. We also present results separately for mathematics and reading (English language arts). For each subject-year-grade grouping, we standardized

---

3 Further changes to the district boundaries occurred the following year, with many of the suburbs of the old SCS leaving the newly merged district and creating their own districts. While these administrative changes were in process, there was little movement of staff or students across the old or new district boundaries.

4 Schools that were originally part of SCS did not receive funding from the Intensive Partnerships initiative until after the merge.

5 Information and a list of schools can be found at Achievement School District, undated.

6 However, in our analysis, we do include the Memphis Innovation Zone (i-Zone) schools.

the scale scores by the within-state student standard deviation so that we could interpret the estimates in effect-size units of the student-level test score distributions in each state.[7]

In addition to average overall scale scores for grades 3 through 8 and high school, we examined nontest outcomes, such as attendance (for MCS only), graduation rates, and dropout rates. In Appendix B, we report the results for these other outcomes.[8] We also examined results for demographic subgroups. Specifically, we generated results for Hispanic, black, and economically disadvantaged groups by grade and subject when these subgroups make up a sufficient proportion of district population. We also report these results in Appendix B. Table 3.1 lists the outcomes and subgroups for each site.

Having preintervention data is important to control for differences between schools and students in treatment districts and those in the rest of the state. Thus, we collected and used three years of preintervention data, from school years 2006–2007 to 2008–2009.[9] In

---

[7] To understand the effect-size concept, consider a simple example. Suppose that students take a test and that the scale score values for this test range from 100 to 500, with a mean of 300. Without further information, an estimated impact of, say, three points would be uninformative. To make sense of this finding, what is needed is information on how much variation there is in the scale score. The standard deviation of the test scale score is the usual way of measuring this variation. Frequently, test scale scores follow a bell-shaped distribution known as a normal distribution. In this common case, about two-thirds of students score within one standard deviation of the mean (300, in this example), and about 95 percent score within two standard deviations of the mean. The effect size is simply the change in the scale score (for example, three points) translated into standard deviation units. If the standard deviation were 10, the effect size would be $3 \div 10 = 0.3$, indicating that the program increased test scores by 0.3 standard deviations. This would be a meaningful impact, which not many education interventions attain. In contrast, if the standard deviation were 100 and the difference in scale score were three points, the effect size would be a more modest 0.03.

[8] We focus on test scores rather than measures of high school completion because high school completion is a cumulative outcome that we do not expect the reforms to affect for several years. In contrast, it is more plausible that the reforms instituted as part of the initiative could have short-run effects on test scores.

[9] We truncated the preintervention data at the 2006–2007 school year to avoid additional changes in tests and so that predictions from the model would better reflect recent trends and changes in states' testing and school demographics. For example, PPS experienced a major change in demographics between 2006 and 2007 that led to a sharp decline in test scores compared with other schools in the state.

Table 3.1
Summary of Data Elements

Data Element	HCPS	MCS	PPS
Grade 3–8 test scores (unless otherwise noted)	Math and reading	Math and reading	Math and reading
High school test scores	Reading (grades 9 and 10)	None[a]	Reading (grade 11)
Relevant subgroup test score	Black, Hispanic, low socioeconomic status	None[b]	Black, low socioeconomic status
Nontest outcomes	Graduation and dropout rates	Graduation, promotion (K–8), attendance (K–12), and dropout rates	Graduation and dropout rates
Covariates	Ethnicity, ELLs, FRPL, students absent more than 21 days,[c] stability rate,[c,d] average preintervention proficiency levels 1, 2, 3, 4, and 5 for mathematics and reading[c]	Ethnicity; FRPL; average preintervention percentages in proficient, advanced, and below proficient in mathematics and reading[c]	Ethnicity; FRPL; average preintervention percentages in proficient, advanced, basic, and below basic in mathematics and reading[c]

NOTE: ELL = English-language learner. FRPL = free and reduced-price lunch.

[a] Tennessee administers several end-of-course exams in high school. However, these exams can be retaken throughout high school; without being able to separate first-time from retested students' scores, it makes these test scores noisy signals of performance. As a result, we exclude these tests from our analysis.

[b] The state department of education provided the overall test score information to RAND, not broken out by subgroup. Thus, we could not complete these analyses by subgroup.

[c] In this analysis, we used an average of all preintervention years of the variable at the district level.

[d] Stability rate indicates the percentage of the October membership survey still present for the February membership survey.

addition to preintervention outcomes, we used other publicly available school-level covariates in the analysis. Table 3.1 also lists these.

Table 3.2 shows the mean values in test scores and demographics for the Intensive Partnership sites and for the rest of the schools in

Table 3.2
Mean Values for District Average Standardized Test Scores and Demographics

	School Year 2008–2009		School Year 2013–2014	
Variable	Intensive Partnership Site	Rest of State	Intensive Partnership Site	Rest of State
HCPS and rest of Florida				
Standardized test scores				
Math, grades 3–8	−0.04	−0.01	−0.03	−0.01
Reading, grades 3–8	−0.07	−0.01	−0.04	−0.01
Reading, high school	−0.03	−0.01	−0.05	−0.01
Percentage of students who are				
Black	0.22	0.23	0.22	0.23
Hispanic	0.28	0.35	0.34	0.30
Asian	0.03	0.02	0.03	0.02
Receiving FRPL	0.52	0.49	0.60	0.54
ELLs	0.15	0.11	0.12	0.09
PPS and rest of Pennsylvania				
Standardized test scores				
Math, grades 3–8	−0.33	0.03	−0.28	0.03
Reading, grades 3–8	−0.35	0.03	−0.35	0.03
Math, high school	−0.32	−0.03	−0.26	−0.03
Reading, high school	−0.32	−0.01	−0.19	−0.02

Table 3.2—Continued

Variable	School Year 2008–2009		School Year 2013–2014	
	Intensive Partnership Site	Rest of State	Intensive Partnership Site	Rest of State
Percentage of students who are				
Black	0.56	0.14	0.53	0.14
Hispanic	0.01	0.07	0.02	0.09
Asian	0.02	0.03	0.03	0.03
Receiving FRPL	0.69	0.40	0.66	0.40
MCS and rest of Tennessee				
Standardized test scores, grades 3–8				
Math	–0.39	0.05	–0.47	0.04
Reading	–0.51	0.07	–0.47	0.04
Percentage of students who are				
Black	0.86	0.17	0.80	0.15
Hispanic	0.06	0.05	0.11	0.08
Asian	0.01	0.02	0.01	0.02
Receiving FRPL	0.79	0.52	0.85	0.59

SOURCE: States' departments of education.

NOTE: We used the following formula to calculate standardized average test scores for each school, grade, and subject: (average school test score – average test score in state) ÷ standard deviation of individual test scores in state. Then we averaged scores within school and subject across grades weighted by grade enrollment. Finally, we averaged these school averages across schools in the Intensive Partnership site and in the rest of the state, weighted by school enrollment. This procedure implies that the sum of the average in the Intensive Partnership site and for the rest of the state does not necessarily equal 0. We calculated demographic variables at the school level by dividing the number of students from a certain category by the total number of students in the school and then averaging across schools in the Intensive Partnership site and in the rest of the state based on student enrollment by school.

the same states. It is important to note that Intensive Partnership sites, with the exception of HCPS, had much larger fractions of students from minority ethnicities than the other districts in their states. Every site has a larger fraction of students in poverty (i.e., those who qualify for FRPL) than the other districts in its state. HCPS also has a higher fraction of students who are ELLs than the rest of Florida.

It should also be noted from Table 3.2 that the Intensive Partnership sites performed worse in math and reading than the rest of the schools in their states in the 2008–2009 school year, before the start of the Intensive Partnerships initiative. By 2013–2014, they were still lagging behind, but, in most cases, the gaps have narrowed. In the rest of this chapter, we describe the empirical methods we employ to discern whether the improvement in the Intensive Partnership sites, relative to the rest of schools in their states, can be attributed to the Intensive Partnerships initiative.

School-Level Difference-in-Differences Methodology

Estimating the program's impact is difficult because the outcomes in the Intensive Partnership sites could differ from those in non–Intensive Partnership sites for reasons other than the Intensive Partnership program itself, such as students in Intensive Partnership sites being less affluent than students in other sites. As shown in Table 3.2, there are clear differences between the distributions of characteristics of students served by schools in the Intensive Partnership sites and those of students in other schools in the same state that are not in the Intensive Partnership sites. To the extent that these differences drive differences in student outcomes, comparisons between the outcomes of students in schools in the Intensive Partnership sites and those in the non–Intensive Partnership sites will be misleading as to the impact of the intervention.

To disentangle the intervention's effects from the effects of student characteristics and other district-specific factors, we employ a DiD approach using school-level data. This approach involves two steps. The first step uses data on school-level outcomes and on demographic characteristics (at the school and district levels) in the preintervention years to forecast what school outcomes are likely to be in the postinterven-

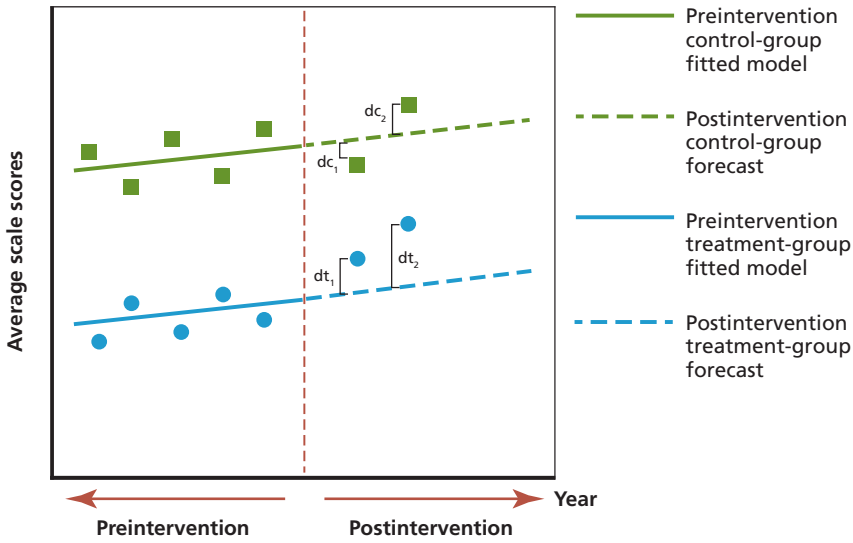
tion years, taking into account any changes in demographic characteristics (at the school and district levels).[10] In the second step, we examine whether differences between the actual outcomes and the forecasted outcomes systematically differ between schools in an Intensive Partnership site and those in the same state's non–Intensive Partnership sites. This DiD can be interpreted as the gap between the performance of schools in Intensive Partnership sites and non–Intensive Partnership schools, net the difference that would be expected given the preintervention outcome patterns and differences in demographics.

The hypothetical example in Figure 3.1 depicts how the first step in this procedure works. The figure shows the relationship between some school-level outcome (average scale scores, in this example) and time. Data points to the left of the red dashed line are from years prior to the Intensive Partnerships initiative, and data points to the right are from years after the Intensive Partnerships initiative went into place. In this example, there are very large differences in the preintervention years between the treated and comparison schools, shown by the difference in the height of the lines along the vertical axis.

To account for the differences between the treated and nontreated schools, our method uses data from before the start of the intervention to form a prediction of what the counterfactual outcomes would be in the postintervention world. This forecast is based on a statistical model that uses preintervention data to estimate linear predictions for postintervention years. We then use the predictions to determine what the outcomes likely would have been had school and district demographics continued to have the same effect on outcomes as they did before the intervention. We depict the preintervention data graphically as squares (for the control group) and circles (for the treatment group) to the left of the dashed red line in Figure 3.1. The solid lines represent the fit of

[10] In previous interim analyses, we have also tried first trimming the sample of non–Intensive Partnership schools to keep only those schools that are more similar in terms of demographics to the schools in the Intensive Partnership site. After selecting these schools, we followed the two estimation steps described here. The estimation results with the trimmed sample were very similar to the results obtained using the full sample of schools in the state (which we present in this report). This suggests that a linear specification does a relatively good job in controlling for differences in observed characteristics.

Figure 3.1
Graphical Depiction of Methodology for Computing Forecasts of Postintervention Trends



RAND *RR1295/3-1-3.1*

the statistical model, and the dashed lines depict the forecasts of the model.

We then compute the difference between what the forecasting model predicts that the outcome will be and the actual outcome. In Figure 3.1, these deviations from this forecast for the first two years after the intervention are equal to the vertical distance from the forecasted outcome (i.e., the dashed line) and the actual outcome. For the comparison group, these deviations are dc_1 and dc_2; for the treatment group, they are dt_1 and dt_2. In this example, the comparison group does a little better than predicted in one year and a little worse in the other. In contrast, the difference between the actual and predicted treatment-group performance is large and positive in both years.

The second step of our method consists of estimating whether these differences are systematically and statistically different between schools in an Intensive Partnership site and those in the comparison non–Intensive Partnership sites. This difference in prediction differ-

ences (or prediction errors) provides our DiD estimation of the Intensive Partnerships initiative's impact. It can be interpreted as the difference in performance between schools in treated districts and other schools in the state, after netting out the difference that would be expected given preintervention outcome patterns and school and district demographics.

To implement this two-step DiD analysis, we use a multivariate regression procedure. Appendix A describes this procedure in detail. In the first step, we fit a linear regression of the school outcomes on a set of demographic variables listed in Table 3.1 measured both at the school and district levels and on indicator variables for each district.[11] To fit this regression, we use information from the preintervention years only (i.e., up to the 2008–2009 school year). Then, using this model, we predict school outcomes for each year (both in the pre- and postintervention periods) using the year's observed school and district demographics and district-specific estimated intercepts. Then, we calculate the differences between the actual outcomes and the predicted values for each school for each year.

In the second step, we fit a separate linear school-level regression for each year of these differences on an indicator variable for whether a school is in the treatment (Intensive Partnership) district or in a non–Intensive Partnership site, on school and district demographic variables, and on a district-level random component (or random effect).[12] We allow the coefficients on the treatment indicator variable to vary with time by estimating separate regressions for each year in the pre- and postintervention periods. This approach allows examining whether a different trend between the treatment and comparison schools existed in the preintervention period. Ideally, the coefficient for the treatment indicator variable in the preintervention period would be close to 0 and

[11] These district indicators control for district-specific characteristics that do not change over time.

[12] In previous interim analyses, we estimated models that did not control for school and district demographic variables in the second step. The estimated effects were very similar to the ones presented in this report, indicating that statewide influence of demographics characteristics on average school performance have been relatively stable over time, at least in the states of Florida, Pennsylvania, and Tennessee.

not systematically trending upward or downward. This would imply that treatment and comparison schools shared a common trend in their outcomes prior to the Intensive Partnerships initiative and any differential deviations from that trend for the treatment schools can be causally interpreted as the initiative's effect. Our method also allows examining whether the treatment effects in the postintervention period change over time. One important hypothesis is that the Intensive Partnerships initiative will take time to generate effects because the reforms it entails require several years to implement. Our empirical evaluation strategy allows us to test this hypothesis.

A significant empirical challenge is to determine whether the usual variability in outcomes that occurs across districts could explain the initiative's estimated impacts. The district-by-time–level random-effect component included in the analysis addresses this problem. We explicitly model that the common shocks to schools' performances in a district in a given year, which would occur regardless of the Intensive Partnerships initiative, follow a normal distribution. Adding this district-year random-effect component to the model allows us to measure the natural variability in outcomes across districts. This allows us to judge whether the initiative's estimated impacts are large enough in comparison with the expected variation in the absence of any intervention. The drawback to this random-effect approach is that it makes very strong assumptions about the way the common shocks are distributed across districts (in each year) that might not be warranted.[13]

We next discuss an alternative methodology for estimating the initiative's effects that makes weaker assumptions about the distributions of the errors but also have some limitations.

Synthetic-Control-Group Methodology

We implemented an additional analysis that uses the SCG methodology to test the intervention's effect. This methodology was first intro-

[13] For example, we assume that district shocks are normally distributed, uncorrelated with observed district and school characteristics, and independent over time.

duced in Abadie and Gardeazabal, 2003, and further developed in Abadie, Diamond, and Hainmueller, 2010. The SCG methodology uses information at the level of the intervention. In our case, we use information aggregated at the school district level. The central idea behind the SCG approach is to construct an SCG made up of weighted observations from other comparison districts. The weights are created so that the weighted average of the comparison district looks as similar as possible to the treatment district in its preintervention characteristics and outcomes.[14] By creating a group that looks just like the treatment group before the intervention, we can be more confident that any differences between the SCG and the treated group are due to the intervention. However, the SCG method cannot construct a similar comparison group if no similar groups actually exist in the data. When a similar comparison group cannot be constructed, it is difficult to interpret the results from an SCG analysis.

The SCG methodology presents an important advantage regarding statistical inference. By working with data aggregated at the same level as the intervention, i.e., the school district, we do not need to model shocks that are common to all schools in a district (such as a change in superintendent) like we did with the DiD methodology. Furthermore, the SCG methodology uses permutation tests (Bertrand, Duflo, and Mullainathan, 2004; Abadie, Diamond, and Hainmueller, 2010) to conduct inference. The idea behind this method is to compare the estimated impact of the intervention to a distribution of placebo effects that are obtained by repeatedly redoing the same analysis but each time using a different comparison district as a placebo treatment site. The distribution of placebo effects mimics the variability in the estimates that would occur naturally because of unobserved factors. If the actual estimate is larger than this natural variability, the estimate is deemed to be statistically significant. The main appeal of the permuta-

[14] The donor pool of comparison districts is trimmed before constructing the weights so that it does not include any district that is very different from the treatment district. We dropped districts where the average enrollment (2005–2014) was less than 1 percent of the average enrollment in the treatment district. We also dropped districts where the average percentage of minority students (2005–2014) was less than 10 percent of the average percentage in the treatment district.

tion tests is that they make no assumptions about the distribution of the placebos but rather use the empirical distribution that the data provide. In comparison, the DiD approach assumes that the distribution of districtwide shocks follow a normal distribution. The drawback of not imposing a distribution assumption is that permutation testing can be quite conservative. In other words, it requires that the impact sizes be relatively large to be considered statistically significant.

We regard the DiD method as our main approach, but we report the results from the SCG method as a second opinion or robustness check. Many reasons led us to choose the DiD method as the main approach, although both methods delivered similar impact estimates.[15] First, the DiD method allows controlling for changes in demographics and in their effects on outcomes, either of which could occur unrelated to the reforms.[16] Conversely, in the SCG methodology, any difference between the treatment group and the SCG after the intervention begins is regarded as a direct effect of the intervention. In certain circumstances, interventions lead to changes in the student body composition, so it should be factored into the impact estimates. This is not the case with the Intensive Partnerships initiative, which did not have any specific goal regarding the composition of students in the participating districts.

Second, the DiD method also allows controlling for changes in the composition of the districts. This was an important issue in the

[15] Another alternative is to work with student-level outcomes and match each student in an Intensive Partnership site with an observationally identical student elsewhere in the state. This would entail accessing confidential individual data from each of the states in which the Intensive Partnership sites are located, going back to the 2006–2007 school year. This approach is costly in terms of negotiating access to the information, data storage, and computational effort, a fact that limits its reproducibility. Moreover, the immediate benefits of the approach are not obvious because we control for all the demographic information that one would use to find a suitable match or control for each student. Because of these issues, we opted for working with publicly available data that can be obtained from state department of education websites or by making requests for such data to the departments of education.

[16] For example, if one group of students in a state—e.g., low-income students—saw a greater gain than the other students, perhaps because of a state policy or economic conditions, an individual district might see a better results than another district if it had a higher proportion of these students, even without any reform effects.

analysis of MCS. As described earlier, just prior to the 2013–2014 school year, MCS merged with SCS. Thus, the new unified district included a significant portion of schools that did not actually receive the intervention, a fact that would result in a contaminated treatment group and biased estimates of the initiative impacts. By working with school-level data, we could retain only the schools that were originally part of MCS and exclude schools that used to be part of SCS from the analysis in all years (and exclude them from the control group as well).

Third, in some instances, the SCG method cannot construct a similar SCG. This is a problem we encountered in MCS because the demographics in this district are very different from those in other districts in Tennessee (see Table 3.2). Regarding this point, the DiD estimator has an advantage because it can accommodate any differences between the treatment group and the comparison group in the preintervention period. The working assumption is that those differences would remain constant over time in the absence of an intervention.

Finally, regarding inference, the SCG method offers the advantage of making no assumptions regarding the distribution of the natural variability in districtwide results versus the normal distribution that the random-effect estimator assumes in the DiD method. However, in simulations, we have found that the random-effect estimator can be an effective way of controlling for districtwide effects, even in circumstances in which they do not follow a normal distribution.

CHAPTER FOUR

Results

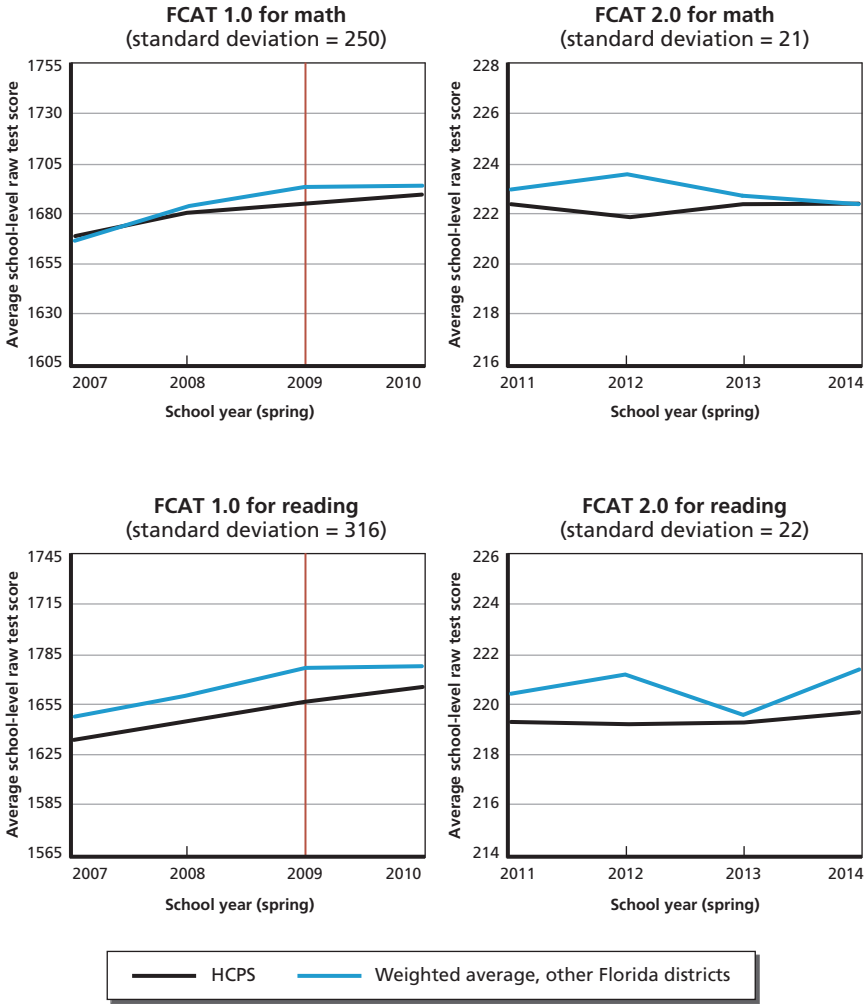
We now present results for each of the Intensive Partnership sites. We show results side by side using both the DiD and SCG methods. Consistent results across the methods should provide more confidence that our findings are not sensitive to the particular methods used. We focus on one key academic outcome: average standardized student scale scores on the state assessments. For ease of exposition, we have averaged standardized student scores in lower grades (3 through 8) and in high school (9 through 12) for each subject and for each school (weighted by enrollment size in each grade). Appendix B contains separate results for each grade and results for other outcomes and subsamples using the DiD methodology.

Hillsborough County Public Schools in Florida

Figure 4.1 shows the average test scores for grades 3 through 8 in mathematics and reading for HCPS and for the rest of the districts in Florida, separately from 2007 to 2010 and from 2011 to 2014 because there was a change in the Florida Comprehensive Assessment Test (FCAT) in 2011.[1] The figure shows that the average scores in HCPS are below

[1] During the end of this grade range, many students in HCPS and other Florida districts also take end-of-course (EOC) exams if they are enrolled in courses for which Florida requires an EOC exam (e.g., algebra). However, these students also take the FCAT end-of-grade exams, so we include them in our analysis. However, to the extent that some schools and districts have more students taking advanced courses, these students might have not been recently exposed to the content that the end-of-grade exams cover. Our analysis does

Figure 4.1
Average School-Level Test Scores on Grade 3 Through 8 Mathematics and Reading, Hillsborough County and Other Florida Districts



NOTE: The solid vertical red line at 2009 indicates the final year before the grant was announced and funding commenced. We weighted school-level test scores by school enrollment.

RAND *RR1295/3-1-4.1*

the average scores in the rest of Florida for most years in the preintervention period. Also, the average scores in HCPS followed roughly the same trend as the average test scores in the rest of the state prior to the intervention. After the intervention, average scores in HCPS follow a slight upward trend, while average scores in the rest of the state seem more erratic, with increases in some years and decreases in others. In general, by 2014, the gap between HCPS average scores in math and those in the rest of the state had closed, whereas, in the case of reading, it is difficult to evaluate visually whether the gap has increased or decreased from its preintervention levels.

Next, we use the DiD and SCG methodologies to evaluate whether we can attribute these gains to the intervention and whether they are large enough to be deemed statistically significant. Figures 4.2 and 4.3 show the Intensive Partnerships initiative's estimated effect sizes (in standard deviations) in HCPS for grades 3 through 8 in mathematics and reading, respectively. The left panel shows the effects using the DiD methodology, while the right panel shows the results using the SCG methodology. The horizontal axis represents the school year, with a solid vertical line at 2009, the final year before the funding for the Intensive Partnerships initiative was issued and the final year we use to calculate the preintervention model. The solid black lines depict the estimated effect size, and the magnitudes (along the vertical scale) are in standard deviations of the student test score distribution. Values greater than 0 indicate that the deviations between the forecasted test scores and actual test scores tend to be larger in HCPS than in comparison schools. The dashed blue lines in the DiD graphs depict the 95-percent confidence intervals from the random-effect model. We consider an estimated effect statistically significant if 0 is not included in the confidence interval. The vertical red and green lines in the SCG graphs show the 95-percent range (i.e., 2.5th percentile to 97.5th percentile) and the 75-percent range (i.e., 12.5th percentile to 87.5th percentile), respectively, of effects that could happen by chance, as estimated by the conservative placebo-test inference strategy. Notice that

---

not account for any changing tendency of HCPS students to take courses that do not cover the end-of-grade exam content.

the distribution of placebo effects does not need to be centered on the estimated effects (i.e., the values plotted in black) and is expected to be centered on 0.[2] For inference purposes, we can characterize estimated effects that are outside of the placebo ranges as being statistically significant either at 95-percent or 75-percent confidence, respectively. In other words, estimated effects that lie outside these intervals are large enough that we can reject the hypothesis that they happened because of natural variation in scores and thus can be attributed as an effect of the program. All figures depicting effects on test scores in this report have the same format.

The purpose of these analytic strategies is to estimate the size of the change that the Intensive Partnerships initiative causes; however, it is challenging to make causal inferences without actually conducting a randomized experiment. The DiD and SCG methods can support strong causal inference under the right conditions, and we can use these figures to assess the extent to which those conditions apply. In the case of the DiD methodology, the main assumption is that, if the intervention had not happened, academic achievement in the Intensive Partnership site in the school years after 2008–2009 would have followed the same trend as in the other districts in the state. Although we cannot test this assumption, we can instead test whether, *prior to the intervention*, the Intensive Partnership sites and the other districts shared a common trend. Graphically, a common trend would imply that the estimates for the preintervention period (school years 2006–2007 to 2008–2009) oscillate around 0 and are not statistically significant. Conversely, if the estimates show a clear pattern (and are statistically different from 0), student achievement at the Intensive Partnership site had a different trend from that in the rest of the state, which makes the assumption

[2] We obtain the placebo effects by redoing our analysis using each comparison district as a placebo treatment site. In other words, we falsely assume that the schools in the placebo districts have received the intervention, and we perform the same empirical analysis. The collection of all placebo effects approximates the potential natural variation in students' proficiency measures that we could expect to occur in the absence of the Intensive Partnerships initiative. The effect calculated in the Intensive Partnership site can fall anywhere on this distribution. In fact, if the estimated effect for the Intensive Partnership site falls in the tails of the distribution, we can conclude that the intervention had a statistically significant effect because we would be unlikely to obtain an estimate of that size caused by natural chance.

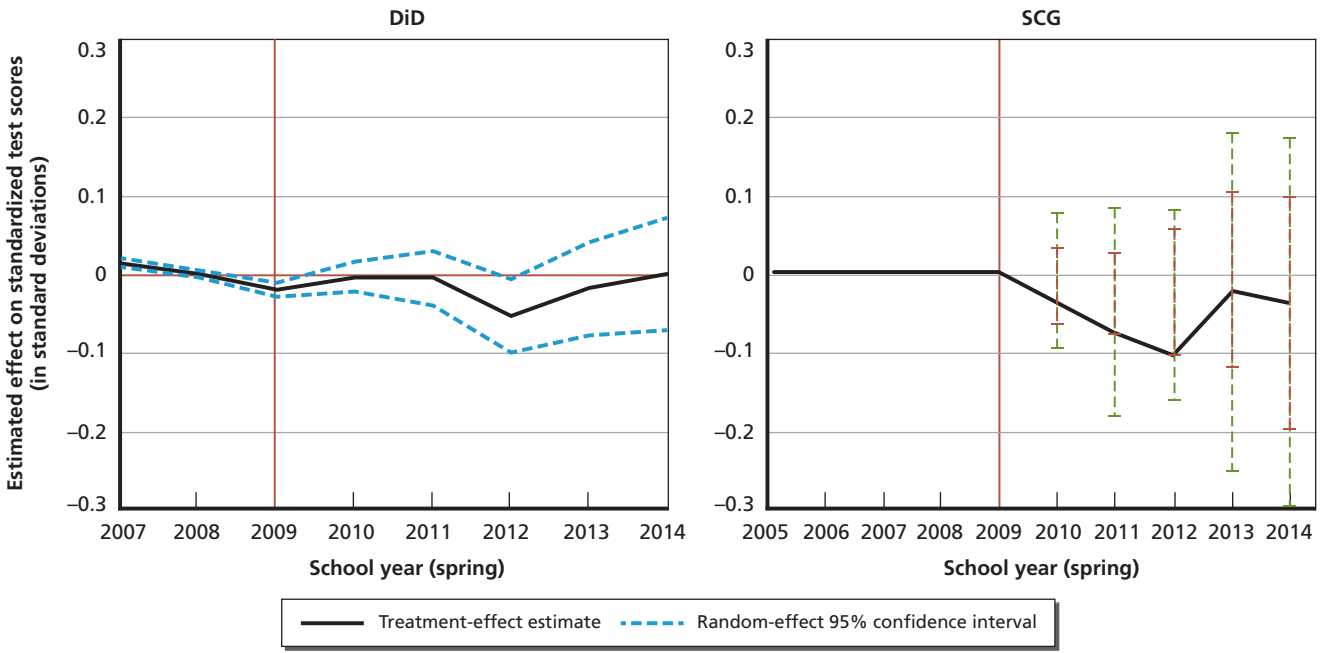
of common trends after school year 2008–2009 (in the absence of the intervention) less plausible. Figures 4.2 and 4.3 show that, prior to the intervention, HCPS had negative trends in math and reading achievement in lower grades relative to other schools in Florida. Although the estimates are statistically significant, the slopes are very mild and thus are not a source of concern about the validity of the DiD methodology.

In the case of the SCG method, we can judge the method's validity graphically by the differences between the achievement at the Intensive Partnership site and at the SCG in the years prior to the intervention. The closer these differences are to 0, the more likely that the SCG would serve as a benchmark of what would have happened in the Intensive Partnership site in the absence of the intervention. In other words, it is more likely that the estimates for the school years after 2008–2009 can be regarded as the causal effect of the intervention. Figures 4.2 and 4.3 indicate that we could construct an SCG that very closely resembled the achievements in HCPS for both math and reading in the preintervention period, ensuring that we can causally interpret the results of the SCG methodology.

Figures 4.2 and 4.3 indicate that HCPS experienced lower achievement gains than comparison schools after the onset of the initiative. However, using the DiD methodology, we observe an increase in the effects of the intervention in recent years, with a positive estimate in 2014 of around 0.03 standard deviations for reading in the lower grades (3–8). However, given the relatively large standard errors, this effect is not statistically significant. In the case of mathematics, in recent years, we also observe an increase in the effects, reaching a positive but small (close to zero) and not statistically significant effect in 2014. The SCG provides similar results: a small improvement in the case of reading in 2014 in the lower grades and an improvement in math achievement (although not statistically significant) when compared with 2012.

A noteworthy pattern, and one that we see in other locations as well, is that the range of the estimates that can be considered as natural variations, based on the placebo treatments, is quite wide. In contrast, the confidence intervals from the random-effect model are much narrower. However, as discussed earlier, the random-effect confidence

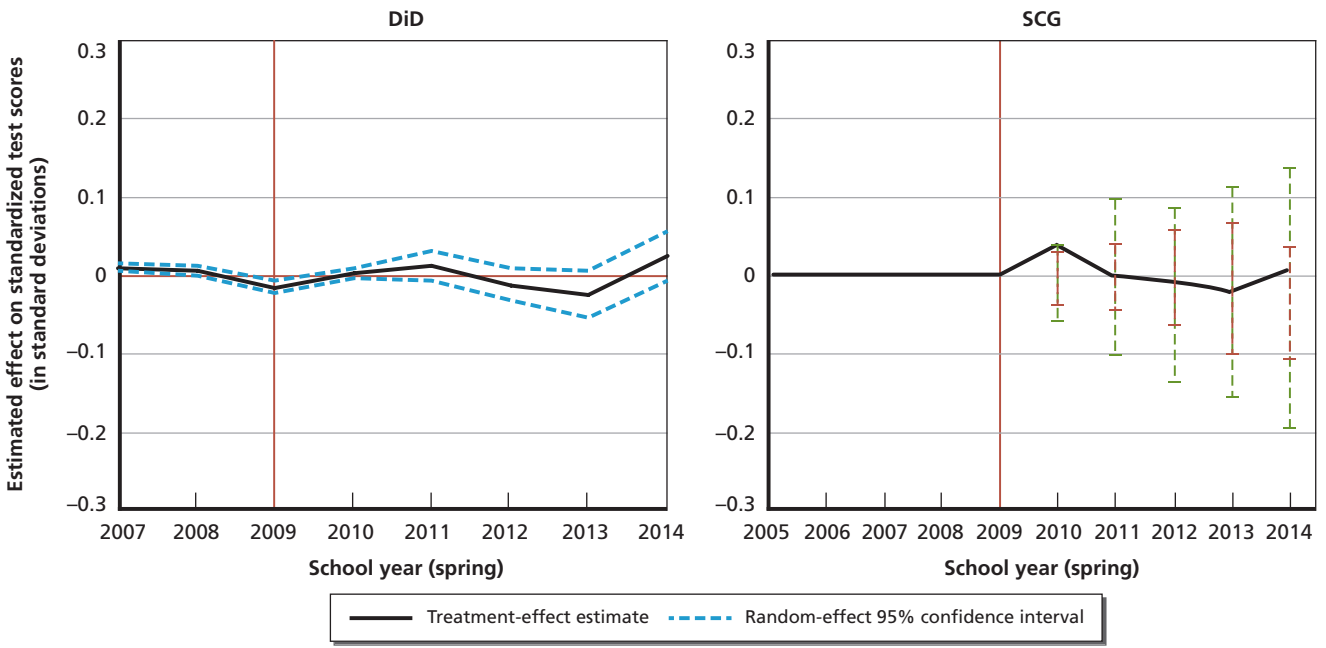
Figure 4.2
Estimates of the Intensive Partnerships Initiative's Effect on Grade 3 Through 8 Math, Hillsborough County, Florida



NOTE: The solid vertical red line at 2009 indicates the final year before the grant was announced and funding commenced. Dashed blue lines in the DiD graph depict the 95-percent confidence intervals from the random-effect model. The vertical red and green lines in the SCG graph depict the 95-percent and 75-percent ranges, respectively, of the placebo-effect distribution.

RAND *RR1295/3-1-4.2*

Figure 4.3
Estimates of the Intensive Partnerships Initiative's Effect on Grade 3 Through 8 Reading, Hillsborough County, Florida



NOTE: The solid vertical red line at 2009 indicates the final year before the grant was announced and funding commenced. Dashed blue lines in the DiD graph depict the 95-percent confidence intervals from the random-effect model. The vertical red and green lines in the SCG graph depict the 95-percent and 75-percent ranges, respectively, of the placebo-effect distribution.

RAND *RR1295/3-1-4.3*

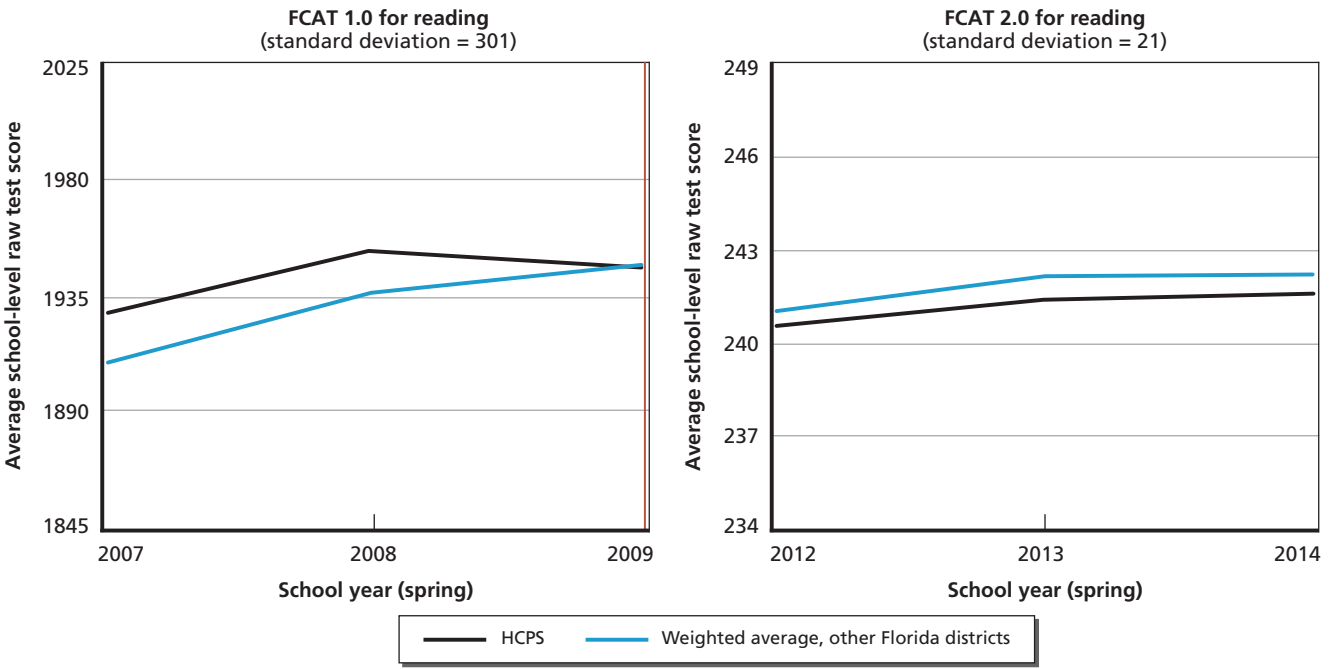
intervals are based on stronger statistical assumptions. On the other hand, though, only very large treatment effects would be detectable using the more conservative placebo method in the SCG graphs. In fact, the placebo confidence intervals are so wide that it would be very unlikely for any reasonable program impact to be detectable. Therefore, the graphs using the two methods should be viewed in conjunction because similar patterns provide evidence of some effect even if the effects are not statistically significant in both graphs. We also evaluate whether the estimated effects are substantively significant in comparison with other interventions in education. We defer that analysis to Chapter Five.

We also estimated the Intensive Partnership initiative's impact on student achievement separately for black students, Hispanic students, and economically disadvantaged students, using the DiD methodology. (Appendix B presents detailed results.) In 2014, the initiative's effects on math in the lower grades (3–8) are not statistically significant for any subgroup (as was the case overall). Regarding reading in the lower grades, the initiative had a significant effect in 2014 for economically disadvantaged students (0.04 standard deviations), which was larger than the effect estimated for the overall population (0.03 standard deviations). The effects for the other subgroups (black and Hispanic students) were not statistically significant and were similar in magnitude to the overall effect.

Figure 4.4 shows the average scores in high school reading in HCPS and in the rest of the state, and Figure 4.5 shows the estimated impact of the Intensive Partnerships initiative.[3] Figure 4.4 does not include school years 2009–2010 and 2010–2011 because there were three large changes in the scale of the scores, from 2010 to 2012, and including the intermediate years complicates the scale of the figures and would not allow us to detect the trends in HCPS and in the rest

[3] In Florida, prior to school year 2010–2011, the main exam for the state was the FCAT, which tested mathematics and reading in grades 3 through 10. During the 2010–2011 school year, the state switched tests to the FCAT 2.0 and the Florida EOC assessments. The FCAT 2.0 continued to test students in reading through grade 10. However, the EOC exams test subject-specific math content in high school. As a result, for consistency with the prior-period testing, we have excluded high school mathematics from our analysis.

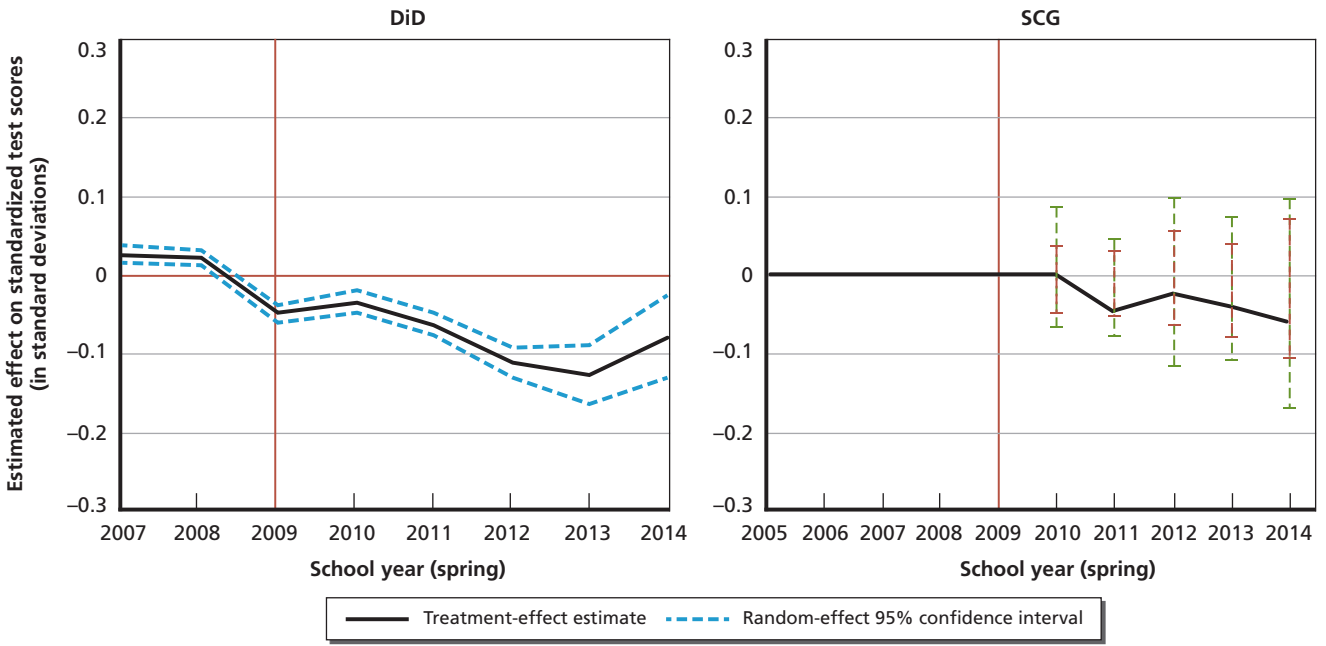
Figure 4.4
Average School-Level Test Scores on High School Reading, Hillsborough County and All Other Florida Districts



NOTE: The solid vertical red line at 2009 indicates the final year before the grant was announced and funding commenced. We weighted school-level test scores by school enrollment.

RAND *RR1295/3-1-4.4*

Figure 4.5
Estimates of the Intensive Partnerships Initiative's Effect on High School Reading, Hillsborough County, Florida



NOTE: The solid vertical red line at 2009 indicates the final year before the grant was announced and funding commenced. Dashed blue lines in the DiD graph depict the 95-percent confidence intervals from the random-effect model. The vertical red and green lines in the SCG graph depict the 95-percent and 75-percent ranges, respectively, of the placebo-effect distribution.

RAND *RR1295/3-1-4.5*

of Florida. Nevertheless, by inspecting Figure 4.4, we observe that average high school reading scores in HCPS were initially higher than those in the rest of the state. However, this situation changed by 2009; thereafter, HCPS underperformed the rest of the state in high school reading. In fact, Figure 4.5 shows that the post-2010 estimated impacts are negative, using both the DiD and SCG methods. Using the DiD methodology, we also found negative and statistically significant effects on high school reading for black students, Hispanic students, and economically disadvantaged students (see Appendix B), with a somewhat larger negative effect for economically disadvantaged students.

In addition to our impact estimation for these three achievement outcomes, we estimated the effect on the dropout rates for high schools.[4] We found a positive impact of more than 3 percentage points on the dropout rate in 2013, i.e., the dropout rate increased, with small and insignificant impact estimates for other years. Appendix B contains more details about all the impact estimates.

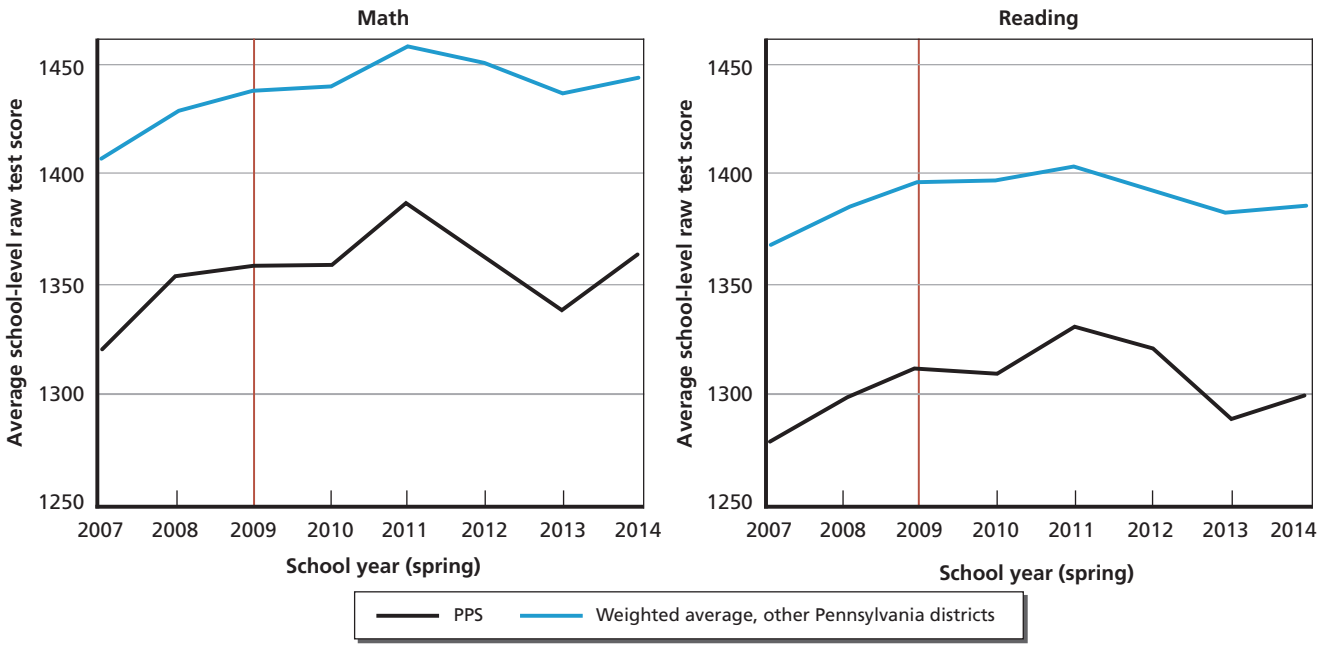
Pittsburgh Public Schools, Pennsylvania

Figure 4.6 shows the average grade 3 through 8 school-level test scores in PPS and in the rest of the state. There was a general upward trend in mathematics and reading achievement, in PPS and in the rest of the state, up to 2011. Then there was a general decrease in scores in 2012 and 2013 and a recovery in 2014.

We apply the DiD and SCG methods to investigate whether PPS outperformed or underperformed districts in the rest of the state after the implementation of the Intensive Partnership initiative. Figures 4.7 and 4.8 show the results for mathematics and reading, respectively, for grades 3 through 8. For math, we found small and statistically insignificant effects for most years after the intervention, with the exception of 2014, which shows larger estimated effects using both methods. The

[4] We chose not to estimate the impact on graduation rates because of a change in how these were calculated after the intervention. This change in calculation formula had different effects for different districts.

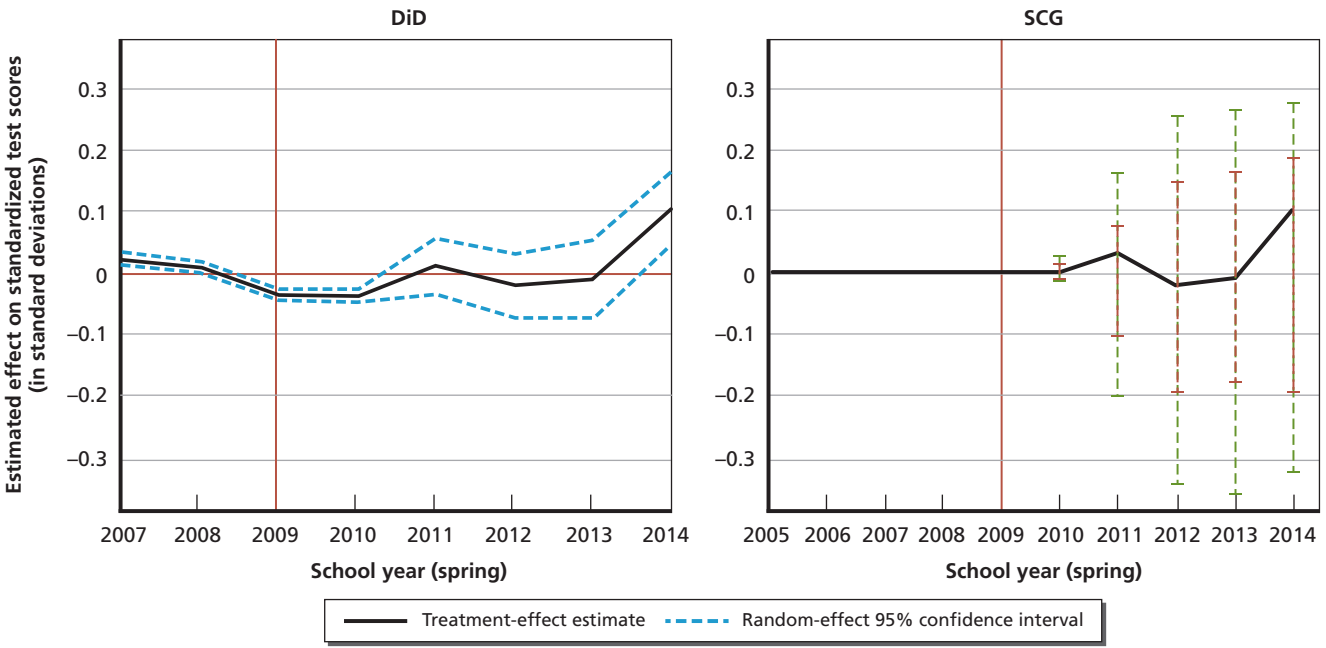
Figure 4.6
Average School-Level Test Scores on Grade 3 Through 8 Mathematics and Reading, Pittsburgh and All Other Pennsylvania Districts



NOTE: The solid vertical red line at 2009 indicates the final year before the grant was announced and funding commenced. We weighted school-level test scores by school enrollment.

RAND *RR1295/3-1-4.6*

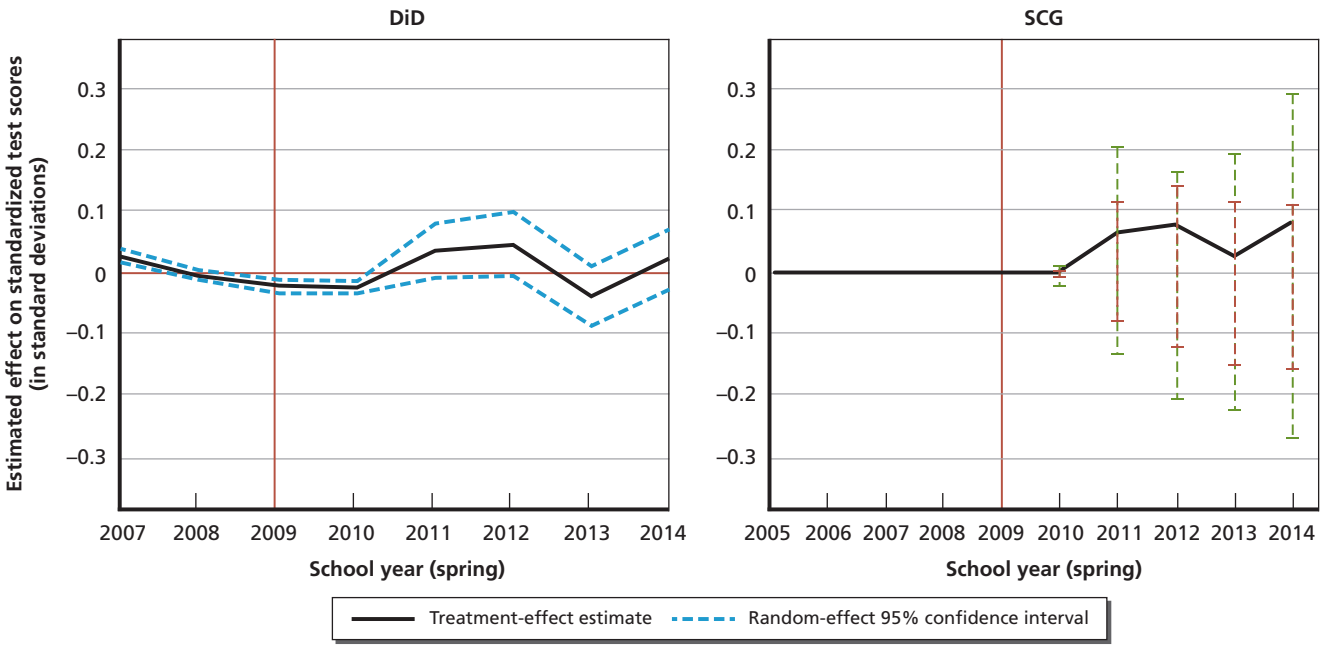
Figure 4.7
Estimates of the Intensive Partnerships Initiative's Effect on Grade 3 Through 8 Math, Pittsburgh, Pennsylvania



NOTE: The solid vertical red line at 2009 indicates the final year before the grant was announced and funding commenced. Dashed blue lines in the DiD graph depict the 95-percent confidence intervals from the random-effect model. The vertical red and green lines in the SCG graph depict the 95-percent and 75-percent ranges, respectively, of the placebo-effect distribution.

RAND *RR1295/3-1-4.7*

Figure 4.8
Estimates of the Intensive Partnerships Initiative's Effect on Grade 3 Through 8 Reading, Pittsburgh, Pennsylvania



NOTE: The solid vertical red line at 2009 indicates the final year before the grant was announced and funding commenced. Dashed blue lines in the DiD graph depict the 95-percent confidence intervals from the random-effect model. The vertical red and green lines in the SCG graph depict the 95-percent and 75-percent ranges, respectively, of the placebo-effect distribution.

RAND *RR1295/3-1-4.8*

DiD method delivers an estimated effect of 0.10 standard deviations in 2014, which is statistically significant.[5] The SCG method estimates a similar effect for 2014 (0.10 standard deviations), although it is not statistically significant. Using the DiD methodology, we also found positive and statistically significant effects in mathematics in grades 3 through 8 in 2014 for black students and economically disadvantaged students; these effects are slightly larger than the overall effect (see Appendix B).

For reading in grades 3 through 8, we observe positive effects in the second and third years after the intervention, a pattern that is also apparent in the average test scores in Figure 4.6. The gap in the average scores between PPS and districts in the rest of the state was reduced after 2009 until a drop-off in 2013. Consequently, using the DiD methodology, we estimate positive effects of 0.03 standard deviations in 2011 and 0.05 standard deviations in 2012. Both methodologies found a reduction in the estimated effects in 2013 and a rebound in 2014. The estimated effects in 2014 are 0.02 standard deviations using the DiD method and 0.08 standard deviations using the SCG method. These effects are not statistically significant given the wide confidence intervals in the case of the DiD method and the wide range of the placebo distribution in the case of the SCG method.

Using the DiD methodology, we also found positive effects on reading in grades 3 through 8 for black students in 2011 (0.03 standard deviations), 2012 (0.07 standard deviations), and 2014 (0.08 standard deviations), with the last two effects being statistically significant. We also found positive effects for economically disadvantaged students in 2011 (0.06 standard deviations) and 2014 (0.03 standard deviations), with the first effect being statistically significant (see Appendix B).

Figure 4.9 shows average high school reading school-level test scores. We present the average scores from 2007 to 2012 (left panel) and from 2013 to 2014 (right panel). In 2013, Pennsylvania intro-

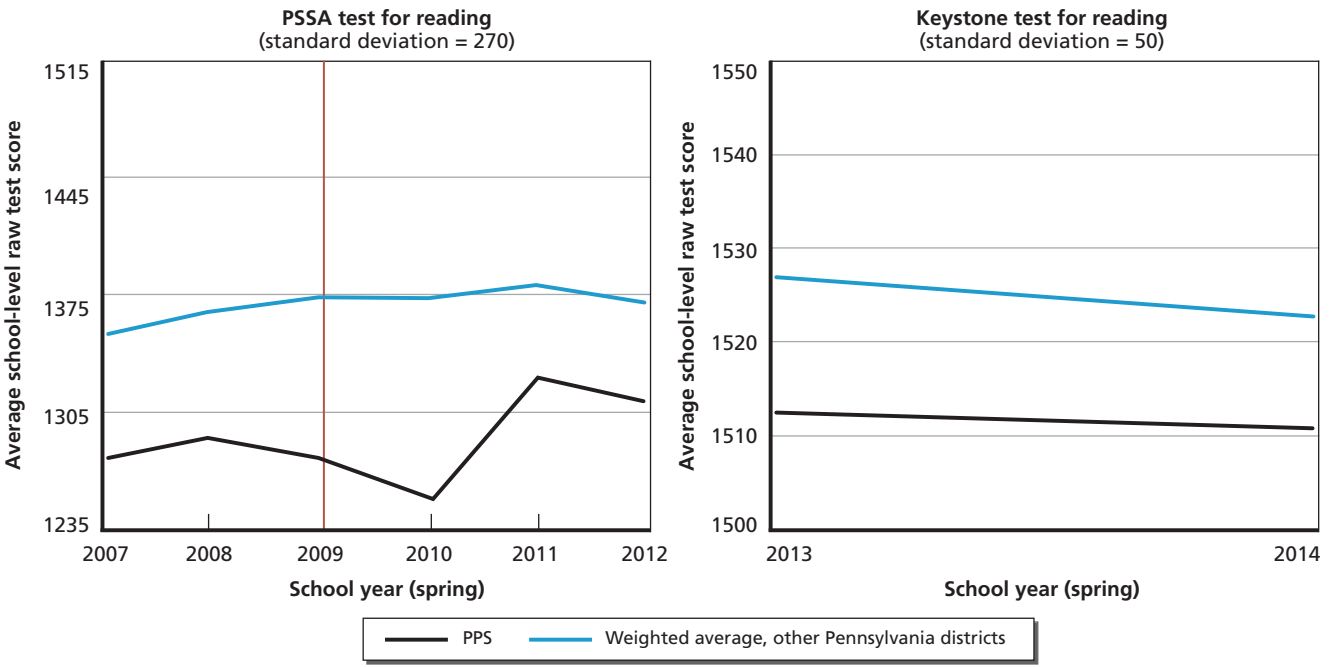
[5] The *p*-value of the DiD effect size for lower grades' math in school year 2013–2014 is 0.001. This effect remains significant after using the Bonferroni correction to maintain an overall 5-percent significance level for the group of five estimates for the postintervention years.

duced the Keystone Exams, which replaced the Pennsylvania System of School Assessment (PSSA) in high schools.[6] The new tests have a smaller standard deviation than the PSSA does, making a similar difference expressed in scores more meaningful in terms of achievement. This makes it difficult to assess from Figure 4.9 whether high schools in PPS improved in 2013 versus 2012 more than districts in the rest of the state. We address this issue by analyzing standardized test scores with our DiD and SCG methodologies. Our findings are mixed with respect to the effects of the initiative on high school reading in PPS (Figure 4.10). On the one hand, using the DiD methods, we found that PPS experienced lower achievement gains in high school reading than comparable schools after the start of the initiative. By 2014, the DiD method found a statistically significant effect of –0.08 standard deviations. However, on the other hand, the SCG method found positive achievement gains in high school reading in PPS in comparison with similar districts in Pennsylvania. By 2014, the SCG method found an effect of 0.07 standard deviations, although not statistically significant. The DiD estimates by subgroups also point to positive effects in achievement in high school reading in 2014. We found a statistically significant gain of 0.15 standard deviations for black students and a statistically significant gain of 0.10 standard deviations for economically disadvantaged students when compared with similar students in similar schools in Pennsylvania (see Appendix B).

In addition to our impact estimation for these three achievement outcomes, we estimated the effect on the graduation and dropout rates for high schools (see Appendix B). We found a positive impact on graduation rates (i.e., an increase) in 2010 through 2014, averaging approximately 6 percentage points. The effect was significant in 2010 through 2013 and almost significant in 2014. We found a negative effect on dropout rates (i.e., a decrease) in 2010 through 2014, averaging more

[6] In contrast with the PSSA, which is an operational test of reading and math, the Keystone Exam tests specific subjects (algebra I for mathematics and literature for reading). There is much less standardization across schools and districts regarding the grade level for algebra I than for literature. Therefore, the literature Keystone results are comparable to the discontinued 11th-grade PSSA reading test, whereas the algebra I Keystone test cannot be used as a comparable replacement of the 11th-grade PSSA math test.

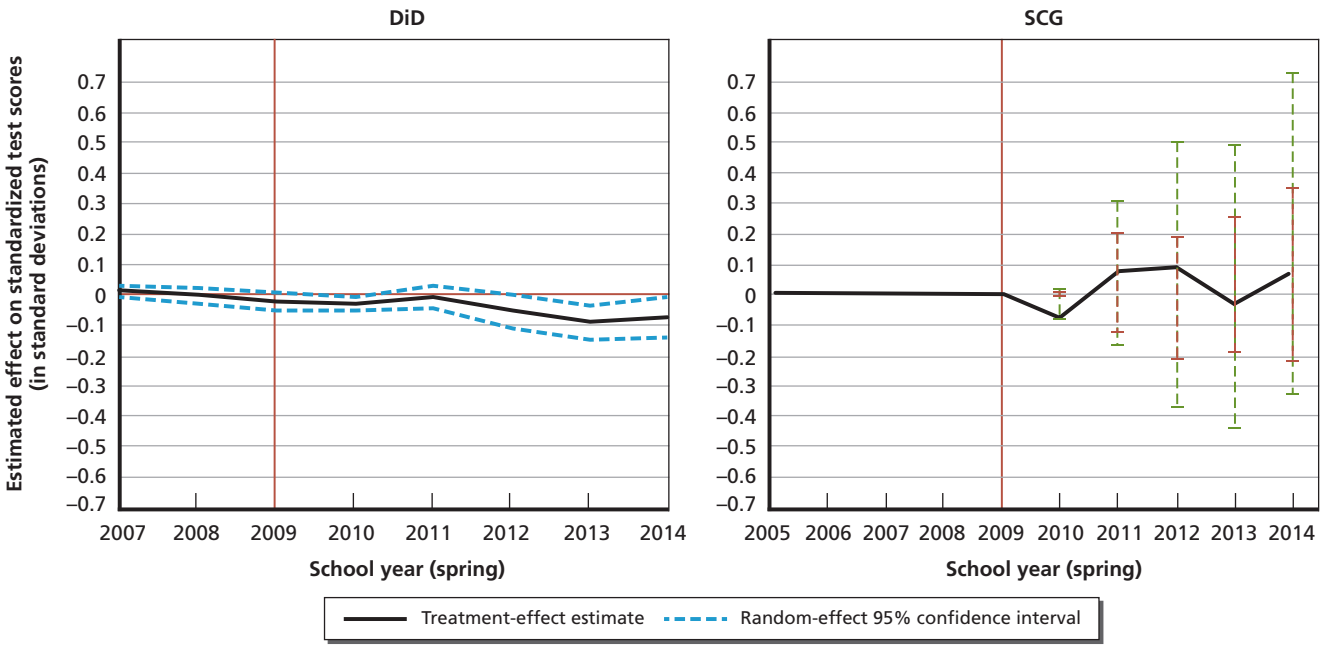
Figure 4.9
Average School-Level Test Scores on High School Reading, Pittsburgh and All Other Pennsylvania Districts



NOTE: The solid vertical red line at 2009 indicates the final year before the grant was announced and funding commenced. We weighted school-level test scores by school enrollment.

RAND *RR1295/3-1-4.9*

Figure 4.10
Estimates of Effect of Intensive Partnerships Initiative on High School Reading, Pittsburgh



NOTE: The solid vertical red line at 2009 indicates the final year before the grant was announced and funding commenced. Dashed blue lines in the DiD graph depict the 95-percent confidence intervals from the random-effect model. The vertical red and green lines in the SCG graph depict the 95-percent and 75-percent ranges, respectively, of the placebo-effect distribution.

RAND *RR1295/3-1-4.10*

than 2 percentage points. The effect was significant in 2010, 2012, 2013, and 2014 and almost significant in 2011.

Memphis City Schools, Tennessee

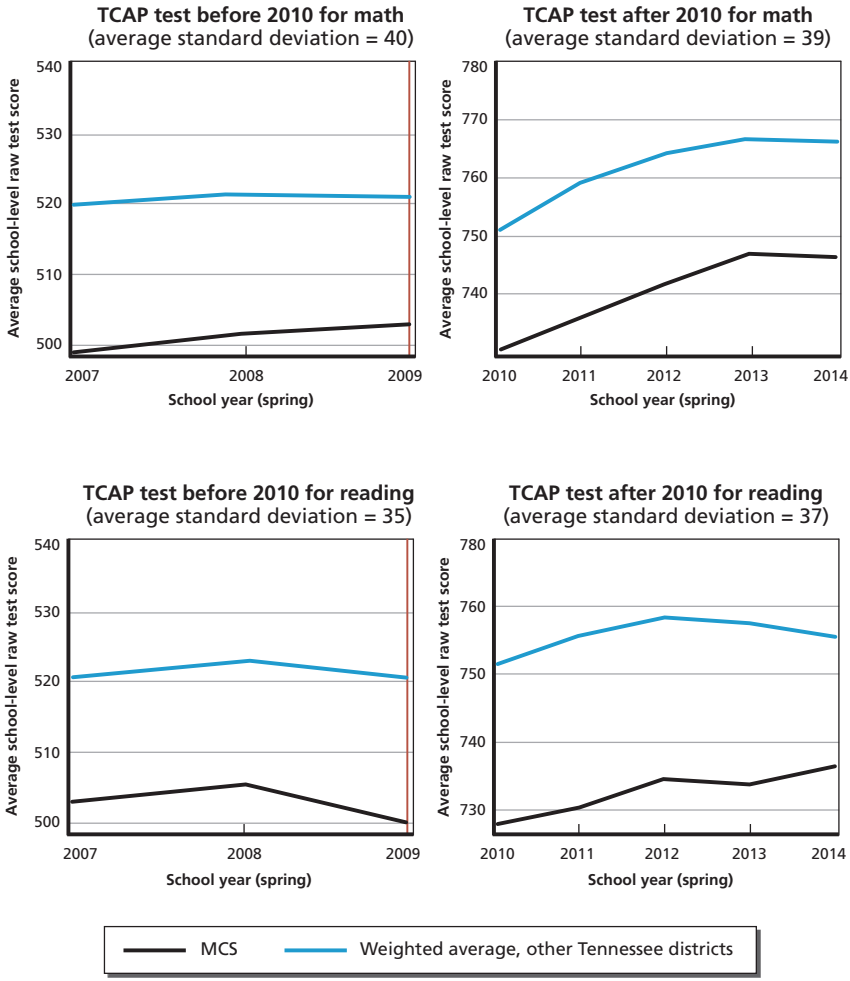
Figure 4.11 shows average grade 3 through 8 mathematics and reading test scores for MCS and all other districts in the state.[7] Because new curriculum standards and assessments were implemented in 2010, we present the averages in two panels, one for 2007 to 2009 and one for 2010 to 2014. We observe that MCS consistently scored lower than districts in the rest of the state did. However, it is not obvious from Figure 4.11 whether MCS has closed or increased the gap with the rest of the state after the Intensive Partnerships initiative.

To evaluate this point, we perform the DiD and SCG analyses, which we present in Figures 4.12 and 4.13, for grade 3 through 8 mathematics and reading, respectively. The estimated impact is less clear for MCS than for the other Intensive Partnership sites. This is because finding an adequate control group is difficult. In the case of the DiD methodology, we found that, after controlling for demographics and baseline factors, there appears to be a strong negative trend for both mathematics and reading before 2009, i.e., schools in MCS were declining more than other schools in Tennessee in the years prior to the initiative. This differential trend between MCS and other schools in the years leading up to the reforms poses a difficulty for evaluating the initiative's impact because it is not possible to know whether the different trend reflected a temporary aberration, with the MCS schools expected to return to their prior performance levels even in the absence of the reform, or whether the trend reflected a new path on which MCS would have continued without the reforms.

The interpretation of the results from the SCG method presents similar challenges. In contrast with the DiD method, the SCG method indicates a positive preinitiative trend. These deviations from 0 during

[7] As noted in Chapter One, MCS merged with SCS, but we retain the MCS notation to emphasize that we are analyzing the schools that were previously in the MCS district.

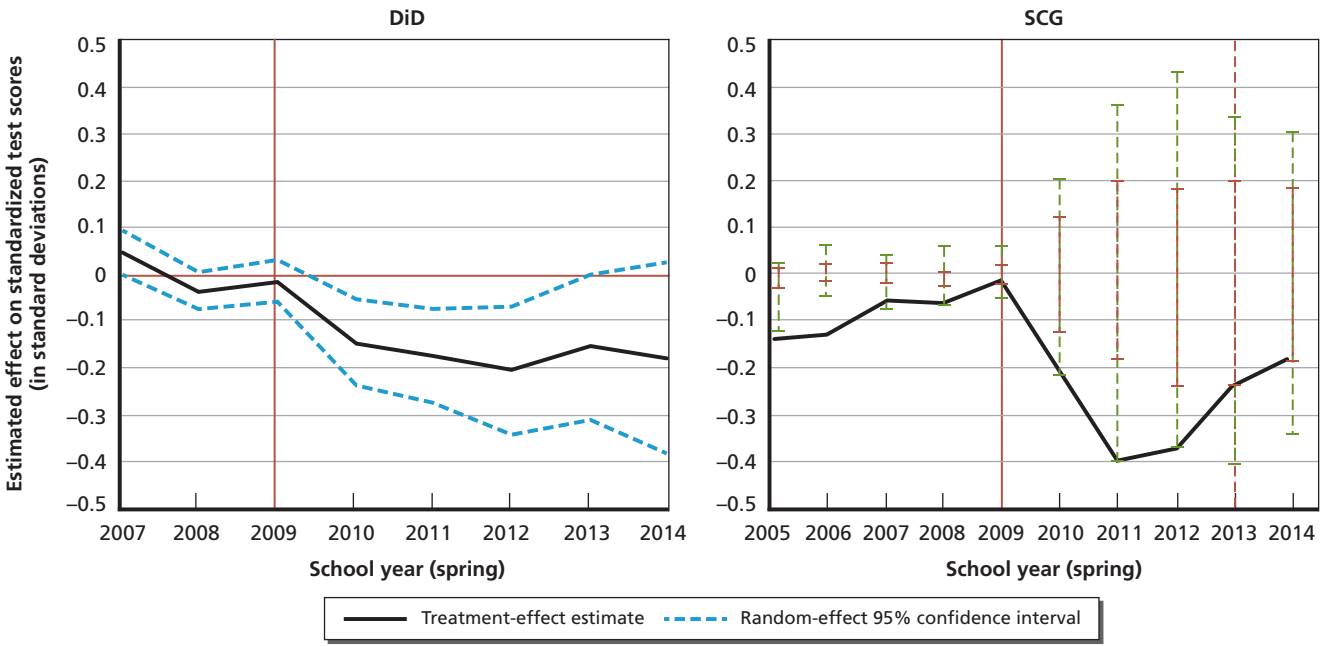
Figure 4.11
Average School-Level Test Scores on Grade 3 Through 8 Math, Memphis City Schools and All Other Tennessee Schools



NOTE: The solid vertical red line at 2009 indicates the final year before the grant was announced and funding commenced. We weighted school-level test scores by school enrollment.

RAND *RR1295/3-1-4.11*

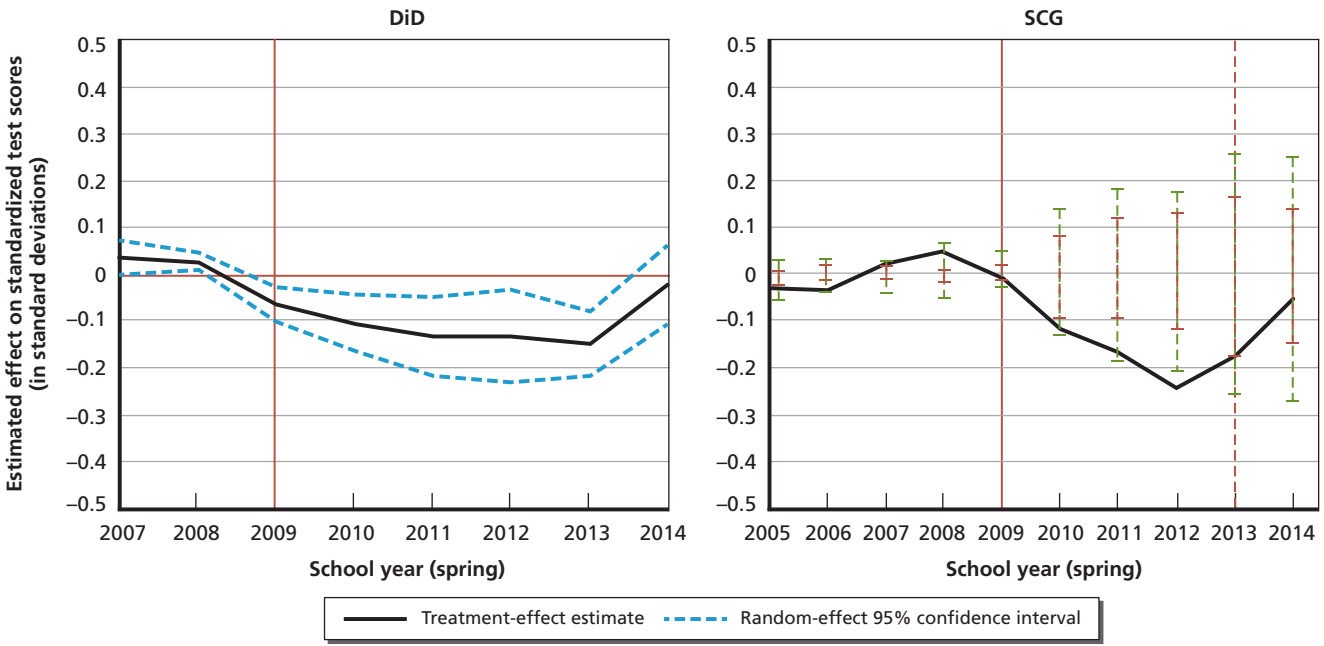
Figure 4.12
Estimates of the Intensive Partnerships Initiative's Effect on Grade 3 Through 8 Math, Memphis, Tennessee



NOTE: The solid vertical red line at 2009 indicates the final year before the grant was announced and funding commenced. Dashed blue lines in the DiD graph depict the 95-percent confidence intervals from the random-effect model. The vertical red and green lines in the SCG graph depict the 95-percent and 75-percent ranges, respectively, of the placebo-effect distribution.

RAND *RR1295/3-1-4.12*

Figure 4.13
Estimates of the Intensive Partnerships Initiative's Effect on Grade 3 Through 8 Reading, Memphis, Tennessee



NOTE: The solid vertical red line at 2009 indicates the final year before the grant was announced and funding commenced. Dashed blue lines in the DiD graph depict the 95-percent confidence intervals from the random-effect model. The vertical red and green lines in the SCG graph depict the 95-percent and 75-percent ranges, respectively, of the placebo-effect distribution.

RAND *RR1295/3-1-4.13*

the preintervention period create difficulties similar to those of the DiD method for forecasting the expected outcomes in the absence of the reform. This reflects the difficulty of creating a control group for MCS, whose composition of minority students is very different from that in the rest of Tennessee. In fact, the SCG methodology used only two districts (out of 53) in the state to construct the SCG. It is worth noting that the SCG method estimates a positive preinitiative trend relative to the control, whereas the DiD estimates a negative trend. This difference in sign indicates that the two methods, one of which uses district-level data and one of which uses school-level data, base their estimates on different samples from the state's districts and schools.

The fact that both the DiD and the SCG methods estimate strong dips in student achievement in MCS relative to other schools after 2009, despite using different control groups and finding different trends in the preinitiative period, suggests that the intervention might have had negative effects in the early years. However, both methods also agree that this situation has started to change in the later years of the intervention and that MCS stopped losing ground relative to other schools in the state in grades 3 through 8, starting in school year 2011–2012. For math, the DiD estimates show that the differences in trends started to level off in 2012, whereas the SCG estimates indicate that MCS started to catch up with the comparison schools. By 2014, both the DiD method and the SCG method estimate negative effects of −0.18 standard deviations, although neither is statistically significant.

We observe results for reading in grades 3 through 8 that are similar to those observed for math. There is a negative trend in standardized scores in MCS (in comparison with the rest of the state) prior to 2009. The DiD method shows that this negative trend continued through 2013, when MCS started to catch up with the rest of the state regarding achievement gains. The SCG method provides similar findings following the intervention. The estimated effects after the intervention are negative—although this method suggests a slightly positive trend before 2009—and show a decline until 2012. After that, we found a strong recovery, especially in school year 2013–2014. The estimated effects for 2014 are −0.02 standard deviations using the DiD method and −0.07 standard deviations using the SCG method. Again,

we should take these estimates with caution because of the difficulty in finding adequate comparison schools.

In addition to our impact estimation for these two achievement outcomes, we estimated the impact on attendance, graduation rates, and dropout rates for high schools and on attendance and promotion for other schools (i.e., schools with grades K through 8). Unfortunately, we do not have impact estimates based on high school state assessment tests because the change to EOC tests disrupted the measurement of high school achievement. We found mixed results for high schools in 2014, with significantly improved attendance rates but significant drops in graduation rates. There was no significant impact on the dropout rate. Attendance was significantly down at other schools in 2014, and promotion rates did not changed significantly. See Appendix B for more details about the impact estimates for all of the outcomes.

CHAPTER FIVE

Putting the Estimates in Context

In this chapter, we discuss whether the magnitudes of the estimated effects are substantially important. To do that, we use two comparative benchmarks recommended in the literature (Hill et al., 2007; Bloom et al., 2008). The first benchmark is based on the typical growth in academic achievement that occurs during a year of life for an average student in the United States. This growth is measured by the change in test scores from spring to spring and divided by the student-level standard deviation. Table 5.1 shows the expected achievement growth, taken from Bloom et al., 2008. It provides a context for judging the importance of the estimated effects that the Intensive Partnerships initiative has on learning by comparing them with the overall learning that would occur naturally in the absence of an intervention.

The second approach is to use as benchmarks the observed effects from past educational interventions. Although no interventions would be strictly comparable to the Intensive Partnerships initiative in term of its broad scope and scale, Table 5.2 shows effects from interventions that have received attention in the literature. We present first the effects of interventions at the school level and then effects from interventions at the district level.

School-Level Interventions

Our first source of information is Borman et al., 2003. The authors performed a meta-analysis of several CSR models. The mean of 1,111 annual effect sizes in their analysis was 0.15 standard deviations.

Table 5.1
Average Annual Gains in Effect Size from Nationally Normed Tests

Grade	Reading	Mathematics
K–1	1.52	1.14
1–2	0.97	1.03
2–3	0.60	0.89
3–4	0.36	0.52
4–5	0.40	0.56
5–6	0.32	0.41
6–7	0.23	0.3
7–8	0.26	0.32
8–9	0.24	0.22
9–10	0.19	0.25
10–11	0.19	0.14
11–12	0.06	0.01
Mean grades 3–8	0.36	0.50
Mean grades 9–11	0.21	0.20

SOURCE: Bloom et al., 2008.

NOTE: Tests include the California Achievement Test, 5th ed.; Stanford Achievement Test, 9th ed.; TerraNova California Test of Basic Skills; Gates-MacGinitie Reading Tests; Metropolitan Achievement Test (MAT 8); and TerraNova California Achievement Test and Stanford Achievement Test, 10th ed.

Panel A in Table 5.2 presents the mean annual effect for the three CSR models that had the most-rigorous empirical evidence: Direct Instruction, School Development Program, and Success for All. Rigorous empirical evidence requires a relatively large number of studies that use comparison groups. Panel A presents the mean annual effect for the subset of those studies that third parties or independent evaluators

Table 5.2
Average Effect Sizes in Educational Interventions

Intervention	Number of Effect Sizes[a]	Mean Effect
School-level interventions		
A. CSR models (Borman et al., 2003)		
Direct Instruction	146	0.15*
School Development Program	7	0.11*
Success for All	85	0.08*
B. Charter schools (Betts and Tang, 2011)		
Elementary school: math/reading	10/9	0.05*/0.02
Middle school: math/reading[b]	10/9	0.06*/0.01
High school: math/reading	8/7	–0.02/0.05
C. Project STAR (Student–Teacher Achievement Ratio) (Schanzenbach, 2007)		
K	n.a.	0.19*
Grade 1	n.a.	0.19*
Grade 2	n.a.	0.14*
Grade 3	n.a.	0.15*
D. Mean effects from meta-analysis studies (Hill et al., 2007)		
Elementary schools	32	0.23
Middle schools	27	0.27
High schools	28	0.24
District-level interventions		
E. New Leaders effects: impact of attending New Leaders school for three years (Gates et al., 2014)		
Lower grades (3–8): math/reading	10	0.03*/0.02*
High school (9–12): math/reading	5	–0.01/0.03
F. CDDRE or data-driven reform (Carlson, Borman, and Robinson, 2011)		
Lower grades (3–8): math/reading	31	0.06*/0.03

Table 5.2—Continued

Intervention	Number of Effect Sizes[a]	Mean Effect
G. DAITs (Strunk and McEachin, 2014)		
Grades 2–11: Hispanic–white gap: math/reading	43	0.05*/0.02*
Grades 2–11: black–white gap: math/reading	43	0.05*/0.02
Grades 2–11: FRPL–non-FRPL gap: math/reading	43	0.02*/0.00
Grades 2–11: ELL–non-ELL gap: math/reading	43	0.04*/0.03*

SOURCES: Borman et al., 2003; Schanzenbach, 2007; Hill et al., 2007; Betts and Tang, 2011; Gates et al., 2014.

NOTE: n.a. = not applicable. * = significantly different from 0 at the 5-percent level or less. CDDRE = Center for Data-Driven Reform in Education. DAIT = District Assistance and Intervention Team.

[a] Number of studies included in the meta-analysis. For the New Leader, CDDRE, and DAIT estimates, the number of effect sizes refers to the number of districts where the program was implemented.

[b] Excludes results from Knowledge Is Power Program middle schools.

conducted. The average annual effect for this subset ranges from 0.08 to 0.15 standard deviations.

A second source of information, presented in panel B of Table 5.2, is Betts and Tang, 2011. The authors performed a meta-analysis of charter schools' effect on student achievements. They found an overall effect size for elementary school reading and mathematics of 0.02 standard deviations and 0.05 standard deviations, respectively, and for middle school mathematics of 0.055 standard deviations. Results are not statistically significant for middle school reading or for high school mathematics or reading.

Panel C in Table 5.2 presents the effect sizes of Tennessee's Project STAR (Student–Teacher Achievement Ratio) on grades K through 3 reading and math test scores, as summarized in Schanzenbach, 2007. The average size effects range from 0.14 to 0.19 standard deviations and are statistically significant.

Panel D presents the results from Hill et al., 2007, summary of 76 meta-analyses of past educational interventions that reported achievement effect sizes for experimental and quasi-experimental stud-

ies. The mean of those effect sizes by grade range (elementary, middle, high school) is 0.25 standard deviations. However, there appears to be considerable variation in the distribution of those effects. The standard deviations of the effect sizes are 0.21 for elementary school interventions, 0.24 for middle school interventions, and 0.15 for high school interventions.

District-Level Interventions

Panel E presents the estimates from a RAND evaluation of the New Leaders program, which recruited and trained school principals in ten districts (Gates et al., 2014). We present the effects on mathematics and reading scores for students who have attended for three years a school that a New Leaders principal leads. We found some moderate effects, as high as 0.03 standard deviations for mathematics in grades 3 through 8.

Panel F presents the one-year impact of a data-driven reform initiative that CDDRE implemented, which was evaluated using district-level random assignment (Carlson, Borman, and Robinson, 2011). Note, however, that not all schools participated in the study—only those that district leadership targeted, usually a subset of the lowest-performing schools. The full intervention consisted of performing quarterly benchmark assessments (mostly in grades 3 through 8), reviewing data, training in leadership and data interpretation, providing reviews of research on effective programs and practices, and assistance in implementing proven programs. The first year of the intervention, the period for the evaluation in Carlson, Borman, and Robinson, 2011, covered only the first three components. The authors found a statistically significant effect of 0.06 standard deviations for math. They also found a positive effect for reading of 0.03 standard deviations, but it was not statistically significant at conventional levels.

Panel G presents the results of a program in California that provided technical assistance to the worst-performing districts, those at the bottom of the list of districts that had failed to make adequate yearly progress for at least four years. The state provided substantial amounts

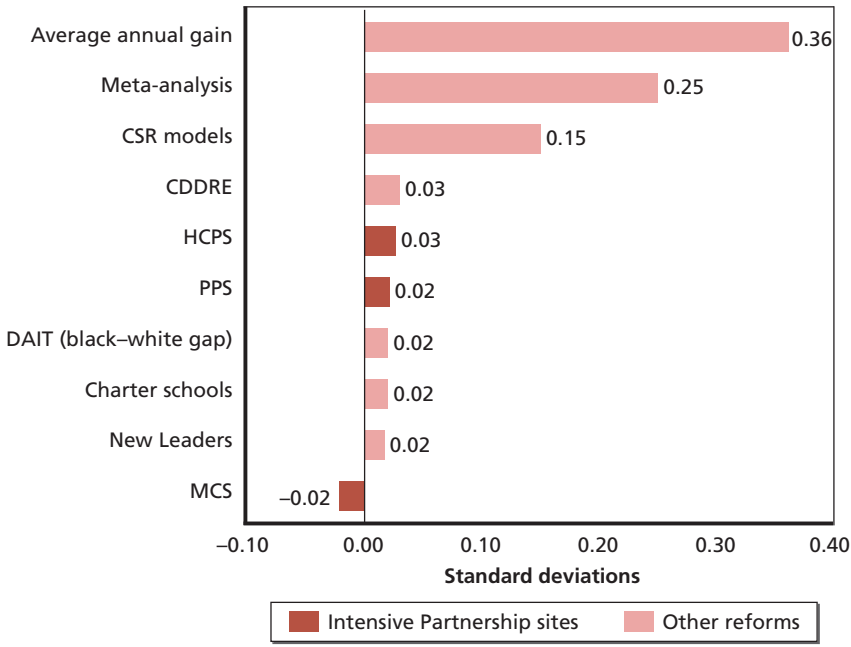
of funding to these districts to contract with state-approved experts, called DAITs, to help them build district capacity and improve student performance (Strunk and McEachin, 2014). The outcomes presented in the report are in terms of reducing achievement gaps for minority and disadvantaged groups. Providing technical assistance to these districts improved Hispanic students' achievement in math in 0.05 standard deviations, relative to white students. It led to a similar improvement for black students and smaller relative improvements for students who qualify for the federal FRPL program and for ELL students. Regarding reading scores, the intervention led to smaller improvements in general, and scores were statistically significant only for reducing the Hispanic–white gap and the ELL–non-ELL gap.

Next, we examine how the effect sizes from the Intensive Partnerships initiative compare with these benchmarks. For this comparison, we average the 2013–2014 DiD estimates for each subject (reading and math) in grades 3 through 8 and in high school. It is important to note that estimates for the 2013–2014 school year are the most-optimistic effects to date in most cases. They also are a better test of the initiative's effectiveness in improving student performance because the reforms needed several years to be implemented in each site. Moreover, the upward trend observed in most sites suggests that there could be bigger impacts in the coming years.

Figure 5.1 presents the results for reading scores in grades 3 through 8. The average effect of PPS and HCPS is 0.02 standard deviations. This effect size represents about 6.6 percent of the expected yearly gain in reading in the absence of any intervention (0.36 standard deviations). The effect size appears small when compared with other benchmarks from school-level interventions but compares favorably with the average effect found for charter schools and for other districtwide interventions, such as the New Leaders program and the provision of technical assistance (DAITs).

The same favorable comparison is found in the case of math for PPS, as shown in Figure 5.2. The average effect of 0.10 standard deviations is about 20.8 percent of the expected learning in a year without any intervention. The effect is smaller than the benchmarks from the other school-level interventions but larger than the average effects from

Figure 5.1
Effect Sizes for Reading Scores, Grades 3 Through 8

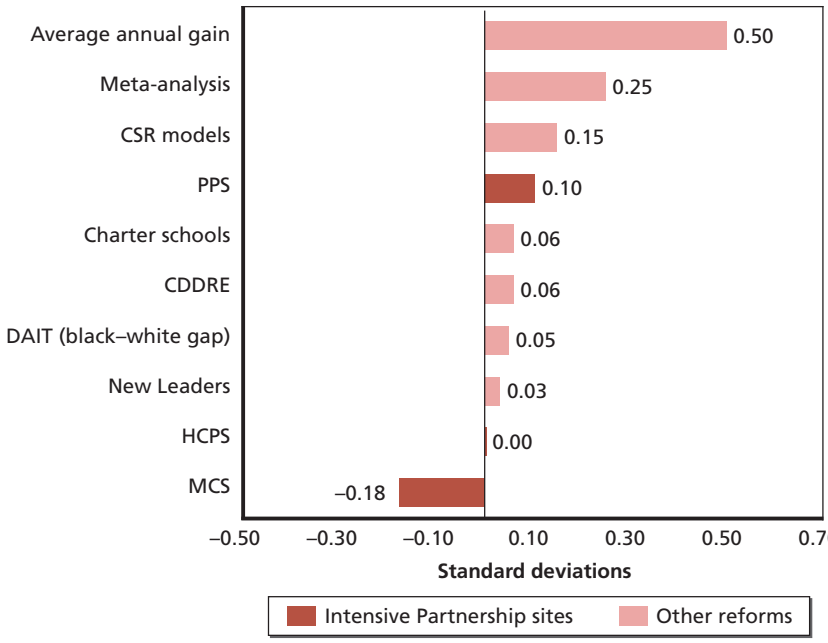


RAND *RR1295/3-1-5.1*

charter schools or the other district-level interventions. The average effect in HCPS is smaller and close to zero but, as described in Chapter Four, we observe an increase in the effects in the later years.

In Figures 5.1 and 5.2, we also found negative effects of the Intensive Partnership initiative in reading and mathematics for MCS for 2013–2014. However, as discussed earlier, we should interpret the results from MCS with caution because we found that there was a less favorable trend in MCS test scores in comparison with other schools in the state even before the start of the Intensive Partnership initiative. Moreover, it was difficult to establish a comparison group for MCS because its composition of minority students is unlike that of any other district in Tennessee. If anything, it appears that the Intensive Partnership initiative has helped to stop the negative trend in mathematics and reading scores in MCS, relative to the trend in the rest of the state.

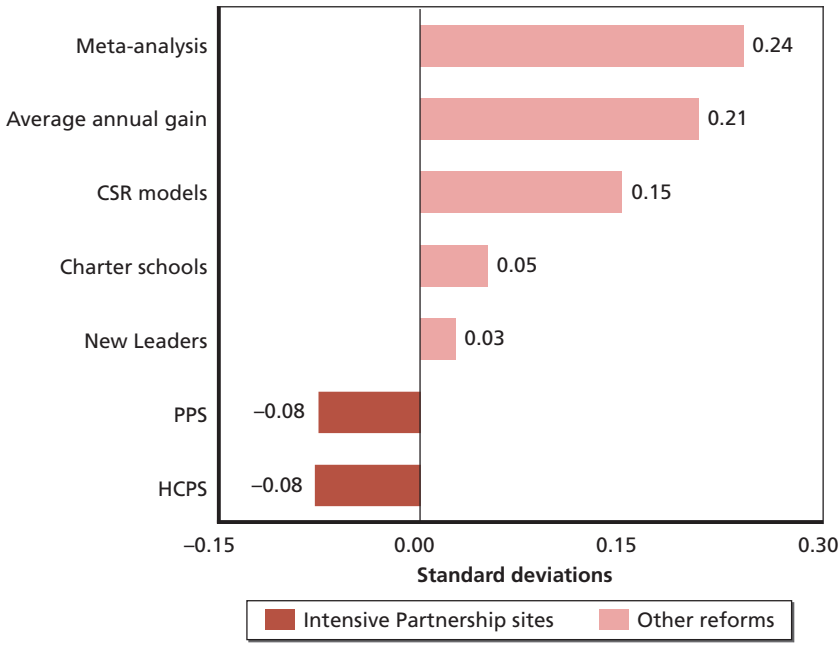
Figure 5.2
Effect Sizes for Math Scores, Grades 3 Through 8



RAND *RR1295/3-1-5.2*

Figure 5.3 presents the effects for reading scores in high school from PPS and HCPS, which are the sites from which we have available estimates. The evidence indicates negative effects of the Intensive Partnership intervention on high school student achievement in both districts. We found that the intervention is associated in 2013–2014 with a reduction in reading test scores of 0.08 standard deviations for high schools in PPS and in HCPS compared with other high schools in Pennsylvania and Florida, respectively.

Figure 5.3
Effect Sizes for Reading Scores, High School



RAND *RR1295/3-1-5.3*

CHAPTER SIX

Summary and Conclusion

This report presents impact estimates of the Bill & Melinda Gates Foundation's Intensive Partnerships for Effective Teaching on student outcomes through the 2013–2014 school year. The sites evaluated are HCPS, MCS, and PPS. The initiative includes a broad series of reforms related to teacher evaluation and how those evaluations are used to inform human-resource practices. The ultimate goal of the reforms is to provide every student access to highly effective teachers.

These reforms centered on improving the teacher-evaluation systems in these sites and using teacher evaluations to make personnel decisions, including providing supports to help teachers improve their effectiveness and incentives to retain effective teachers and place them with the students who are most in need. Although the reforms took time to implement, all sites had implemented many aspects by the last two years of the period that this report covers. Our analysis compares the progress in student achievement that the three school districts made and the performance that would be expected in the absence of the initiative. We estimate their expected performance using the actual performance of schools and districts within the same states. One factor that must be taken into account is that these states were, to varying degrees, making policy changes to encourage similar reforms in all districts. Therefore, our impact estimates reflect gains over and above those made by other schools and districts that were responding to statewide policy.

We found mixed but mostly insignificant effects of the initiative on student performance in the lower grades (3 through 8), with the

exception of MCS, which fared significantly worse after the start of the initiative. However, impact estimates were increasing in 2013–2014 for almost all achievement outcomes in the sites. This should not be surprising because the sites needed several years to implement the broad set of reforms that the initiative promoted.

If the more-recent trends continue, the sites could observe significant positive impact in the next years. For example, focusing on the 2013–2014 school year (latest available data) and using the DiD methodology, we found that, on average, the schools of PPS experienced greater achievement gains in lower-grade (3–8) mathematics (0.10 standard deviations) and reading (0.02 standard deviations) than comparable schools did in other Pennsylvania districts.[1] Similarly, schools of HCPS experienced greater achievement gains in reading (0.03 standard deviations) in the lower grades (3–8) and performed similarly in mathematics when compared with comparable schools in other Florida districts. However, the estimates fell short of the conventional levels of statistical significance in most cases. We found that schools in MCS experienced lower achievement gains in school year 2013–2014 in mathematics in the lower grades (–0.18 standard deviations) than comparable schools in other Tennessee districts did, although this represents a rebound in the recent years after a very large dip in the first three years following the start of the initiative. We observe a similar dip and rebound in reading in MCS but still find an average negative effect (–0.02 standard deviations) in school year 2013–2014 in comparison with similar schools in Tennessee. However, neither of these estimates is statistically significant.

We found evidence of negative effects of the Intensive Partnership intervention on high school student achievement in both districts where we could estimate this effect. We found that the intervention is associated with a statistically significant reduction in reading test scores of 0.08 standard deviations for high schools in PPS and in

[1] We express impact as fractions of a student-level statewide standard deviation of the relevant test score. In Chapter Five, we provide more detail on this measure. For example, an average year of school for students in grades 3 through 8 is equivalent to approximately one-third of a standard deviation for reading and one-half of a standard deviation for mathematics.

HCPS compared with other high schools in Pennsylvania and Florida, respectively.[2]

We also estimated the effects on academic achievement for black students, Hispanic students, and economically disadvantaged students, where these data were available. In most cases, the subgroup effects followed the same pattern as the overall effects. The only exception is for high school reading in PPS, where the overall effect in 2014 is negative and statistically significant, but the effects for black students and economically disadvantaged students are positive and statistically significant (0.15 and 0.10 standard deviations, respectively).

With the exception of the estimates for high school reading in PPS, the intervention's estimated impacts are similar using two distinct methodologies: a DiD analysis and an SCG analysis. This lends credence to the robustness of our results and strengthens our confidence in the overall findings, including the recent upward trajectory in the impact estimates across Intensive Partnership sites. This trajectory suggests that the reforms might be on the way to having a positive impact after a few transition years with little or small negative impact. In our next report, we will update these estimates for 2014–2015 and 2015–2016 to see whether the trend continues.

---

[2] We do not estimate results for high school test scores for MCS because Tennessee administers only end-of-course exams for high school students. The sample of students taking these end-of-course exams is determined by many factors that we have not measured; therefore, the test results are not useful for school-level comparisons.

APPENDIX A

Estimation Methods

School-Level Difference-in-Differences Methodology

This appendix provides details concerning the DiD method used in this report.

Model

In the first step of the method, we develop a forecasting model that uses preintervention data to predict the outcomes in postintervention years under the counterfactual assumption that the intervention had not happened. The prediction model accounts for separate intercepts for each district and for differences in school and district demographics. The equation we estimate is given by[1]

$$Y_{sdt} = \alpha_d + \beta_X X_{sdt} + \beta_{\bar{X}} \bar{X}_{dt} + \in_{sdt}, \quad \text{(A.1)}$$

where Y_{sdt} is the outcome for school s in district d by year t (note that the outcomes can pertain to a specific grade or student subgroup); the term α_d denotes district-specific intercepts; the term X_{sdt} denotes the school demographic characteristics each year, including ethnicity composition and percentage of students in FRPL plans. It also includes some time-invariant characteristics, such as average preintervention proficiency levels in mathematics and reading (i.e., in school years 2006–2007 to

[1] We do not weight these models by school size, so each school is weighted equally in the analysis.

2008–2009).[2] For a list of specific covariates, see Table 3.1 in Chapter Three. The term $\bar{X}_{dt}$ contains the time-varying variables in X_{sdt} but aggregated at the district level (we do not include time-constant district-level variables because they are perfectly collinear with the district-specific intercept).

An extension of our model would be to use linear district-specific time trends to predict postintervention counterfactual outcomes. However, we found that, as we predict several years into the future, maintaining the trends from before the intervention leads to large prediction errors and imprecise estimates of the initiative's impact. Thus, in this report, we do not include linear district-specific trends in our model.[3]

We estimate the model in Equation A.1 using only information from school years 2006–2007 to 2008–2009. We use the estimated model to form a forecast of the outcome for each school in the postintervention period. We then compute the difference between the forecast and the actual value for all schools. This difference reflects how the outcome differs from what is expected given the preintervention pattern in outcomes.

The second step in the analysis examines whether the differences between the forecasted and actual values are systematically different in the Intensive Partnership sites and the comparison districts. We estimate the following regression:

$$dif_{sdt} = \gamma + \eta_t treatment_d + \theta_{t,X} X_{sdt} + \theta_{t\bar{X}} \bar{X}_{dt} + \mu_{dt} + \upsilon_{sdt}. \quad \text{(A.2)}$$

The variable dif_{sdt} denotes the difference between the forecasted and actual values of the outcome. The vectors X_{sdt} and $\bar{X}_{dt}$ are the same vectors of school-level and district-level demographics as in Equa-

---

[2] Because we use average preintervention proficiency levels in the forecasting model, schools that opened after 2009 are not included in the analysis sample.

[3] We examined the error of the predicted outcomes for schools in the comparison group. Because it was not exposed to the intervention, it is expected that past trends at the district level are a good predictor of future outcomes. We found that the model without trends delivered smaller prediction errors, as measured by the root mean-squared error.

tion A.1 (excluding time-invariant variables). The term μ_{dt} is a district random component distributed normally, i.e., $\mu_{dt} \sim N\left(0,\sigma^2\right)$. The variable $treatment_d$ is an indicator variable that equals 1 if the schools are in the intervention district. The coefficient of interest is η_t, which captures the difference in the prediction error (the difference in the difference) between schools in the intervention and comparison districts in year *t*.

Notice that we allow η_t to vary with time. In practice, we estimate Equation A.2 separately for every year, before and after the intervention. Also note that we control for demographic factors both in Equation A.1, the forecasting model, and in Equation A.2, the model that explains the difference between the forecasted and the actual values. The reasoning for following this approach is that, in Equation A.1, we assume the effects of demographic factors to be constant over time. In other words, we assume that the influence that different factors (such as the ethnicity composition) have on the achievement outcome do not vary over time. In reality, however, this might not be true, and it adds to the prediction error. We acknowledge this by adding demographic factors to Equation A.2 and by letting them have differential impacts in every year (because we estimate separate regressions for Equation A.2 for each time period). The key assumption behind this approach is that changes in demographics at the school level and, more importantly, at the district level, are independent of or unrelated to the Intensive Partnerships initiative.

Statistical Inference

Conducting statistical inference—that is, performing statistical tests of hypotheses, such as "the Intensive Partnership reforms had no effect on a particular student achievement metric"—is not straightforward with the school-level DiD approach described above because the outcomes of schools within a district are likely to be correlated with each other because of the likely presence of "common shocks" that affect the outcomes of all schools within a district. Treating them as statistically independent units would lead one to overstate the statistical precision of a given estimate. Moulton, 1986, and Bertrand, Duflo, and Mullainathan, 2004, discuss this problem. These articles show that

ignoring this problem will lead to drastically understated estimated standard errors of the effect of an intervention that occurs at an aggregate level. Furthermore, recent research has shown that a conventional approach to inference, basing confidence intervals using standard errors computed using an adjustment that corrects for clustering at the level of the treatment assignment (in our case, districts), performs very poorly when the number of clusters receiving the treatment is small (Conley and Taber, 2011). In our case, only one district in any state is treated, so this approach to statistical inference is not appropriate. Because of this intradistrict correlation, we added the random-effect component μ_{dt} to Equation A.2. The idea is that this term would capture all sources of within-district residual correlation across schools. However, the drawback to this approach is that it imposes a strong assumption about the structure of the residual correlation, and there is no way to assess whether such an assumption is warranted. Because of that, in this report, we have also implemented an SCG methodology. This methodology aggregates school data to the district level, avoiding the problem of having to model common shocks among schools within a district. It also deals with the issue of a small number of treated units (one district) by using nonparametric—or free-of-distributional-assumptions—inference. We explain more about this method, and its limitations, in the next section.

Synthetic-Control-Group Methodology

The SCG methodology uses information at the level of the intervention. In our case, we use information aggregated at the school district level. The central idea behind the SCG approach is to construct an SCG made up of weighted observations from other comparison districts. The weights are created so that the weighted average of the SCG looks as similar as possible to the treatment group in its preintervention characteristics and outcomes.

Model

The SCG method is based on finding weights that minimize the differences between the treatment site and the weighted comparison group. More specifically, the method minimizes the distance between a ($k \times 1$) vector of preintervention covariates for the treated district, X_1, and a weighted combination of the vector for comparison sites, WX_0. Specifically, the vector of weights W is chosen to minimize the distance defined by the expression

$$\sqrt{\left(X_1 - WX_0\right)' V \left(X_1 - WX_0\right)},$$

where V is a ($k \times k$) positive semidefinite matrix. The role of the matrix V is to assign the relative importance of the matching covariates based on their predictive power of the outcome. A simple way to choose V is the diagonal matrix in which the elements on the diagonal are the standardized coefficients of a regression of the outcome on the covariates. However, a better choice is to weight each matching covariate so as to minimize the mean-squared prediction error of the outcome over the preintervention periods. We found that using a fully nested optimization procedure that searched among all possible positive semidefinite matrices for a minimum diagonal V yielded a higher-quality match. Thus, all estimations in our analysis use this enhanced approach to calculate the V matrix. Once the weights have been identified, they are applied to the outcome of interest, and one can visually assess whether outcomes appear close in the preintervention period and diverge posttreatment.

One disadvantage in comparison with the DiD approach is that the SCG methodology does not allow controlling for changes in student characteristics after the intervention. For instance, there might be important changes in the percentage of minority students or in poverty levels in the treated district or in the districts that are part of the SCG. If these changes in student characteristics are not a result of the intervention, we need to account for their impact in the outcomes of interest in order to obtain unbiased impact estimates of the intervention.

Informed by prior RAND experience with SCG methods, as well as suggestions by Abadie, Diamond, and Hainmueller, 2010, to minimize bias that might occur by giving weights to districts completely different from those of the treatment sites, we restrict the donor pool (i.e., the potential controls) to those districts that were most similar to the treatment site. Across all the Intensive Partnership states, the treatment districts tend to be quite different from other districts in both size and demographics. Thus, prior to performing any analysis, we eliminated districts that were very different from the Intensive Partnership district. We dropped districts where average enrollment (2005–2014) was less than 1 percent of the average enrollment in the Intensive Partnership district. We also dropped districts where average percentage of minority students (2005–2014) was less than 10 percent of the average percentage in the Intensive Partnership district.

After trimming the data in this manner, we decide on the outcomes of interest and the preintervention period variables to include. In Abadie, Diamond, and Hainmueller, 2010, the authors indicate that the weighted control for the SCG is calculated so that it approximates the unit exposed for the outcome predictors and M linear combinations of the preintervention outcome. This should result in the smallest mean-squared predicted difference between the treatment and weighted control groups. However, little guidance is given about how to select between various linear combinations of the preintervention outcome variable. For example, in their paper, the authors have 18 preintervention years but chose three specific years from these to match. We found that, by using the outcome variables for the five preintervention years and the same covariates that we use in the DiD analysis, the mean-squared prediction error for PPS and HCPS were close to 0. We did not find a good match for MCS, but this was because of the characteristics of MCS and not the model. Therefore, we chose to model the SCG using the outcomes of the five preintervention years and the average of the five preintervention years of the covariates listed in the last row of Table 3.1 in Chapter Three. We use a user-written command called *synth* to run the SCG models in Stata software (Abadie, Diamond, and Hainmueller, 2011). In addition to inputting the covariates listed above, we add the option to use the nested optimization pro-

cedure described above and the option to calculate the mean-squared prediction error for 2005 through 2009 (our preintervention years).

The number of control districts used to construct the weights varied for each state, depending on the total number of districts in the state and the trimming procedure described above. In Table A.1, we show, for each outcome in each state, the number of districts in the donor pool used to construct the SCG and the number of districts that received a nonzero weight. We observe that, for HCPS and PPS, most of the districts that we retain after the trimming take part in constructing the SCG. In contrast, in the MCS case, only two districts out of the 53 available districts had nonzero weights. This reflects the difficulty in finding districts in Tennessee that are similar to MCS and can act as controls.

Statistical Inference

We adapt the inference method that Abadie, Diamond, and Hainmueller, 2010, recommends for the SCG estimation procedure and Bertrand, Duflo, and Mullainathan, 2004, recommends for DiD

Table A.1
Synthetic-Control-Group Methodology: Number of Districts in the Construction of the Control Group and of the Placebo Distribution

Site	Outcome	Districts in Donor Pool	Districts with Positive Weight in SCG	Districts in Placebo Distribution
HCPS	Math grades 3–8	54	54	28
	Reading grades 3–8	54	54	29
	Reading high school	47	47	15
MCS	Math grades 3–8	53	2	53
	Reading grades 3–8	53	2	53
PPS	Math grades 3–8	117	116	48
	Reading grades 3–8	117	115	59
	Math high school	72	68	30
	Reading high school	72	70	35

models. Specifically, we use exact inference with permutation tests based on placebo treatments (Abadie, Diamond, and Hainmueller, 2010; Bertrand, Duflo, and Mullainathan, 2004). The idea behind this method is to compare the intervention's actual estimated impact and a distribution of placebo effects that we obtain by repeatedly redoing the same analysis but each time using a different comparison district as a placebo treatment site. The distribution of placebo effects mimics the variability in the estimates that would occur naturally because of unobserved factors. If the actual estimate is larger than this natural variability, we deem the estimate to be statistically significant. The main appeal of the permutation tests is that they make no assumptions about the distribution of the placebos but rather use the empirical distribution that the data provide. However, permutation testing can be quite conservative, requiring that the impact sizes be relatively large to be considered statistically significant.

After we calculate the placebo treatment effects for each district in the comparison group, we use these estimations to create the placebo distribution. Abadie, Diamond, and Hainmueller, 2010, recommends restricting the distribution to placebo treatment effects that were calculated with a reasonable degree of error. The authors argued that, when a plausible comparison group cannot be formed, the SCG method is not the correct one to use, so districts that do not have an adequate root mean-squared prediction error (RMSPE) should not be used in creating the placebo distributions. The authors do not make a case for any specific range of adequate error, but they show two sets of results: (1) restricting the placebo treatment effects to those whose RMSPE is less than 20 times the RMSPE of the actual treatment district and (2) restricting the placebo treatment effects to those whose RMSPE is less than five times the RMSPE of the actual treatment district. We create our placebo ranges by restricting the distribution to include only placebo treatment districts with RMSPEs less than five times that of the Intensive Partnership site. The rightmost column of Table A.1 shows the number of districts used to generate the distribution of placebo effects.

APPENDIX B

Results for Additional Outcomes

This appendix contains the results for additional outcomes, including test scores for student subgroups (specifically, by demographic characteristic and grade) and indicators of high school persistence (graduation and dropout rates). Table B.1 presents results for HCPS, Table B.2 for PPS, and Table B.3 for MCS.

For each outcome, we report the estimated treatment effect for a given postintervention year in the first row and its *p*-value in the second row in brackets. In this appendix, we report only DiD estimates. The column titled "2009 (Placebo Effect)" presents the effects for school year 2008–2009, the last school year *before* the implementation of the initiative. Because the initiative had not yet started, we would expect these placebo estimates to be small and not statistically significant. If, on the contrary, these effects are statistically significant, that would indicate the existence of differential trends in outcomes between the Intensive Partnership site and those in the rest of the state, prior to the start of the initiative. In those cases, the impact estimates for the other years (2010–2014) should be interpreted with caution because it is difficult to extrapolate with certainty what those trends would have been in the absence of the initiative.

Table B.1
Hillsborough County Public Schools Impact Estimates, by Grade, Subgroup, and Year

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
3	Math	All	Difference	–0.036**	–0.014	0.016	–0.038	–0.029	0.024
			p-value	0.010	0.449	0.536	0.293	0.446	0.618
		Black	Difference	–0.053**	–0.035	–0.001	–0.011	–0.073	–0.014
			p-value	0.041	0.153	0.979	0.758	0.106	0.808
		High poverty	Difference	–0.027	0.001	0.009	–0.039	–0.023	0.031
			p-value	0.228	0.972	0.764	0.423	0.594	0.571
		Hispanic	Difference	–0.034	–0.040	–0.007	–0.034	–0.097*	–0.013
			p-value	0.195	0.109	0.851	0.504	0.073	0.836
	Reading	All	Difference	–0.056***	0.020**	0.057***	0.029**	–0.013	0.076***
			p-value	0.001	0.030	0.001	0.011	0.562	0.001
		Black	Difference	–0.051**	0.022	–0.005	0.044***	–0.015	0.065**
			p-value	0.022	0.105	0.730	0.007	0.453	0.011
		High poverty	Difference	–0.047***	0.014	0.049***	0.030**	–0.006	0.076***
			p-value	0.004	0.090	0.001	0.009	0.735	0.001

Table B.1—Continued

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
3, continued		Hispanic	Difference	–0.038	–0.030	0.102***	0.038	–0.009	0.098***
			p-value	0.249	0.140	0.001	0.201	0.694	0.004
4	Math	All	Difference	0.014	0.027	–0.007	–0.074***	–0.058**	–0.061*
			p-value	0.240	0.106	0.741	0.006	0.043	0.064
		Black	Difference	0.021	0.033*	0.026	–0.023	–0.036	–0.013
			p-value	0.220	0.070	0.382	0.481	0.232	0.744
		High poverty	Difference	–0.003	0.074***	0.019	–0.078***	–0.042	–0.042
			p-value	0.895	0.001	0.368	0.008	0.181	0.247
		Hispanic	Difference	0.011	–0.006	–0.031*	–0.097***	–0.065**	–0.138***
			p-value	0.471	0.774	0.091	0.001	0.035	0.001
	Reading	All	Difference	–0.008	–0.017***	0.058***	0.035***	0.017	0.033***
			p-value	0.413	0.001	0.001	0.001	0.281	0.006
		Black	Difference	–0.021	0.002	0.067***	0.048	0.011	0.103***
			p-value	0.101	0.934	0.001	0.099	0.569	0.001
		High poverty	Difference	–0.025	0.011	0.065***	0.025*	0.023	0.061***
			p-value	0.213	0.088	0.001	0.068	0.170	0.001

Table B.1—Continued

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
4, continued		Hispanic	Difference	–0.035	–0.063***	0.027*	0.045***	0.008	–0.007
			p-value	0.152	0.005	0.056	0.001	0.701	0.651
5	Math	All	Difference	–0.032	–0.003	–0.012	–0.046**	–0.005	–0.035
			p-value	0.116	0.722	0.630	0.019	0.856	0.311
		Black	Difference	0.007	–0.008	0.068*	–0.020	0.065*	0.015
			p-value	0.850	0.450	0.085	0.474	0.052	0.721
		High poverty	Difference	–0.038**	–0.029***	0.007	–0.043**	–0.010	–0.023
			p-value	0.032	0.001	0.750	0.024	0.722	0.514
		Hispanic	Difference	–0.071***	0.003	–0.009	–0.056**	–0.033	–0.058
			p-value	0.001	0.845	0.755	0.015	0.224	0.139
	Reading	All	Difference	–0.032***	0.010	0.008	–0.012	0.002	0.060***
			p-value	0.001	0.191	0.468	0.310	0.890	0.002
		Black	Difference	–0.016	–0.019	0.029	0.009	0.007	0.065**
			p-value	0.557	0.104	0.268	0.504	0.809	0.021
		High poverty	Difference	–0.042***	–0.003	0.020	–0.004	–0.019	0.092***
			p-value	0.002	0.698	0.101	0.750	0.174	0.001

Table B.1—Continued

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
5, continued		Hispanic	Difference	–0.074***	–0.019	0.001	–0.026	–0.031	0.042
			p-value	0.001	0.248	0.983	0.240	0.495	0.186
6	Math	All	Difference	–0.067***	–0.028***	0.020	–0.104***	–0.073***	–0.073***
			p-value	0.001	0.001	0.288	0.001	0.001	0.002
		Black	Difference	–0.096***	–0.012	0.010	–0.099***	–0.079**	–0.074
			p-value	0.006	0.309	0.588	0.004	0.035	0.125
		High poverty	Difference	–0.048**	–0.016*	0.001	–0.066***	–0.099***	–0.032
			p-value	0.015	0.074	0.960	0.001	0.001	0.338
		Hispanic	Difference	–0.061***	0.024	0.077***	–0.019	–0.055	–0.022
			p-value	0.002	0.268	0.001	0.608	0.120	0.558
	Reading	All	Difference	–0.027*	–0.023***	–0.038***	–0.105***	–0.085***	–0.115***
			p-value	0.064	0.001	0.006	0.001	0.001	0.001
		Black	Difference	–0.035	0.000	–0.050***	–0.090***	–0.100***	–0.121***
			p-value	0.360	0.980	0.001	0.001	0.001	0.001
		High poverty	Difference	–0.021	–0.001	–0.034*	–0.088***	–0.080***	–0.081***
			p-value	0.259	0.926	0.056	0.001	0.001	0.001

Table B.1—Continued

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
6, continued		Hispanic	Difference	–0.050*	0.021	0.044**	–0.015	–0.023	–0.053***
			p-value	0.059	0.229	0.025	0.498	0.402	0.004
7	Math	All	Difference	–0.031***	–0.041***	–0.051***	–0.042**	–0.030	–0.002
			p-value	0.002	0.001	0.001	0.045	0.325	0.960
		Black	Difference	0.028	–0.034	–0.033	–0.103***	–0.167***	–0.099**
			p-value	0.340	0.155	0.324	0.001	0.001	0.010
		High poverty	Difference	–0.042***	–0.016	–0.023	–0.034*	–0.022	–0.002
			p-value	0.004	0.289	0.131	0.078	0.470	0.968
		Hispanic	Difference	–0.060**	–0.058***	–0.005	–0.048**	–0.061	0.000
			p-value	0.019	0.003	0.876	0.034	0.169	0.994
	Reading	All	Difference	–0.040***	–0.019**	–0.060***	–0.093***	–0.124***	–0.083***
			p-value	0.008	0.039	0.001	0.001	0.001	0.001
		Black	Difference	0.004	0.003	–0.031	–0.143***	–0.164***	–0.140***
			p-value	0.880	0.908	0.330	0.001	0.001	0.001
		High poverty	Difference	–0.032**	0.000	–0.035***	–0.082***	–0.087***	–0.076***
			p-value	0.044	0.976	0.001	0.001	0.003	0.009

Table B.1—Continued

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
7, continued		Hispanic	Difference	−0.050*	−0.004	−0.041***	−0.104***	−0.142***	−0.110***
			p-value	0.077	0.830	0.002	0.001	0.001	0.002
8	Math	All	Difference	−0.085***	0.007	−0.009	−0.034**	0.115*	0.163***
			p-value	0.001	0.468	0.259	0.047	0.065	0.002
		Black	Difference	−0.018	0.003	−0.061***	−0.042*	−0.001	0.055
			p-value	0.288	0.840	0.009	0.086	0.982	0.135
		High poverty	Difference	−0.084***	−0.030***	−0.026**	−0.069***	0.043	0.115**
			p-value	0.001	0.004	0.019	0.001	0.392	0.012
		Hispanic	Difference	−0.108***	−0.016	−0.008	−0.014	0.100	0.183**
			p-value	0.001	0.206	0.830	0.691	0.139	0.018
	Reading	All	Difference	−0.022*	0.018**	−0.072***	−0.099***	−0.083***	−0.086***
			p-value	0.045	0.032	0.001	0.001	0.001	0.001
		Black	Difference	−0.019	−0.005	−0.064	−0.088***	−0.087***	−0.018
			p-value	0.365	0.756	0.130	0.001	0.005	0.522
		High poverty	Difference	−0.030**	−0.009	−0.096***	−0.119***	−0.119***	−0.094***
			p-value	0.015	0.459	0.001	0.001	0.001	0.001

Table B.1—Continued

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
8, continued		Hispanic	Difference	–0.078***	0.000	–0.045	–0.095***	–0.107**	–0.068
			p-value	0.002	0.985	0.332	0.002	0.030	0.110
3–8	Math	All	Difference	–0.032***	–0.002	–0.004	–0.052**	–0.017	0.002
			p-value	0.001	0.840	0.841	0.030	0.580	0.962
		Black	Difference	0.009	0.002	–0.010	–0.064***	–0.039*	0.007
			p-value	0.517	0.893	0.520	0.006	0.093	0.833
		High poverty	Difference	–0.029***	0.007	–0.011	–0.061**	–0.027	0.004
			p-value	0.001	0.400	0.534	0.016	0.366	0.910
		Hispanic	Difference	–0.039***	–0.016***	–0.020*	–0.059**	–0.053*	–0.031
			p-value	0.001	0.007	0.067	0.026	0.068	0.408
	Reading	All	Difference	–0.028***	0.003	0.013	–0.012	–0.025	0.026
			p-value	0.001	0.477	0.161	0.255	0.106	0.102
		Black	Difference	–0.012	0.021**	–0.012	–0.027*	–0.031**	0.024
			p-value	0.309	0.010	0.338	0.088	0.035	0.199
		High poverty	Difference	–0.037***	0.014***	0.007	–0.013	–0.018	0.044*
			p-value	0.001	0.001	0.510	0.328	0.285	0.050

Table B.1—Continued

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
3–8, continued		Hispanic	Difference	–0.030***	–0.010	0.019**	–0.010	–0.026	0.021
			p-value	0.005	0.172	0.031	0.633	0.232	0.427
9	Reading	All	Difference	–0.079***	–0.027**	–0.004	–0.106***	–0.127***	–0.132***
			p-value	0.001	0.031	0.700	0.001	0.001	0.001
		Black	Difference	–0.123***	0.001	–0.054***	–0.093**	–0.087**	–0.211***
			p-value	0.001	0.974	0.002	0.010	0.044	0.001
		High poverty	Difference	–0.093***	–0.003	–0.011	–0.070***	–0.122***	–0.151***
			p-value	0.001	0.812	0.499	0.007	0.001	0.001
		Hispanic	Difference	–0.119***	–0.031	–0.006	–0.087***	–0.096***	–0.140***
			p-value	0.001	0.124	0.794	0.001	0.001	0.001
10	Reading	All	Difference	–0.038***	–0.077***	–0.173***	–0.124***	–0.163***	–0.104***
			p-value	0.001	0.001	0.001	0.001	0.001	0.001
		Black	Difference	–0.078***	–0.087***	–0.184***	–0.134***	–0.222***	–0.133***
			p-value	0.001	0.001	0.001	0.001	0.001	0.001

Table B.1—Continued

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
10, continued		High poverty	Difference	–0.072***	–0.052***	–0.189***	–0.162***	–0.181***	–0.151***
			p-value	0.001	0.002	0.001	0.001	0.001	0.001
		Hispanic	Difference	–0.024	–0.051***	–0.132***	–0.101***	–0.162***	–0.073*
			p-value	0.352	0.002	0.001	0.001	0.001	0.084
High school test	Reading	All	Difference	–0.070***	–0.034***	–0.063***	–0.110***	–0.126***	–0.078***
			p-value	0.001	0.001	0.001	0.001	0.001	0.004
		Black	Difference	–0.116***	–0.036***	–0.108***	–0.083***	–0.115***	–0.113***
			p-value	0.001	0.001	0.001	0.001	0.001	0.002
		High poverty	Difference	–0.080***	–0.012	–0.061***	–0.099***	–0.135***	–0.118***
			p-value	0.001	0.211	0.001	0.001	0.001	0.001
		Hispanic	Difference	–0.077***	–0.016	–0.037***	–0.066***	–0.102***	–0.074***
			p-value	0.001	0.402	0.001	0.001	0.001	0.001

Table B.1—Continued

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
High school	Dropout rate (%)	All	Difference	0.880**	0.030	–2.100	–0.200	3.380***	0.720
			p-value	0.011	0.860	0.295	0.576	0.001	0.187
	Grad rate (%)	All	Difference	N/A	N/A	N/A	N/A	N/A	N/A
			p-value						

NOTE: Difference rows represent the difference between the treatment and its comparison group for each outcome in each year. Difference values of 0.000 indicate differences of less than 0.0005. High school tests indicate the average for tests taken in grade 9 or 10. For the graduation and dropout rates, we used a logit model to estimate the predicted trends to take into account the bounded range of these estimates. The shaded column presents the effects for school year 2008–2009, the last school year before implementation. Statistically significant estimates for 2008–2009 indicate the existence of differential trends between the Intensive Partnership site and those in the rest of the state prior to the start of the initiative. In those cases, the impact estimates for the other years should be interpreted with caution because it is difficult to extrapolate with certainty what those trends would have been in the absence of the initiative. In 2011, Florida started to implement the FCAT 2.0. The FCAT 2.0 does not administer the math exam in grade 9 or 10. *** = statistically significant at the 1% level. ** = statistically significant at the 5% level. * = statistically significant at the 10% level.

Table B.2
Pittsburgh Public Schools Impact Estimates, by Grade, Subgroup, and Year

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
3	Math	All	Difference	−0.028	−0.070***	−0.067*	−0.095	0.112***	0.137***
			p-value	0.439	0.001	0.096	0.137	0.006	0.001
		Black	Difference	−0.017	−0.071***	0.024	−0.078*	0.140***	0.149***
			p-value	0.738	0.001	0.692	0.095	0.004	0.001
		High poverty	Difference	−0.005	−0.043***	−0.032	−0.108**	0.166***	0.182***
			p-value	0.860	0.005	0.478	0.031	0.001	0.001
	Reading	All	Difference	−0.022	−0.066***	−0.115***	−0.092*	−0.075**	−0.009
			p-value	0.497	0.001	0.001	0.063	0.020	0.773
		Black	Difference	−0.027	−0.073***	−0.013	−0.074**	0.005	0.033
			p-value	0.595	0.001	0.777	0.029	0.870	0.326
		High poverty	Difference	0.000	−0.019	0.013	−0.096**	0.007	0.072**
			p-value	0.991	0.132	0.744	0.019	0.840	0.032

Table B.2—Continued

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
4	Math	All	Difference	−0.144***	−0.072***	−0.031	−0.066*	−0.072*	0.114***
			p-value	0.001	0.001	0.326	0.086	0.087	0.001
		Black	Difference	−0.183***	−0.103***	0.033	0.028	0.016	0.289***
			p-value	0.001	0.001	0.423	0.546	0.740	0.001
		High poverty	Difference	−0.079***	−0.046***	−0.019	−0.096**	−0.043	0.157***
			p-value	0.001	0.001	0.563	0.029	0.375	0.001
	Reading	All	Difference	−0.090***	−0.051***	0.038	0.038	−0.055*	0.040
			p-value	0.001	0.001	0.111	0.342	0.094	0.194
		Black	Difference	−0.122***	−0.089***	0.096**	0.174***	0.004	0.156***
			p-value	0.001	0.001	0.017	0.001	0.915	0.001
		High poverty	Difference	−0.074***	−0.038***	0.072***	−0.043	−0.031	0.065
			p-value	0.001	0.001	0.006	0.241	0.390	0.114

Table B.2—Continued

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
5	Math	All	Difference	0.027	–0.003	–0.005	0.011	–0.024	–0.002
			p-value	0.307	0.738	0.867	0.828	0.578	0.962
		Black	Difference	0.040	–0.033*	0.010	0.048	0.035	0.123***
			p-value	0.325	0.059	0.808	0.437	0.423	0.001
		High poverty	Difference	0.041	–0.003	0.024	–0.150***	–0.008	0.010
			p-value	0.107	0.845	0.457	0.001	0.854	0.720
	Reading	All	Difference	0.055***	0.035***	0.106***	0.146***	–0.055	0.005
			p-value	0.001	0.001	0.001	0.005	0.183	0.893
		Black	Difference	0.090**	0.015	0.117***	0.215***	–0.007	0.116***
			p-value	0.021	0.465	0.007	0.001	0.853	0.004
		High poverty	Difference	0.055***	0.024**	0.140***	–0.028	–0.015	0.019
			p-value	0.005	0.012	0.001	0.506	0.707	0.643

Table B.2—Continued

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
6	Math	All	Difference	−0.096***	−0.113***	−0.139***	−0.083**	−0.158***	−0.032
			p-value	0.001	0.001	0.001	0.044	0.002	0.527
		Black	Difference	−0.100**	−0.095***	−0.114***	0.018	−0.140***	0.073
			p-value	0.017	0.001	0.004	0.628	0.001	0.135
		High poverty	Difference	−0.029	−0.066***	−0.063**	0.007	−0.086*	0.037
			p-value	0.314	0.001	0.021	0.901	0.078	0.410
	Reading	All	Difference	−0.023	−0.049***	0.034	0.082***	−0.011	−0.075***
			p-value	0.162	0.001	0.119	0.004	0.715	0.009
		Black	Difference	−0.002	−0.054***	0.046	0.181***	−0.074***	0.066*
			p-value	0.948	0.001	0.215	0.001	0.006	0.056
		High poverty	Difference	0.008	−0.018	0.072***	0.109***	0.009	−0.013
			p-value	0.676	0.174	0.002	0.001	0.819	0.622

Table B.2—Continued

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
7	Math	All	Difference	−0.024	0.032**	−0.004	−0.110**	−0.086***	0.119***
			p-value	0.395	0.027	0.822	0.021	0.003	0.002
		Black	Difference	0.061*	0.051***	−0.067*	−0.072	−0.094**	0.108**
			p-value	0.069	0.002	0.053	0.168	0.015	0.014
		High poverty	Difference	−0.028	0.039**	0.016	0.021	−0.038	0.189***
			p-value	0.304	0.016	0.412	0.777	0.215	0.001
	Reading	All	Difference	0.043	0.026**	0.023	0.081**	−0.109***	0.048*
			p-value	0.140	0.041	0.130	0.043	0.001	0.054
		Black	Difference	0.159***	−0.002	−0.028	0.152***	−0.071**	0.072*
			p-value	0.001	0.913	0.264	0.001	0.047	0.051
		High poverty	Difference	0.024	0.014	0.054***	0.083**	−0.045***	0.141***
			p-value	0.423	0.272	0.001	0.014	0.002	0.001

Table B.2—Continued

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
8	Math	All	Difference	0.005	0.001	−0.001	0.023	−0.133***	0.050*
			p-value	0.803	0.883	0.955	0.385	0.001	0.052
		Black	Difference	0.002	0.002	0.005	−0.019	−0.123**	0.082**
			p-value	0.961	0.850	0.888	0.693	0.029	0.035
		High poverty	Difference	−0.025	−0.001	0.038	−0.042	−0.027	0.149***
			p-value	0.238	0.930	0.219	0.584	0.378	0.001
	Reading	All	Difference	−0.010	0.027***	−0.014	0.068***	−0.083***	−0.030
			p-value	0.380	0.003	0.431	0.005	0.001	0.183
		Black	Difference	−0.028	0.004	−0.013	0.012	−0.080	0.005
			p-value	0.383	0.835	0.709	0.755	0.101	0.897
		High poverty	Difference	−0.037*	−0.006	−0.009	−0.012	−0.017	0.057***
			p-value	0.098	0.661	0.711	0.711	0.262	0.002

Table B.2—Continued

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
3–8	Math	All	Difference	–0.042***	–0.036***	0.011	–0.020	–0.009	0.104***
			p-value	0.001	0.001	0.631	0.455	0.774	0.001
		Black	Difference	–0.056	–0.060***	–0.022	–0.018	–0.009	0.147***
			p-value	0.154	0.001	0.454	0.563	0.786	0.001
		High poverty	Difference	–0.018	–0.017***	0.022	–0.077*	0.005	0.113***
			p-value	0.157	0.094	0.319	0.083	0.874	0.001
	Reading	All	Difference	–0.015	–0.025***	0.035	0.046*	–0.040	0.021
			p-value	0.182	0.001	0.136	0.084	0.107	0.391
		Black	Difference	–0.019	–0.040***	0.028	0.065**	–0.033	0.076***
			p-value	0.395	0.001	0.262	0.018	0.161	0.005
		High poverty	Difference	0.004	–0.004**	0.055**	–0.022	–0.021	0.031
			p-value	0.789	0.702	0.023	0.542	0.375	0.219

Table B.2—Continued

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
11	Reading	All	Difference	–0.037	–0.029**	–0.007	–0.055**	–0.092***	–0.076**
			p-value	0.234	0.016	0.718	0.039	0.001	0.023
		Black	Difference	–0.039	–0.071***	0.021	0.075*	–0.033	0.152***
			p-value	0.339	0.002	0.449	0.052	0.557	0.006
		High poverty	Difference	0.054	–0.016	0.006	–0.065*	–0.050	0.104**
			p-value	0.141	0.277	0.801	0.068	0.204	0.038
High school	Dropout rate (%)	All	Difference	2.070***	–1.950***	–1.140*	–3.550***	–2.400***	–2.890***
			p-value	0.001	0.001	0.067	0.001	0.001	0.001
	Grad rate (%)	All	Difference	–14.430***	3.700***	11.910***	4.940**	4.670***	3.890*
			p-value	0.001	0.001	0.001	0.049	0.001	0.084

NOTE: Difference rows represent the difference between the treatment and its comparison group for each outcome in each year. Difference values of 0.000 indicate differences of less than 0.0005. The test scores for grade 11 apply to the Keystone Exam, which is an EOC exam that tests for specific subjects (algebra I for math and literature for reading). The Keystone Exam for math is less standardized across schools, so we do not include it in the analysis. For the graduation and dropout rates, we used a logit model to estimate the predicted trends to take into account the bounded range of these estimates. The shaded column presents the effects for school year 2008–2009, the last school year before implementation. Statistically significant estimates for 2008–2009 indicate the existence of differential trends between the Intensive Partnership site and those in the rest of the state prior to the start of the initiative. In those cases, the impact estimates for the other years should be interpreted with caution because it is difficult to extrapolate with certainty what those trends would have been in the absence of the initiative. *** = statistically significant at the 1% level. ** = statistically significant at the 5% level. * = statistically significant at the 10% level.

Table B.3
Memphis City Schools Impact Estimates, by Grade, Subgroup, and Year

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
3	Math	All	Difference	0.022	−0.277***	−0.162	−0.129	−0.124*	−0.032
			p-value	0.650	0.002	0.140	0.165	0.063	0.762
	Reading	All	Difference	−0.067	−0.166**	−0.212**	−0.145**	−0.128***	0.016
			p-value	0.138	0.026	0.043	0.040	0.008	0.856
4	Math	All	Difference	0.002	−0.029	−0.165*	−0.083	−0.024	−0.134
			p-value	0.977	0.728	0.071	0.517	0.886	0.254
	Reading	All	Difference	−0.174**	−0.054	−0.103	−0.119	−0.166***	−0.125***
			p-value	0.016	0.208	0.113	0.178	0.004	0.003
5	Math	All	Difference	0.025	−0.166	−0.212**	−0.322**	−0.253	−0.048
			p-value	0.603	0.103	0.042	0.020	0.116	0.821
	Reading	All	Difference	−0.029	−0.132***	−0.184***	−0.085	−0.201**	0.027
			p-value	0.234	0.001	0.004	0.327	0.037	0.836

Table B.3—Continued

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
6	Math	All	Difference	−0.062	−0.194***	−0.306***	−0.300***	−0.142	−0.296**
			p-value	0.195	0.001	0.001	0.001	0.178	0.039
	Reading	All	Difference	−0.134***	−0.089***	−0.192***	−0.192***	−0.103**	−0.214***
			p-value	0.001	0.002	0.001	0.007	0.034	0.004
7	Math	All	Difference	0.029	−0.140***	−0.270***	−0.224**	−0.099	−0.264**
			p-value	0.642	0.004	0.001	0.018	0.344	0.027
	Reading	All	Difference	−0.086	−0.110**	−0.243***	−0.160***	−0.152**	−0.034
			p-value	0.182	0.035	0.001	0.006	0.026	0.466
8	Math	All	Difference	0.019	−0.098	−0.238***	−0.288***	−0.205*	−0.190
			p-value	0.732	0.118	0.001	0.001	0.063	0.178
	Reading	All	Difference	−0.023	−0.012	−0.159***	−0.080	−0.091	−0.077
			p-value	0.716	0.708	0.001	0.197	0.223	0.189
3–8	Math	All	Difference	0.024	−0.146***	−0.173***	−0.204***	−0.154*	−0.177*
			p-value	0.526	0.002	0.001	0.003	0.051	0.092
	Reading	All	Difference	−0.095***	−0.103***	−0.133***	−0.132***	−0.148***	−0.022
			p-value	0.003	0.001	0.002	0.008	0.001	0.617

Table B.3—Continued

Grade	Subject	Group	Statistic	2009 (Placebo Effect)	2010	2011	2012	2013	2014
Elementary school	Attendance (%)	All	Difference	0.023	–0.320	–0.880***	–0.810***	–1.290**	–1.320***
			p-value	0.308	0.141	0.001	0.004	0.014	0.004
	Promotion (%)	All	Difference	2.180***	1.020	–1.690***	–1.790***	5.000	–2.060
			p-value	0.008	0.118	0.008	0.001	0.474	0.841
High school	Dropout Rate (%)	All	Difference	2.390	–0.040	–3.890**	4.150*	–0.430	0.540
			p-value	0.382	0.984	0.026	0.061	0.762	0.834
	Graduation Rate (%)	All	Difference	–0.790	–7.780*	0.490	–5.640**	–9.880***	–11.500***
			p-value	0.774	0.053	0.854	0.013	0.002	0.001
	Attendance (%)	All	Difference	0.080	1.240	0.670	0.990*	0.570	3.780***
			p-value	0.871	0.104	0.351	0.085	0.447	0.002

NOTE: Difference rows represent the difference between the treatment and its comparison group for each outcome in each year. Difference values of 0.000 indicate differences of less than 0.0005. For the graduation, dropout, attendance, and promotion rates, we used a logit model to estimate the predicted trends to take into account the bounded range of these estimates. The shaded column presents the effects for school year 2008–2009, the last school year before implementation. Statistically significant estimates for 2008–2009 indicate the existence of differential trends between the Intensive Partnership site and those in the rest of the state prior to the start of the initiative. In those cases, the impact estimates for the other years should be interpreted with caution because it is difficult to extrapolate with certainty what those trends would have been in the absence of the initiative. *** = statistically significant at the 1% level. ** = statistically significant at the 5% level. * = statistically significant at the 10% level.

APPENDIX C

Specific Practices for Levers of Implementation

This appendix lists the practices evaluated for each lever of implementation of the Intensive Partnerships initiative. The information comes from Stecher et al., 2016.

Teacher Evaluation

- Have principals or other administrators observe teachers.
- For at least some teachers, have an additional set of evaluators observe.
- Use student or parent surveys or other measures of teacher effectiveness.
- For subjects and grades with state tests, use individual value-added models or student growth percentile scores.
- For subjects and grades with no state tests or other measures of student growth, use individual value-added models or student growth percentile scores.
- Combine multiple measures using weights.
- Establish a data warehouse for teacher-evaluation data.

Staffing

- Conduct early or expedited recruiting or hiring for high-need positions.
- Conduct early hiring for all vacancies.

- Have schools make final hiring decisions
- Train administrators to make good hiring decisions.
- Use a new applicant screening model based on the teacher-effectiveness rubric.
- Offer incentives to work in high-need schools and classrooms.
- Do not let seniority heavily influence transfers or furloughs.
- Have school leaders make final decisions about which teachers are placed in their schools.
- Link tenure retention to effectiveness ratings.
- Use effectiveness ratings as a basis for dismissal.
- Have schools make final decisions about teacher retention and dismissal.

Professional Development

- Use evaluation data to identify teacher development needs.
- Offer professional development designed to improve specific teaching skills measured in the evaluation.
- Link coaching and mentoring feedback to evaluation components.
- Provide instruction, mentoring, coaching, or academies for new teachers.
- Have supervisors oversee teachers' professional-development participation.
- Create an electronic system for professional-development data collection.

Compensation

- Award bonuses, stipends, and salary increments based on individual effectiveness measures.
- Do not use a traditional step-based salary schedule exclusively.
- Give bonuses or salary increments for high-need positions.
- Give incentives for desired teacher behavior.
- Create positions for effective teachers with different responsibilities.

Bibliography

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller, "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, Vol. 105, No. 490, June 1, 2010, pp. 493–505.

———, "SYNTH: Stata Module to Implement Synthetic Control Methods for Comparative Case Studies," Boston College Department of Economics, Statistical Software Component S457334, October 6, 2011, revised January 28, 2014.

Abadie, Alberto, and Javier Gardeazabal, "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, Vol. 93, No. 1, 2003, pp. 113–132.

Abadie, Alberto, and Guido W. Imbens, "On the Failure of the Bootstrap for Matching Estimators," *Econometrica*, Vol. 6, No. 76, November 2008, pp. 1537–1557.

———, *Matching on the Estimated Propensity Score*, Cambridge, Mass.: National Bureau of Economic Research, Working Paper 15301, August 2009. As of March 30, 2016:
http://www.nber.org/papers/w15301

Achievement School District, home page, undated. As of April 19, 2016:
http://achievementschooldistrict.org/

Baird, Matthew D., John Engberg, Gerald Paul Hunter, and Benjamin K. Master, *Access to Effective Teaching: The Intensive Partnerships for Effective Teaching Through 2013–2014*, Santa Monica, Calif.: RAND Corporation, RR-1295/4-BMGF, 2016.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics*, Vol. 119, No. 1, 2004, pp. 249–275.

Betts, Julian, and Y. Emily Tang, *The Effect of Charter Schools on Student Achievement: A Meta-Analysis of the Literature*, University of Washington Bothell, Center on Reinventing Public Education, National Charter School Research Project, October 2011. As of March 30, 2016:
http://www.crpe.org/publications/
effect-charter-schools-student-achievement-meta-analysis-literature

Bill & Melinda Gates Foundation, "Foundation Commits $335 Million to Promote Effective Teaching and Raise Student Achievement: Bill & Melinda Gates Foundation," press release, November 2011. As of March 30, 2016:
http://www.gatesfoundation.org/Media-Center/Press-Releases/2009/11/
Foundation-Commits-$335-Million-to-Promote-Effective-Teaching-and-Raise-Student-Achievement

Bloom, Howard S., Carolyn J. Hill, Alison Rebeck Black, and Mark W. Lipsey, *Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions*, MDRC Working Papers on Research Methodology, October 2008. As of March 30, 2016:
http://www.mdrc.org/publication/performance-trajectories-and-performance-gaps-achievement-effect-size-benchmarks

Borman, Geoffrey D., Gina M. Hewes, Laura T. Overman, and Shelly Brown, "Comprehensive School Reform and Achievement: A Meta-Analysis," *Review of Educational Research*, Vol. 73, No. 2, Summer 2003, pp. 125–230.

Carlson, Deven, Geoffrey D. Borman, and Michelle Robinson, "A Multistate District-Level Cluster Randomized Trial of the Impact of Data-Driven Reform on Reading and Mathematics Achievement," *Educational Evaluation and Policy Analysis*, Vol. 33, No. 3, September 2011, pp. 378–398.

Center on Great Teachers and Leaders, American Institutes for Research, "Databases on State Teacher and Principal Evaluation Policies," undated. As of March 31, 2016:
http://resource.tqsource.org/stateevaldb/

Conley, Timothy G., and Christopher R. Taber, "Inference with 'Difference in Differences' with a Small Number of Policy Changes," *Review of Economics and Statistics*, Vol. 93, No. 1, February 2011, pp. 113–125.

Dehejia, Rajeev H., and Sadek Wahba, "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, Vol. 94, No. 448, 1999, pp. 1053–1062.

Doherty, Kathryn M., and Sandi Jacobs, *State of the States 2013: Connect the Dots—Using Evaluations of Teacher Effectiveness to Inform Policy and Practice*, Washington, D.C.: National Council on Teacher Quality, October 2013. As of March 31, 2016:
http://www.nctq.org/dmsStage/
State_of_the_States_2013_Using_Teacher_Evaluations_NCTQ_Report

Gates, Susan M., Laura S. Hamilton, Paco Martorell, Susan Burkhauser, Paul Heaton, Ashley Pierson, Matthew D. Baird, Mirka Vuollo, Jennifer J. Li, Diana Lavery, Melody Harvey, and Kun Gu, *Preparing Principals to Raise Student Achievement: Implementation and Effects of the New Leaders Program in Ten Districts*, Santa Monica, Calif.: RAND Corporation, RR-507-NL, 2014. As of March 30, 2016:
http://www.rand.org/pubs/research_reports/RR507.html

Gutierrez, Italo, Gabriel Weinberger, and John Engberg, *The Overall Impact of the Intensive Partnership for Effective Teaching Through 2012–13*, Santa Monica, Calif.: RAND Corporation, March 25, 2015. Not available to the general public.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd, "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, Vol. 64, No. 4, October 1997, pp. 605–654.

Hill, Carolyn J., Howard S. Bloom, Alison Rebeck Black, and Mark W. Lipsey, *Empirical Benchmarks for Interpreting Effect Sizes in Research*, MDRC Working Papers on Research Methodology, July 2007. As of March 30, 2016:
http://www.mdrc.org/publication/
empirical-benchmarks-interpreting-effect-sizes-research

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder, "Efficient Estimation of Average Treatment Effect Using the Estimated Propensity Score," *Econometrica*, Vol. 71, No. 4, July 2003, pp. 1161–1189.

Martorell, Paco, Ethan Scherer, and John Engberg, *Interim Findings from the Evaluation of the Intensive Partnership for Effective Teaching: Results on the Overall Impact Through 2011*, Santa Monica, Calif.: RAND Corporation, November 8, 2012. Not available to the general public.

———, *Interim Findings from the Evaluation of the Intensive Partnership for Effective Teaching: An Investigation into Using School-Level Models for the Overall Impact Analysis*, Santa Monica, Calif.: RAND Corporation, February 12, 2013a. Not available to the general public.

———, *Interim Findings from the Evaluation of the Intensive Partnership for Effective Teaching: Results on the Overall Impact Through 2012 for the College Ready Promise*, Santa Monica, Calif.: RAND Corporation, February 13, 2013b. Not available to the general public.

———, *The Overall Impact of the Intensive Partnership for Effective Teaching Through 2012–13*, Santa Monica, Calif.: RAND Corporation, September 1, 2014. Not available to the general public.

Moulton, Brent R., "Random Group Effects and the Precision of Regression Estimates," *Journal of Econometrics*, Vol. 32, No. 3, August 1986, pp. 385–397.

National Council on Teacher Quality, "State-by-State Summary," c. 2015. As of February 12, 2016:
http://www.nctq.org/statePolicy/2015/statePolicyNationalSummary.do

NCTQ—*See* National Council on Teacher Quality.

Public Law 89-10, Elementary and Secondary Education Act, April 11, 1965.

Rosenbaum, Paul R., and Donald B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, Vol. 70, No. 1, 1983, pp. 41–55.

Schanzenbach, Diane Whitmore, "What Have Researchers Learned from Project STAR?" *Brookings Papers on Education Policy*, No. 9, 2006–2007, pp. 205–228.

Stecher, Brian M., Michael S. Garet, Laura S. Hamilton, Elizabeth D. Steiner, Abby Robyn, Jeffrey Poirier, Deborah Holtzman, Eleanor S. Fulbeck, Jay Chambers, and Iliana Brodziak de los Reyes, *Implementation: The Intensive Partnerships for Effective Teaching Through 2013–2014*, Santa Monica, Calif.: RAND Corporation, RR-1295-BMGF, 2016.

Strunk, Katharine O., and Andrew McEachin, "More Than Sanctions: Closing Achievement Gaps Through California's Use of Intensive Technical Assistance," *Educational Evaluation and Policy Analysis*, Vol. 36, No. 3, September 2014, pp. 281–306.

This interim report presents estimates of the overall effect that the Bill & Melinda Gates Foundation's Intensive Partnerships for Effective Teaching initiative has had on student outcomes through the 2013–2014 school year. The initiative's aim is to encourage and support strategic human-capital reforms that are intended to improve the ways in which "teachers are recruited, evaluated, supported, retained, and rewarded." The reform's cornerstone is the development and implementation of teacher-evaluation systems based on student achievement growth; structured classroom observations by principals or trained peers; and other inputs, such as student or parent surveys. These evaluations are used to guide personnel practices in staffing, professional development, and compensation and career-ladder decisions with the goal of giving every student access to highly effective teachers. The report covers Hillsborough County Public Schools (HCPS) in Florida, Memphis City Schools (MCS) in Tennessee, and Pittsburgh Public Schools (PPS) in Pennsylvania. The initiative has not had the dramatic positive effects on student outcomes for which the foundation had hoped. The authors' estimates of impact are small and not statistically significant in the first three or four years after the intervention (with the exception of Memphis City Schools, which fared significantly worse in the first years). However, impact estimates were increasing in 2013–2014 (the fifth year after the intervention began) in many sites, which suggests that the reforms might be on the way to having a positive effect.

EDUCATION

www.rand.org

$24.50

ISBN-10 0-8330-9552-8
ISBN-13 978-0-8330-9552-7

RR-1295/3-1-BMGF