Improving Teaching Effectiveness

FINAL REPORT

The INTENSIVE PARTNERSHIPS *for* EFFECTIVE TEACHING *Through 2015–2016*

BRIAN M. STECHER, DEBORAH J. HOLTZMAN, MICHAEL S. GARET, LAURA S. HAMILTON, JOHN ENGBERG, ELIZABETH D. STEINER, ABBY ROBYN, MATTHEW D. BAIRD, ITALO A. GUTIERREZ, EVAN D. PEET, ILIANA BRODZIAK DE LOS REYES, KAITLIN FRONBERG, GABRIEL WEINBERGER, GERALD PAUL HUNTER, JAY CHAMBERS

RAND CORPORATION

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

For more information on this publication, visit www.rand.org/t/RR2242

Library of Congress Cataloging-in-Publication Data is available for this publication.
ISBN: 978-1-9774-0079-6

Published by the RAND Corporation, Santa Monica, Calif.
© Copyright 2018 RAND Corporation
RAND® is a registered trademark.

Cover: julief514/GettyImages with DragonImages/GettyImages.

This work is licensed under a Creative Commons Attribution 4.0 International License. All users of the publication are permitted to copy and redistribute the material in any medium or format and transform and build upon the material, including for any purpose (including commercial) without further permission or fees being required. For additional information, please visit http://creativecommons.org/licenses/by/4.0/

The RAND Corporation is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Support RAND
Make a tax-deductible charitable contribution at
www.rand.org/giving/contribute

www.rand.org

Preface

The Bill & Melinda Gates Foundation launched the Intensive Partnerships for Effective Teaching (IP) in the 2009–2010 school year.[1] After thorough screening, the foundation identified seven IP sites—three school districts and four charter management organizations—to implement teacher-evaluation systems and related human-relations reforms patterned on a common design over a six-year period.[2] The foundation also selected the RAND Corporation and its partner, the American Institutes for Research (AIR), to evaluate the IP efforts. The evaluation began in July 2010 and collected data annually for six school years, from 2010–2011 through 2015–2016.

The RAND/AIR team produced three interim reports and two journal articles:[3]

- Brian M. Stecher, 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, *Improving Teaching Effectiveness: Implementation—The Intensive Partnerships for Effective Teaching*

[1] In the interest of concision, we use the year pair alone, without *school year*, to indicate a school year (e.g., *2009–2010* is the school year running from the fall of 2009 through the spring of 2010.

[2] We use the word *site* to describe any of the three school districts or four charter management organizations that received funding from the foundation to implement the IP initiative.

[3] We also prepared a series of internal reports to the foundation and the sites.

Through 2013–2014, Santa Monica, Calif.: RAND Corporation, RR-1295-BMGF, 2016

- Matthew D. Baird, John Engberg, Gerald Hunter, and Benjamin Master, *Improving Teaching Effectiveness: Access to Effective Teaching—The Intensive Partnerships for Effective Teaching Through 2013–2014*, Santa Monica, Calif.: RAND Corporation, RR-1295/4-BMGF, 2016
- Italo A. Gutierrez, Gabriel Weinberger, and John Engberg, *Improving Teaching Effectiveness: Impact on Student Outcomes—The Intensive Partnerships for Effective Teaching Through 2013–2014*, Santa Monica, Calif.: RAND Corporation, RR-1295/3-1-BMGF, 2016
- Gema Zamarro, John Engberg, Juan Esteban Saavedra, and Jennifer Steele, "Disentangling Disadvantage: Can We Distinguish Good Teaching from Classroom Composition?" *Journal of Research on Educational Effectiveness*, Vol. 8, No. 1, 2015, pp. 84–111
- Brian Stecher, Mike Garet, Deborah Holtzman, and Laura Hamilton, "Implementing Measures of Teacher Effectiveness," *Phi Delta Kappan*, Vol. 94, No. 3, November 2012, pp. 39–43.

The present report encompasses the full IP initiative, describing its design, implementation, and impact on teaching effectiveness and student outcomes. This study was undertaken by RAND Education and AIR. The Bill & Melinda Gates Foundation sponsored the research. It should be of interest to researchers, policymakers, and practitioners who want to understand the potential benefits and challenges of adopting new teacher-evaluation systems and associated strategic human-relations policies.

RAND Education is a unit of the RAND Corporation that conducts research on prekindergarten, kindergarten–12th grade, and higher education issues, such as assessment and accountability, choice-based and standards-based school reform, vocational training, and the value of arts education and policy in sustaining and promoting well-rounded communities.

More information about RAND can be found at www.rand.org. Questions about this report should be directed to Brian_Stecher@rand.org, and questions about RAND Education should be directed to education@rand.org.

Established in 1946, AIR is an independent, nonpartisan, not-for-profit organization that conducts behavioral and social science research on important social issues and delivers technical assistance, both domestically and internationally, in the areas of education, health, and workforce productivity.

More information about AIR can be found at www.air.org.

Contents

Preface iii
Figures xiii
Tables xxiii
Summary xxv
Acknowledgments li
Abbreviations lv

CHAPTER ONE
Introduction 1
In Brief 1
Overview of the IP Initiative 2
The IP Theory of Action 4
Overview of the IP Evaluation 12
Remainder of the Report 19
Next Steps 21

CHAPTER TWO
The IP Sites 23
In Brief 23
Introduction 23
Demographic Characteristics of the IP Sites 24
The Districts 24
CMOs 40

CHAPTER THREE
Measures of Effective Teaching 53

In Brief 53
Introduction 54
Best Practices for Measuring TE as Part of a Teacher-Evaluation System 54
IP Sites' TE Measures 57
Distribution of TE Ratings over Time 75
Implementation Challenges 78
Teacher and SL Reactions to Sites' Evaluation Systems 83
Summary 109

CHAPTER FOUR
Recruitment, Hiring, Placement, and Transfer Policies 113
In Brief 113
Introduction 114
Best Practices with Respect to Recruitment and Hiring 115
IP Recruitment, Hiring, Placement, and Transfer Policies 119
Implementation Challenges 133
Teacher and SL Reactions to Recruitment, Hiring, Placement, and Transfer Policies 136
Teacher Outcomes Associated with Recruitment, Hiring, Placement, and Transfer Policies 151
Summary 152

CHAPTER FIVE
Tenure and Dismissal Policies 155
In Brief 155
Introduction 156
Best Practices for Tenure and Dismissal 156
IP Tenure and Dismissal Policies 162
Implementation Challenges 172
Teacher and SL Reactions to Tenure and Dismissal Policies 175
Summary 196

CHAPTER SIX
PD Policies 199
In Brief 199

Introduction 200
Best Practices for Evaluation-Linked PD 200
Intensive Partnership PD Policies 206
Implementation Challenges 216
Teacher and SL Reactions to PD Policies 219
Summary 243

CHAPTER SEVEN
Compensation Policies 247
In Brief 247
Introduction 248
Best Practices for Teacher Compensation 249
IP Compensation Policies 252
Implementation Challenges 266
Teacher and SL Reactions to Compensation Policies 268
Did the IP Compensation Policies Increase Pay for Effective Teachers? 282
Summary 289

CHAPTER EIGHT
CL Policies 291
In Brief 291
Introduction 292
Best Practices with Respect to CLs 293
IP CL Policies and How the Sites Defined Them 296
Implementation Challenges 313
Teacher and SL Reactions to CL Policies 317
Teacher Outcomes Associated with CL Policies 330
Summary of the CL Lever 330

CHAPTER NINE
Resources Invested in the IP Initiative 333
In Brief 333
Introduction 334
Evidence on Resources Needed to Implement the Best Practices Described in Previous Chapters 335

Total Spending on the IP Initiative 340
Proportion of Sites' Total Budgets Spent on the IP Initiative in 2015–2016 344
Change over Time in Funding and Funding Source 347
Spending by Implementation Lever 350
Teachers' and SLs' Allocation of Time 353
Comprehensive Cost of Operating the Teacher-Evaluation System in 2014–2015 362
Summary 364

CHAPTER TEN
The Effectiveness of Newly Hired Teachers 367
In Brief 367
Introduction 367
IP Policies Designed to Influence New-Hire Effectiveness 368
SL Perceptions 370
Changes in the Measured Effectiveness of New Hires 375
Summary 388

CHAPTER ELEVEN
Retention of Effective Teachers 389
In Brief 389
Introduction 390
IP Policies Designed to Influence Retention and Issues Assessing Their Impact 390
Teacher Intentions and SL Perceptions 393
Changes in Retention 399
Summary 422

CHAPTER TWELVE
The Initiative's Effects on TE and on LIM Students' Access to Effective Teaching 427
In Brief 427
Introduction 428
Changes in the Distribution of TE 429
SL Perceptions About Student Access to Effective Teachers 433

Estimates of LIM Students' Access to Effective Teachers 438
Summary 448

CHAPTER THIRTEEN
The Initiative's Impact on Student Outcomes 451
In Brief 451
Introduction 451
Teachers' and SLs' Perceptions of Impact 452
Impact on Measured Student Outcomes 458
Summary 481

CHAPTER FOURTEEN
Summary and Conclusions 487
In Brief 487
Introduction 489
What Changes Did the Sites Make? 489
What Outcomes Were Achieved? 492
How Much Did the Sites Vary in Implementation and Outcomes? 493
Are the IP Levers Being Sustained? 496
Potential Limitations of the Methods Used to Estimate Impact 497
Discussion 498
Final Thought 504

References 505

Figures

1.1. The IP Theory of Action 5
1.2. How Levers Are Expected to Improve the Distribution of Effective Teaching 9
1.3. Average Proportion of IP Levers Implemented, Springs 2010–2016 16
3.1. Effectiveness Rating Distributions, 2012–2016, by Site 76
3.2. Teachers Reporting That Student Achievement or Growth Was a Component of Their Evaluation, Springs 2013–2016 86
3.3. Teachers Reporting That Evaluation Components Were Valid Measures of Effectiveness to a Large or Moderate Extent, Spring 2016 88
3.4. Teachers' Agreement That the Way Their Teaching Was Being Evaluated Accurately Reflected the Quality of Their Teaching, Springs 2013–2016 89
3.5. Teachers Reporting That Prior-Year Overall Evaluation Ratings Were Moderately or Highly Accurate, by Prior-Year TE Rating, Spring 2015 90
3.6. Teachers' Agreement That the Teacher-Evaluation System Did a Good Job Distinguishing Effective from Ineffective Teachers, Springs 2013–2016 91
3.7. SLs' Agreement That the Teacher-Evaluation System Did a Good Job Distinguishing Effective from Ineffective Teachers, Springs 2012–2016 92
3.8. Teachers' Agreement with Statements About Observations, Spring 2016 96
3.9. Teachers' Agreement with Statements About the Use of Student Feedback in Teachers' Evaluations, Spring 2015 100

3.10. Teachers' Agreement with Statements About the Fairness of the Evaluation System, Spring 2016 103
3.11. PUC Teachers' Agreement with Statements About the Fairness of the Evaluation System, Springs 2014–2016 104
3.12. Teachers' Agreement with Statements About the Evaluation System's Effects on Their Teaching, Spring 2016 105
3.13. Teachers' Agreement with Statements About the Usefulness of Feedback from Evaluation Components, Spring 2016 107
3.14. Teachers' Agreement That, "After My Teaching Is Observed, I Receive Useful and Actionable Feedback," by Teacher Experience Level, Spring 2016 108
4.1. Teachers' Agreement That "Teachers' Evaluation Results Should Be Factored in to Decisions About How Teachers Are Placed at Schools," Springs 2011, 2013, and 2015 143
4.2. Teachers' Agreement That "I Would Rather Leave Teaching Than Be Forced to Transfer to a School That Is Not of My Choosing," by Teacher Experience Level, Spring 2015 145
4.3. Teachers' Agreement That "I Would Rather Leave Teaching Than Be Forced to Transfer to a School That Is Not of My Choosing," by Prior-Year TE Rating, Spring 2015 146
4.4. Teachers' Agreement That "I Would Rather Leave Teaching Than Be Forced to Transfer to a School That Is Not of My Choosing," by School's Percentage of Students with LIM Status, Spring 2015 147
4.5. District SLs' Agreement That, "More Often Than Is Good for My School, Good Teachers Are Forced to Leave My Staff Because They Are 'Bumped' Due to Seniority or Other Policies," Springs 2013–2016 150
4.6. SLs Reporting That Their School Had Teachers Assigned to the School Based on Site Policy, Springs 2013–2016 151
4.7. District SLs' Agreement That, Overall, They Were Satisfied with the Performance of Teachers Who Transferred into Their School from Elsewhere in the District That Year, Springs 2011–2016 152

5.1. Teachers Reporting That Their Site Granted Tenure, Springs 2011, 2013, and 2015 176
5.2. Teachers' Agreement with Statements About Receiving Tenure, Springs 2011, 2013, and 2015 179
5.3. Teachers' and SLs' Agreement That Keeping Tenure Had Become More Difficult, Spring 2015 180
5.4. Teachers Reporting That Their Evaluation Results Would Be Used to a Large, Moderate, or Small Extent to Determine Whether They Would Enter into Some Type of Probationary Status (e.g., Employee Improvement Plan), Springs 2013–2016 181
5.5. Teachers Reporting That Their Evaluation Results Would Be Used to a Large, Moderate, or Small Extent to Determine Whether They Were Qualified to Continue Teaching, Springs 2013–2016 182
5.6. Teachers' Agreement That They Had a Clear Understanding of the Criteria Used to Make Decisions About Teacher Termination, Springs 2011, 2013, and 2015 185
5.7. Teachers' and SLs' Agreement That Tenure Should Be Linked to Teachers' Evaluation Results, Springs 2011, 2013, and 2015 187
5.8. SLs' Agreement That the Termination Procedures in Their Site Were So Burdensome That Most School Administrators Tried to Avoid Using Them, Springs 2011–2016 188
5.9. SLs' Agreement That "Teachers Who Are in Danger of Being Dismissed Based on Performance Currently Have Sufficient Opportunities and Time to Improve Prior to Actually Being Dismissed," Springs 2011–2016 190
5.10. Teachers' and SLs' Agreement That, as Implemented in Their District, Tenure Protected Bad or Ineffective Teachers, Springs 2011, 2013, and 2015 192
5.11. Teachers' Agreement That "I Am Really Worried About Being Dismissed," Springs 2011, 2013, and 2015 193
5.12. Teachers' Agreement That, "If I Received a Very Low Evaluation Rating, I Would Seriously Consider Leaving Teaching," Springs 2013–2016 195

6.1. Teachers' Responses to the Survey Question, "Has Support (Coaching, Professional Development, etc.) Been Made Available to You to Address the Needs Identified by Your [Previous Year's] Evaluation Results?" Springs 2013–2016 222
6.2. Teachers' Responses to the Survey Question, "To What Extent Did Each of the Following Influence What Professional Development You Participated in This Year (2015–2016, Including Summer 2016)?" 224
6.3. Teachers' Agreement That Their PD During the 2015–2016 Year Was Aligned with Various Sources, Spring 2016..... 225
6.4. Teachers' Agreement That They Had Had Easy Access to a Catalog of PD Opportunities Aligned with Their Site's Teacher-Observation Rubric, Springs 2013–2016 226
6.5. SLs Reporting That They Provided Teachers with Suggestions for PD Based on Their Observations of the Teachers' Instruction, Springs 2013–2016..... 227
6.6. Teachers' Responses About the Availability of, and Their Participation in, "Instructional Coaching (Provided by School-Based Coaches *or* District Coaches)," Springs 2013–2016 229
6.7. Teachers' Responses About "Receiving Instructional Coaching (Provided by School-Based Coaches *or* District Coaches)," Springs 2013–2016..... 230
6.8. Teachers' Agreement That Their PD Experiences in That School Year Had Been Relevant to the Needs of Their Students, Springs 2011–2016..... 232
6.9. SLs' Agreement That Site-Provided PD Offerings Had Enough Slots and Availability to Accommodate All Teachers Who Wished or Needed to Participate, Springs 2012–2016 234
6.10. Teachers' Agreement That Their PD Experiences in That School Year Had Been Useful for Improving Their Instruction, Springs 2011–2016..... 235
6.11. Alliance Teachers Reporting Enhanced Skills and Knowledge, in Various Areas, Resulting from PD, Springs 2011–2016..... 237

6.12. Teachers' Perceptions of the Usefulness of Three Forms of PD, Spring 2016 239
6.13. SLs' Agreement That the Site Did a Good Job of Meeting the PD Needs of Teachers New to the Profession, Springs 2011–2016..... 244
7.1. Teachers Reporting That Their Evaluation Results Would Be Used to a Large, Moderate, or Small Extent to Determine Whether They Would Receive a Monetary Bonus on Top of Their Salary, Springs 2013–2016..... 269
7.2. Teachers Reporting That Their Evaluation Results Would Be Used to a Large, Moderate, or Small Extent to Determine How Much of a Salary Increase They Would Receive for the Subsequent Year, Springs 2013–2016..... 270
7.3. Teachers' Agreement That Teachers Should Receive Additional Compensation for Demonstrating Outstanding Teaching Skills, Springs 2013–2015..... 273
7.4. Teachers' Agreement That Their Site's Teacher Compensation System Was "Reasonable, Fair, and Appropriate," Springs 2014–2016..... 275
7.5. Teachers' Agreement That Their Site's Teacher Compensation System Was "Reasonable, Fair, and Appropriate," by Perception of Salary Adequacy, Spring 2016..... 278
7.6. Teachers' Agreement That Their Site's Compensation System Motivated Them to Improve Their Teaching, Springs 2014–2016 279
7.7. SLs' Agreement That Their Site's Compensation System Was Helping Their School Retain Good Teachers, Springs 2015 and 2016 281
7.8. HCPS Total Compensation, by District Composite TE Level and Study VAM Score..... 284
7.9. PPS Total Compensation, by District Composite TE Level and Study VAM Score..... 285
7.10. HCPS Change in Compensation 288
8.1. Teachers Reporting That Their Site Had or Was Phasing in a CL or Specialized Instructional Positions, Springs 2011–2016..... 319

8.2. Teachers Reporting That There Were Teachers at Their School Who Held Higher-Level CL or Specialized Instructional Positions, Springs 2013–2016 320
8.3. Teachers' Agreement That the Teachers Who Held Higher-Level Career Ladder and Specialized Positions at Their School Were Effective Educators, Selected Site-Years 323
8.4. Teachers' Agreement That the Process by Which Teachers in Their Site Were Selected for the Various CL and Specialized Positions Was Fair, Selected Site-Years 326
8.5. Teachers' Agreement with Statements About Advancing to Higher or Special Positions, Spring 2015 327
8.6. Teachers' Agreement That the Teachers Who Held Higher-Level Positions at Their School Had Helped Them Improve Their Instruction, Selected Site-Years 329
9.1. Overall Expenditures, by Funding Source, November 2009–June 2016 342
9.2. Districts' Per-Pupil Expenditures on the Initiative: Gates Foundation Funding Versus Matching Funds, November 2009–June 2016 348
9.3. CMOs' Per-Pupil Expenditures on the Initiative: Gates Foundation Funding Versus Matching Funds, November 2009–June 2016 349
9.4. Overall Proportion and Total Per-Pupil Expenditures, by Implementation Lever, November 2009–June 2016 351
9.5. Teachers' Overall Time Allocation, 2012–2013 and 2014–2015 354
9.6. SLs' Overall Time Allocation, 2012–2013 and 2014–2015 358
9.7. District Principals' and APs' Time Allocation, 2012–2013 and 2014–2015 360
9.8. CMO Principals' and APs' Time Allocation, 2012–2013 and 2014–2015 361
9.9. Comprehensive Cost of the Teacher-Evaluation Lever in 2014–2015 363
10.1. SLs' Agreement That, Overall, They Were Satisfied with the Performance of Teachers Who Were *Newly Hired* to Their School That Year, Springs 2011–2016 371

10.2. SLs' Satisfaction with Recently Hired Beginning Teachers from Traditional Teacher-Preparation Programs, Springs 2013–2016 373
10.3. SLs' Satisfaction with Recently Hired Beginning Teachers from Alternative Teacher-Preparation Programs, Springs 2013–2016 374
10.4. New Hires in Each Site 377
10.5. HCPS's New-Hire Effectiveness, by VAM Score and Composite TE Level 379
10.6. PPS's New-Hire Effectiveness, by VAM Score and Composite TE Level 381
10.7. SCS's New-Hire Effectiveness, by VAM Score and Composite TE Level 383
10.8. Alliance's New-Hire Effectiveness, by Composite TE Level 384
10.9. Aspire's New-Hire Effectiveness, by VAM Score and Composite TE Level 386
10.10. Green Dot's New-Hire Effectiveness, by Composite TE Level 387
11.1. Teachers Planning to Still Be Teaching in Five Years, Springs 2011, 2013, and 2015 394
11.2. Teachers Planning to Still Be Teaching in Their Current Site in Five Years, by Perception of Salary Adequacy, Spring 2015 396
11.3. Teachers Planning to Still Be Teaching in Their Current Site in Five Years, by Prior-Year TE Rating, Spring 2015 397
11.4. SLs' Agreement That, "More Often Than Is Good for My School, Good Teachers Leave My Staff Because They Perceive Better Opportunities Elsewhere," Springs 2013–2016 398
11.5. District SLs' Agreement That, "More Often Than Is Good for My School, Good Teachers Leave My Staff Because They Perceive Better Opportunities Elsewhere," by School's Percentage of Students with LIM Status, Springs 2014–2016 400
11.6. Numbers of HCPS Teachers and of Those with District Composite TE Levels or Study-Calculated VAM Scores 402
11.7. HCPS Teachers with District Composite TE Levels, by Composite TE Level 403

11.8. Teachers Remaining in HCPS, by Period and by Composite TE Level 405
11.9. Teachers Remaining in HCPS, by Period and VAM Score ... 406
11.10. Numbers of PPS Teachers and of Those with District Composite TE Levels or Study-Calculated VAM Scores 407
11.11. PPS Teachers with District Composite TE Levels, by TE Level 408
11.12. Teachers Remaining in PPS, by Period and Composite TE Level 409
11.13. Teachers Remaining in PPS, by Period and VAM Score 410
11.14. Numbers of SCS Teachers and of Those with District Composite TE or Study-Calculated VAM Scores 411
11.15. SCS Teachers with District Composite TE Levels, by TE Level 412
11.16. Teachers Remaining in SCS, by Period and by District Composite TE Level 413
11.17. Teachers Remaining in SCS, by Period and VAM Score 414
11.18. Numbers of Alliance Teachers and of Those with CMO Composite TE Levels 415
11.19. Alliance Teachers with CMO Composite TE Levels, by TE Level 415
11.20. Teachers Remaining in Alliance, by Period and by CMO Composite TE Level 416
11.21. Numbers of Aspire Teachers and of Those with CMO Composite TE Levels or Study-Calculated VAM Scores 417
11.22. Aspire Teachers with CMO Composite TE Levels, by TE Level 418
11.23. Teachers Remaining in Aspire, by Period and by Composite TE Level 419
11.24. Teachers Remaining in Aspire, by Period and VAM Score ... 420
11.25. Numbers of Green Dot Teachers and of Those with CMO Composite TE Levels 421
11.26. Green Dot Teachers with CMO Composite TE Levels, by TE Level 421
11.27. Teachers Remaining in Green Dot, by Period and by Composite TE Level 422
12.1. District Average Teacher VAM Scores, over Time 431

12.2. SLs' Agreement That They Had Taken Steps to Ensure That Students with the Greatest Needs Were Taught by the Most Effective Teachers, Springs 2014–2016 434
12.3. SLs' Agreement That, "In My School, the Highest-Achieving Students Typically Get the Best Teachers," by School Level, Spring 2016 436
12.4. SLs Reporting That Teacher-Evaluation Results Would Be Used "to Assign Teachers to Classes/Students" (Within Their School), Springs 2013–2016 437
12.5. Teachers Reporting That Their Evaluation Results Would Be Used "to Determine What Classes or Students Within Your School You Will Teach Next Year," Springs 2013–2016 438
12.6. SLs' Responses About the Preference Given to Teachers with Higher Effectiveness Ratings in Assigning Them to Teach Honors and Remedial Classes, Spring 2016 439
12.7. Difference in Access to Effective Teaching in HCPS Between LIM Students and Students Who Are Not LIM Students 442
12.8. Difference in Access to Effective Teaching in PPS Between LIM Students and Students Who Are Not LIM Students 443
12.9. Difference in Access to Effective Teaching in SCS Between LIM Students and Students Who Are Not LIM Students 444
12.10. HCPS Decomposition of Change in Access Coefficients 447
12.11. PPS Decomposition of Change in Access Coefficients 448
12.12. SCS Decomposition of Change in Access Coefficients 449
13.1. Teachers' Agreement That, "in the Long Run, Students Will Benefit from the Teacher-Evaluation System," Springs 2011–2016 453
13.2. SLs' Agreement That, "in the Long Run, Students Will Benefit from the Teacher-Evaluation System," Springs 2011–2016 454
13.3. SLs' Agreement That, "Overall, I'm Satisfied with the Quality of the Whole Teaching Staff in My School This Year," Springs 2011–2016 457

13.4. SLs' Responses About the Proportion of Teachers in Their School Who Had "the Skills Needed to Foster Meaningful Student Learning," Springs 2013–2016 458
13.5. Estimates of the Initiative's Effect on Average Achievement in HCPS Grades 3 Through 8 461
13.6. Estimates of the Initiative's Effects on Average Reading Achievement in HCPS HSs 463
13.7. Estimates of the Initiative's Effect on Average Achievement in PPS Grades 3 Through 8 465
13.8. Estimates of the Initiative's Effect on Reading Achievement in PPS HSs 467
13.9. Estimates of the Initiative's Effect on Average Achievement in SCS Grades 3 Through 8 469
13.10. Estimates of the Initiative's Effect on Average Achievement in CMO Schools Grades 3 Through 8 474
13.11. Estimates of the Initiative's Effects on Reading Achievement in CMO HSs 477
13.12. Estimates of the Initiative's Effect on Average Achievement in Aspire Grades 3 Through 8 478
13.13. Estimates of the Initiative's Effects on Reading in Green Dot HSs 479

Tables

S.1. Change over Time in Rates of Retention of High- and Low-Performing Teachers xxxviii
S.2. Interrupted Time-Series Estimates of Change in LIM Students' Prereform and Postreform Access to Effective Teaching xl
S.3. Estimated Impact of the IP Initiative on Student Achievement in 2014–2015, by Grade Span and Site xli
2.1. Characteristics of the IP Sites, 2009–2010 and 2014–2015 25
3.1. TE Measures in the Districts Through 2015–2016..... 58
3.2. TE Measures in the CMOs Through 2015–2016..... 65
4.1. Recruitment, Hiring, Transfer, and Placement Policies in the Districts Before and During the IP Initiative..... 120
4.2. Recruitment, Hiring, Transfer, and Placement Policies in the CMOs Before and During the IP Initiative 125
5.1. District Tenure and Dismissal Policies Through 2015–2016 163
5.2. CMO Tenure and Dismissal Policies Through 2015–2016... 168
6.1. District PD Policies Through 2015–2016..... 208
6.2. CMO PD Policies Through 2015–2016..... 213
7.1. District Effectiveness-Based Bonus Policies Through 2015–2016 257
7.2. District Effectiveness-Based Salary Schedule Policies Through 2015–2016 262
7.3. CMO Effectiveness-Based Bonus Policies Through 2015–2016 264
7.4. CMO Effectiveness-Based Salary Schedule Policies Through 2015–2016 265
8.1. District CL Positions Through 2015–2016 299

8.2. CMO CL Positions Through 2015–2016 305
9.1. Costs of Four Evaluation Programs 338
9.2. Proportion of Districts' Total Budgets Spent on the Initiative, 2015–2016 345
9.3. Proportion of CMOs' Total Budgets Spent on the Initiative, 2015–2016 346
11.1. Change over Time in Retention of High-Performing and Low-Performing Teachers 424
12.1. Relationship Between Standard Deviations of Student Achievement and Percentile Growth from the Median Student 430
12.2. Interrupted Time-Series Estimates of Change in LIM Students' Access to Effective Teaching from Before the Initiative to After the Initiative 445
13.1. Visualization of the Initiative's Impacts on Achievement in HCPS, by Year, Grade, and Subject 462
13.2. Visualization of the Initiative's Impacts on Achievement in Pittsburgh Public Schools, by Year, Grade, and Subject ... 466
13.3. Visualization of the Initiative's Impact on Achievement in SCS, by Year, Grade, and Subject 471
13.4. Visualization of the Initiative's Impacts on Achievement in CMO Schools, by Year, Grade, and Subject 476
13.5. Visualization of the Initiative's Impacts on Achievement in Aspire, by Year, Grade, and Subject 480
13.6. Summary of the Initiative's Impacts on Achievement Across Sites, 2014–2015, in Standard Deviations of State Test Scores 482
14.1. The Initiative's Estimated Impact on Student Achievement in 2014–2015, by Grade Span and Site 493

Summary

The Intensive Partnerships for Effective Teaching (IP) initiative, designed and funded by the Bill & Melinda Gates Foundation, was a multiyear effort to dramatically improve student outcomes—particularly high school (HS) graduation and college attendance among low-income minority (LIM) students—by increasing students' access to effective teaching.[1] The theory of action was centered on each participating site's development of a robust measure of teaching effectiveness (TE). This measure was to include, at a minimum, both a teacher's contribution to growth in student achievement and assessment of his or her teaching practices with a structured observation rubric. The measure of effectiveness was to be used to

- improve staffing actions, including
  - revising recruitment and hiring practices to obtain the new teachers likely to be effective
  - adjusting placement and transfer procedures to give students with the greatest needs access to the most-effective teachers
  - reforming tenure and dismissal policies to promote retention of more-effective teachers and facilitate removal of persistently ineffective ones
- identify teaching weaknesses and overcome them through effectiveness-linked professional development (PD)

[1] We define *low-income* as eligible for free or reduced-price meals, and *minority* refers to students classified in districts' administrative data sets as black, Hispanic, or Native American, or some combination of the three.

- employ compensation and career ladders (CLs) as incentives to retain the most-effective teachers and have them support the growth of other teachers.

The developers believed that these mechanisms, referred to as *levers*, would lead, over time, to more-effective teaching, greater access to effective teaching for LIM students, and, as a result, greatly improved academic outcomes.

Beginning in 2009–2010, three school districts and four charter management organizations (CMOs) participated in the IP initiative. The three school districts were Hillsborough County Public Schools (HCPS) in Florida; Memphis City Schools (MCS) in Tennessee (which merged with Shelby County Schools, or SCS, during the initiative); and Pittsburgh Public Schools (PPS) in Pennsylvania. The four CMOs were Alliance College-Ready Public Schools, Aspire Public Schools, Green Dot Public Schools, and Partnerships to Uplift Communities (PUC) Schools. RAND and the American Institutes for Research conducted a six-year evaluation of the initiative, documenting the policies and practices each site enacted and their effects on student outcomes.

We found that the sites succeeded in implementing measures of effectiveness to evaluate teachers and made use of the measures in a range of human-resource (HR) decisions; overall, however, the initiative did not achieve its goals for student achievement or graduation.[2] The findings suggest that, with minor exceptions, by 2014–2015, student achievement, LIM students' access to effective teaching, and dropout rates were not dramatically better than they were for similar sites that did not participate in the IP initiative. There are several possible reasons that the initiative failed to produce the desired dramatic improvement in outcomes across all years: incomplete implementation of the key policies and practices; the influence of external factors, such as state-level policy changes during the IP initiative; insufficient time for effects to appear; a flawed theory of action; or a combination of

[2] We were not able to examine college-going rates.

these factors.[3] Yet, the initiative did produce benefits, and the findings suggest some valuable lessons for districts and policymakers.

Overview of the Initiative

The Gates Foundation awarded grants to the seven sites midway through the 2009–2010 school year, and funding and support continued through the end of 2015–2016. As noted earlier, the school districts were HCPS in Florida, SCS in Tennessee, and PPS in Pennsylvania.[4] The CMOs, all based in California, were Alliance College-Ready Public Schools, Aspire Public Schools, Green Dot Public Schools, and PUC.[5] The grants, which were paid incrementally over the period, totaled more than $200 million and ranged in size from $3.8 million to $81 million (with amounts roughly proportional to enrollment).

The seven sites represented a heterogeneous mix in terms of size, but each had a substantial proportion of students with LIM status. At the start of the initiative, the three districts ranged in size from about 27,000 students to about 185,000 students; at least 55 percent of the students in each district were eligible for free or reduced-price lunch, and at least 51 percent were identified as black, Hispanic, or American Indian. The CMOs ranged in size from about 2,500 students to about 7,500 students; at least 70 percent of the students were eligible for free or reduced-price lunch, and at least 80 percent were from minority population groups.

---

[3] Because the participating sites volunteered to participate rather than being randomly assigned, we had to use quasi-experimental techniques that had limited ability to draw causal inferences about effects.

[4] The initial grant was made to MCS, but MCS merged with SCS in 2013, and the initiative continued under the authority of SCS. For the sake of simplicity, we typically refer to the site as *SCS* in this report.

[5] Initially, five CMOs were affiliated for the purposes of the grant under an umbrella organization called the College-Ready Promise. The fifth CMO, Inner City Education Foundation Public Schools, dropped out of the program after the second year and was not a part of the evaluation.

After a one- to two-year planning phase, most sites began implementing parts of the initiative by 2011–2012 and had implemented all the policy levers to some degree by 2013–2014. Specifically, we found the following:

- Each site adopted an observation rubric that established a common understanding of effective teaching. Sites devoted considerable time and effort to train and certify classroom observers and to observe teachers on a regular basis.
- Each site implemented a composite measure of TE that included scores from direct classroom observations of teaching and a measure of growth in student achievement.
- Each site used the composite measure to varying degrees to make decisions about HR matters, including recruitment and hiring; placement and transfer; tenure and dismissal; PD; and compensation and CLs.

Evaluation Design and Methods

We based our evaluation on interviews with central-office administrators, annual surveys of teachers and school leaders (SLs), case studies of schools, analyses of site records and student test scores, and reviews of relevant documentation.[6]

We used a variety of methods to analyze the data to assess teachers' impact on student achievement over the course of the initiative. In particular, we estimated changes in the levels and distribution of TE in two ways: one using our study-calculated measure of value added and the other using the site-developed measure of effectiveness.

[6] This report extends the results of three interim studies published in 2016 and serves as the final report of the evaluation of the initiative.

Site HR Policies

Each site implemented changes in teacher management policies aligned with the IP theory of action. Each began by adopting a measure of TE, which served as the basis for changes in other HR policies. As described in the following sections, these policy areas (or *levers*) included recruitment and hiring, placement and transfer, tenure and dismissal, PD, and compensation and CLs.

Measures of Effective Teaching

What measures did sites adopt to assess TE, and what did educators think of their sites' teacher-evaluation systems?

Each of the IP sites designed a teacher-evaluation system consisting of multiple measures. Every site included at least two factors in its composite score: (1) rubric-based ratings of teaching based on classroom observations and (2) a measure of student achievement growth calculated using either value-added modeling (VAM) or student growth percentiles. Some sites included other factors, such as input from students or parents. In all of the systems, measures based on observations of teachers' practices and on student achievement growth had the largest weights in each teacher's overall rating, although sites took different approaches to measuring and weighting these and other components.

Most teachers were rated as effective or higher. Over the course of the initiative, an increasing proportion of teachers received ratings that placed them in the higher categories, and fewer teachers received ratings that placed them in the lower categories. (After 2013, each site's lowest category contained 2 percent or less of its teachers.) This shift in the rating distribution might reflect actual improvement in teaching, but there is some evidence that it is due to other factors, such as more-generous ratings on subjective components (e.g., classroom observations).

The evaluation system raised some practical challenges that sites addressed in different ways. One challenge was that the observations placed considerable additional burden on principals' time; over time, some sites reduced the observations' length or frequency or allowed other school administrators to conduct observations. Another chal-

lenge that most of the sites faced was that many teachers did not receive individual value-added scores because there were no standardized tests in their subjects or grade levels. Some sites handled this issue by assigning school-level average scores to those teachers; others adopted alternative approaches to measuring student growth.

Despite some concerns about fairness, the surveys found that majorities of teachers were aware of the components of the evaluation measures and thought that those components, particularly the classroom-observation component, were valid measures of their effectiveness as teachers. Most reported that the evaluation system had helped improve their teaching. Although many teachers did not agree that their site's evaluation system was "fair to all teachers, regardless of their personal characteristics or those of the students they teach," most reported that the evaluation system had been fair to *them*. This might reflect the fact that large majorities of teachers received high TE ratings. Indeed, teachers who received high ratings were more likely than low-rated teachers to agree that the evaluation system had been fair to *them*.

Recruitment, Hiring, Placement, and Transfer Policies

How did sites modify their recruitment, hiring, placement, and transfer policies to achieve the initiative's goals?

Our analysis found that all sites modified their recruitment and hiring policies somewhat during the IP initiative—for example, by increasing recruitment efforts, automating the application and screening process, or facilitating hiring for hard-to-staff schools. In addition, some sites developed residency programs and partnerships with local colleges and universities to increase and diversify the pool of teacher candidates. Even so, the sites had difficulty attracting effective teachers to high-need schools, and persistent teacher turnover was a particular problem for the CMOs. Also, changes to placement and transfer policies were relatively uncommon.

SLs generally thought that the processes by which teachers were hired to their school worked well, and, in PPS and SCS, perceptions of the hiring processes improved over time. SLs perceived residency programs as having beneficial effects, and, in sites that offered incentives

for effective teachers to transfer to high-need schools, many, but not all, SLs at high-need schools perceived benefits. In addition, over time, fewer and fewer SLs reported experiencing the loss of good teachers from, and the assignment of unwanted teachers to, their schools.

Tenure and Dismissal Policies

Did the sites reform their tenure and dismissal policies in an effort to reduce the number of low-performing teachers? How did teachers react to their sites' tenure and dismissal policies?

The three districts set specific criteria based on their new evaluation systems to identify low-performing teachers who might be denied tenure, placed on an improvement plan, or considered for dismissal or nonrenewal of their contracts. In contrast, the CMOs (which have never offered tenure) did not establish specific criteria that triggered action for teachers with performance problems, but the CMOs did take teacher-evaluation results into account when considering improvement plans or contract renewal.

Across the sites for which data were available, about 1 percent of teachers were dismissed for poor performance in 2015–2016.[7] At least in part, sites dismissed few teachers because their evaluation systems identified very few poor performers. The sites struggled to balance reforms aimed at dismissing low-performing teachers with those aimed at improving teacher performance. In general, the sites tended to favor trying to help teachers improve rather than dismissing them. Using evaluation results as the basis for tenure and dismissal decisions might have led some principals to avoid giving low observation ratings that would reduce a teacher's composite effectiveness score.

Survey data suggest that teachers understood the sites' tenure policies but that, over time, fewer teachers thought that tenure should be linked to evaluation results. On the other hand, even by the end of the reform, only about half the teachers reported that they understood the criteria for dismissal. As the initiative progressed, SLs in the districts were less likely to report difficulties in dismissing low-performing

[7] We do not have data on dismissal rates prior to the initiative, so we cannot determine whether this represents a change.

teachers, but the opposite was true in the CMOs. In general, SLs saw the dismissal policies as fair. Relatively few teachers reported that they were worried about dismissal, but teachers with low effectiveness ratings were more likely than higher-rated teachers to report worry.

PD Policies

What kinds of PD activities did the sites plan and deliver, especially to meet needs identified through the teacher-evaluation system? Did the sites monitor teachers' participation in PD? Did teachers view the PD as useful in improving their instruction and student learning?

Every site offered multiple types of PD, including coaching, workshops, school-based teacher collaboration, and online and video resources. Principals and other staff recommended PD based on teachers' evaluation results, but sites generally did not require teachers to participate, monitor their participation, or examine whether participants' effectiveness improved as a result of PD.

The sites confronted some challenges in moving toward effectiveness-linked PD systems. They struggled to figure out how to individualize PD to address performance problems identified during a teacher's evaluation. Some also found it difficult to develop a coherent system of PD offerings. Although every site implemented an online repository of PD schedules and materials, the repositories were not entirely successful, at least in part because teachers had technical difficulties accessing the online materials.

Teachers in all the sites indicated awareness of the available PD activities and generally believed that the PD in which they had participated was useful for improving student learning. Teachers' participation in PD was influenced by their evaluations to some degree but also by their own interests, by standards, and by the curriculum. Most teachers viewed school-based collaboration as useful, while fewer teachers said the same of workshops and in-services. Most teachers had access to some form of coaching, on which the sites often relied to individualize PD, and the percentage increased over time. Teachers with lower effectiveness ratings were more likely than those with higher ratings to report having received *formal, individualized* coaching or mentoring. However, teachers with lower effectiveness ratings were gener-

ally no more likely than higher-rated teachers to say that the support they had received had helped them.

Compensation Policies

Did the sites modify their compensation policies to reward effective teaching, and did these policies lead to a more equitable distribution of effective teachers?

All seven IP sites implemented effectiveness-based compensation reforms, although they varied in terms of timing, eligibility criteria, dollar amounts, and the proportion of teachers earning additional compensation—factors that would be likely to affect the impact of the policies.

Changes in compensation policies evoked some concerns. Teachers, for example, wanted assurances that their base salaries would not be reduced under the new compensation system. In addition, some sites faced uncertainty about their overall levels of state funding, which made it difficult to change policies that affected salaries. And where teacher unions or professional organizations existed, sites had to work carefully with them to ensure that their concerns were addressed.

Survey responses suggest that teachers were generally aware of their sites' compensation policies and endorsed the idea of additional compensation for teachers demonstrating outstanding teaching. In most years, a majority of teachers in PPS and the CMOs, but only a minority in HCPS and SCS, thought that site compensation policies were reasonable, fair, and appropriate. However, except in Alliance and Aspire, most teachers did not think that their sites' policies would motivate them to improve their teaching. In addition, some teachers objected to compensation decisions being made on the basis of effectiveness measures they perceived to be flawed.

We were able to analyze the relationship between teacher salaries and effectiveness (controlling for age, experience, education, gender, and race) in two sites: PPS and HCPS. In PPS, teachers with higher effectiveness scores received greater total compensation than teachers with lower effectiveness scores during the two years for which we have data, 2012–2013 and 2013–2014; we could not track changes in the size of that gap over time. In HCPS, we had salary and effectiveness

data from 2010–2011 through 2014–2015, and we found no noticeable change in the salary premium attached to effectiveness during this period. Although more-effective teachers in HCPS tended to receive greater total salaries, the gap in salaries between more and less effective teachers did not increase after the implementation of the effectiveness-based policies, suggesting that the new policies had little effect on how much money effective teachers earned relative to less effective teachers.

CL Policies

To what extent did sites implement CL policies offering teachers additional responsibilities and leadership opportunities, and how did staff react to these policies?

Each of the seven sites introduced specialized roles, with additional pay, open to effective teachers who were willing to accept additional responsibility to provide instructional or curricular support. None, however, implemented a CL with the sequential steps and growing responsibility that many advocates recommend and that the theory of action envisioned. The districts and CMOs took somewhat different approaches; the districts tended to create a few positions that focused on coaching and mentoring new teachers in struggling schools, while the CMOs tended to create a larger number of positions with a wider range of duties as needs shifted over time.

Most teachers knew whether CL options were available in any given year, and, for the most part, teachers viewed CL policies and teachers in CL roles favorably. For example, most teachers thought that the CL teachers with whom they worked were effective educators who had helped them improve their instruction and deserved the additional compensation they received. Majorities of teachers in most sites also reported that the opportunity to become a CL teacher motivated them to improve their instruction and increased the chances they would remain in teaching. PPS was an exception. For a variety of reasons—including problems that occurred during the initial implementation of the CL policies—teachers in PPS were less likely than teachers in the other sites to express positive opinions about CL policies and roles.

Resources Invested in the Initiative

How much did the sites spend, in terms of both fiscal and human resources, to implement the IP initiative?

From 2009 through 2016, total IP spending (i.e., expenditures that could be directly associated with the components of the IP initiative) across the seven sites was $575 million, with per-pupil expenditures ranging from $868 in Green Dot to $3,541 in PPS. As a percentage of total site budget, the expenditures on the initiative ranged from 1.7 percent in Green Dot to 6.6 percent in Aspire. Activities related to compensation and CLs received the largest share of the total funding allocated to the IP initiative in five of the seven sites. The exceptions, SCS and PUC, spent the largest shares of their total IP funding—42 and 37 percent, respectively—on PD.

The largest source of funding for the initiative was the Gates Foundation, which awarded approximately $212 million across the seven sites. Across the sites, the percentage of overall expenditures on the initiative that the foundation funded ranged from 28 percent in Aspire to 53 percent in Green Dot. For HCPS, PPS, Green Dot, and PUC, the largest source of non–Gates Foundation funds was site funds. For SCS, Alliance, and Aspire, federal funding provided the largest source of other funds.

An additional cost of the IP initiative, beyond direct expenditures, was the time teachers and SLs spent on activities related to the teacher-evaluation component. By assigning a dollar value to their time, we estimated that IP costs for teacher-evaluation activities totaled nearly $100 million across the seven sites in 2014–2015: The value of teacher and SL time devoted to evaluation was about $73 million, and the direct expenditures on evaluation made up approximately an additional $26 million. In per-pupil terms, the overall cost for evaluation activities averaged almost $280 per pupil: $201 for the value of time spent on evaluation activities and $78 for fiscal expenditures.

Intermediate and Long-Term Outcomes

The IP initiative was designed to improve TE through specific mechanisms with the goal of improving student achievement and graduation rates. The next sections highlight our findings with respect to the effectiveness of newly hired teachers, the retention of effective teachers, the effectiveness of the teacher workforce overall, LIM students' access to effective teaching, and, finally, the impact of the initiative, as a whole, on student outcomes.

Effectiveness of Newly Hired Teachers

Did the effectiveness of newly hired teachers improve over time?

Each of the sites modified its recruitment and hiring practices with the goal of attracting new teachers who would be more effective than other candidates. Our analysis found little evidence that new policies related to recruitment, hiring, and new-teacher support led to sites hiring more-effective teachers. Although the site TE scores of newly hired teachers increased over time in some sites, these changes appear to be a result of inflation in the TE measure rather than improvements in the selection of candidates. We drew this conclusion because we did not observe changes in effectiveness as measured by study-calculated VAM scores, and we observed similar improvements in the site TE scores of more-experienced teachers.

Most SLs were satisfied with the quality of new hires at the beginning of the initiative, and this level of satisfaction was largely unchanged over the next six years.

Retention of Effective Teachers

Did new staffing policies related to recruitment and hiring, dismissal and tenure, and compensation and CLs collectively increase the likelihood that sites would retain effective teachers?

One of the intermediate goals of the IP initiative was to improve the rate of retention of effective teachers, and some of the levers described previously were intended to have this effect. Because the various levers were implemented at roughly the same time and applied to all or almost all teachers, we could not disentangle the effects of indi-

vidual levers. However, we could examine the extent to which more-effective teachers were retained at higher rates as the initiative progressed and the levers were more fully implemented. Our analysis drew on two sources of evidence related to teacher retention: survey results and empirical analyses of employment records.

On surveys administered from 2011 through 2015, less than half the teachers in most of the sites expected to be still working as a teacher in their current site in five years' time—a substantial challenge for schools if teachers followed through with their plans. In HCPS and PPS, more-effective teachers were more likely than less effective teachers to say they would continue teaching, but, in Aspire, the opposite was true.

In all the IP sites, empirical analyses of employment records from recent years show that teachers with higher effectiveness ratings were generally more likely than lower-rated teachers to continue teaching. However, we found little evidence that the policies designed, in whole or in part, to improve the level of retention of effective teachers had the intended effect. The rate of retention of effective teachers did not increase over time as relevant policies were implemented (see the leftmost TE column of Table S.1). A similar analysis based only on measures of value added rather than on the site-calculated effectiveness composite reached the same conclusion (see the leftmost VAM column of Table S.1).

On the other hand, the implementation of the levers might have been responsible for encouraging less effective teachers to leave the field: In most sites, the exit rate for less effective teachers increased over the course of the initiative, as intended by the new HR policies (see rightmost TE column and rightmost VAM column of Table S.1).

TE and LIM Students' Access to Effective Teaching

Did the overall IP initiative improve the effectiveness of the teacher workforce, and did LIM students have greater access to effective teaching?

The key mechanism through which the IP initiative hoped to improve student outcomes was to increase the effectiveness of teaching on a large scale and to increase LIM students' access to effective teaching. For the most part, neither of these changes occurred, although

Table S.1
Change over Time in Rates of Retention of High- and Low-Performing Teachers

	Change in Retention Rates of High-Performing Teachers		Change in Retention Rates of Low-Performing Teachers	
Site	**TE (Change from the Early to Late IP Period)**	**VAM (Change from the Pre- to Late IP Period)**	**TE (Change from the Early to Late IP Period)**	**VAM (Change from the Pre to Late IP Period)**
HCPS	N/S	N/S	–	–
PPS	+	N/S	–	N/S
SCS	–	–	N/S	–
Alliance	N/S	N/A	N/S	N/A
Aspire	N/S	–	N/S	–
Green Dot	N/S	N/A	N/S	N/A

NOTE: + denotes positive and statistically significant at $p < 0.05$ (green cells). – denotes negative and statistically significant at $p < 0.05$ (red cells). N/S = not significant. N/A = not applicable because of data limitations (gray cells). We could not do this analysis for PUC.

the results of surveys and the results of empirical analyses are not in agreement.

Most SLs in all sites reported on surveys that they had taken steps to assign the most-effective teachers to students with the greatest needs. In addition, most SLs in HCPS and SCS reported using teacher evaluations to some extent to assign teachers to students or classes.

On the other hand, an analysis of the distribution of TE based on our measures of value added found that TE did not consistently improve in mathematics or reading in the three IP districts. There was very small improvement in effectiveness among mathematics teachers in HCPS and SCS and larger improvement among reading teachers in SCS, but there were also significant declines in effectiveness among reading teachers in HCPS and PPS.[8]

In addition, in HCPS, LIM students' overall access to effective teaching and LIM students' school-level access to effective teaching declined in reading and mathematics during the period of the initiative (see Table S.2). In the other districts, LIM students did not have consistently greater access to effective teaching before, during, or after the IP initiative.

Finally, we examined three mechanisms by which sites could increase LIM students' access to effective teaching: (1) improving the effectiveness of teachers in high-LIM classrooms, (2) moving effective teachers from low-LIM to high-LIM classrooms, and (3) increasing the number of LIM students in the classrooms of effective teachers.[9] The analysis found no evidence that any of these changes consistently occurred in the sites.

Student Outcomes

As a whole, did the IP initiative improve student outcomes?

The IP initiative was designed to test whether the effectiveness-based teaching policies described above would lead to dramatic

---

[8] We do not report these results for the CMOs because they had fewer teachers and were missing outcome data for a key year when California discontinued its testing program.

[9] We use *high LIM* and *low LIM* as shorthand for *having a high percentage of students who are LIM* and *having a low percentage of students who are LIM*, respectively.

Table S.2
Interrupted Time-Series Estimates of Change in LIM Students' Prereform and Postreform Access to Effective Teaching

District	Subject	Overall	Between Schools	Within Schools
HCPS	Mathematics	–0.15***	–0.16***	–0.05
	Reading	–0.07***	–0.08***	–0.01
PPS	Mathematics	–0.05	–0.06	0.12
	Reading	0.02	0.03	–0.04
SCS	Mathematics	–0.03	–0.05	0.05
	Reading	–0.05	–0.04	–0.08

NOTE: Prereform period is 2009–2010 and earlier; postreform is 2013–2014 and later. We weighted regressions using VAM standard errors. We clustered standard errors at the school level. We adjusted determination of statistical significance for multiple hypothesis tests using the Benjamini–Hochberg correction. Red indicates a significant negative estimate. Gray indicates a nonsignificant estimate. *** denotes statistical significance at $p < 0.01$.

improvement in student achievement and graduation, particularly for LIM students. We did not find this to be the case when we compared student achievement and graduation in IP schools with achievement and graduation in similar, non–IP schools within each IP site's state.

Teachers were not optimistic that the IP reforms would lead to gains in student outcomes. In the three districts, typically less than 50 percent of teachers agreed that, in the long run, students would benefit from the teacher-evaluation system, and that percentage declined over the course of the reform. In the CMOs, agreement about the benefit of the evaluation system was notably higher. SLs were more likely than teachers to think that the reform would benefit students in the long run, but they also became less likely to believe this over time.

Our analyses of student test results and graduation rates showed no evidence of widespread positive impact on student outcomes six years after the IP initiative was first funded in 2009–2010. As in previous years, there were few significant impacts across grades and subjects in the IP sites.

Table S.3 shows the effects on student achievement, by grade span and site. In most sites and grade-level ranges, the initiative did not have a significant impact on student achievement in mathematics or reading (the gray cells), although there was a significant positive effect in HS English language arts in the CMOs and PPS and a significant negative effect in mathematics in grades 3 through 8 in the CMOs.

In HCPS, the IP initiative had no effect on academic achievement in reading or mathematics in grades 3 through 8 in any year. It was associated with negative effects on HS achievement in reading initially, but, for 2014–2015, we found no effects on reading achievement or dropout rates. In PPS in 2014–2015, the initiative had no effects on average student achievement in reading or mathematics in the lower grade levels. However, it had a positive effect on HS (grade 11) reading achievement in three of the four most-recent years. In SCS, after negative effects in the lower grade levels during the first years of the initiative, there were no statistically significant effects on mathematics or reading achievement in lower grades in 2014–2015. In the CMOs, the initiative had negative effects on average achievement in lower-grade mathematics and no effect on lower-grade reading, but it had positive effects on HS reading achievement in 2014–2015.

Table S.3
Estimated Impact of the IP Initiative on Student Achievement in 2014–2015, by Grade Span and Site

	Grades 3–8		HS	
Site	**Mathematics**	**Reading**	**Mathematics**	**Reading**
HCPS	–0.05	0.00	N/A	0.00
PPS	0.01	–0.01	N/A	0.09
SCS	–0.14	–0.02	N/A	N/A
CMOs	–0.17	–0.08	N/A	0.19

NOTE: We report values in standard deviations of state test scores. We could not estimate the impact on HS mathematics because students did not take the same secondary mathematics tests. Red indicates statistically significant ($p < 0.05$) negative effects, gray indicates non–statistically significant effects, and green indicates statistically significant positive effects.

We should note two caveats about these results. First, other districts in each state were enacting reforms similar to those of the IP initiative. In particular, new teacher-evaluation measures with consequences were enacted in three of the states at the same time the IP initiative was taking place. Our impact estimates reveal the extent to which the IP initiative improved student outcomes *over and above* these statewide efforts. Second, the reforms might require more time to take effect, and we will continue to monitor student outcomes for two more years to see whether there are increased effects.

Sustained Policies

Although we observed few effects on student outcomes, the sites found many aspects of the initiative to be positive. Site administrators, SLs, and teachers we interviewed described several positive consequences. For example, most teachers and SLs we interviewed told us that the adoption of the observation rubrics helped to develop a shared understanding of the elements of effective teaching that fostered communication about instructional improvement. Furthermore, majorities of teachers we surveyed reported using the information from the TE measures to change their instruction. For the most part, the sites are also maintaining the information systems they developed to support the measures of effective teaching. In fact, as of 2016–2017, the sites had maintained multiple elements of the initiative, which we consider an indication that they perceived these reforms to be beneficial. In some cases, the impetus for continuing came from the site, but, in other cases, state law or regulation now requires the practice; disentangling these two factors is difficult. Nevertheless, the key elements that have become established policy in the sites include the following:

- incorporating systematic teacher evaluation into their regular practice. The sites continue to use research-based rubrics for classroom observations to gather information about teaching practice and to structure conversations about improvement. All sites continue to use multiple evaluation measures. All but one continue

to compute a composite TE measure that includes (at a minimum) observation scores and student achievement growth, and they continue to use this measure to identify teachers who need to improve.
- modifying their recruitment and hiring policies to improve the potential effectiveness of new teachers and offer them more initial support.

In addition, many sites continue to offer the CL positions they created, particularly those related to mentoring and coaching. Some sites are retaining central-office staff positions and departments that were created to support the initiative.

Other elements of the initiative are not being continued because they were not perceived to provide sufficient benefits or were too costly. For example, HCPS no longer uses peer evaluators. Most sites have reduced the number, frequency, or length of classroom observations to reduce the burden on administrators and, in some cases, to expedite the rapid sharing of information with teachers. Most sites continue various types of performance-based bonuses, but they are not large and are usually available only to teachers in high-need schools. None of the sites created true CLs with positions arranged in a defined hierarchy of increasing responsibility and salary, and none appears interested in doing so.

Discussion

The IP initiative heightened the sites' attention to TE; however, measuring effectiveness and using it as the basis for teacher management and incentives did not appear to lead to gains in student achievement or graduation rates. The evaluation does not tell us why these outcomes were not achieved, but we are willing to speculate—informed by our observations of the sites and the foundation during the past seven years—about potential factors that might explain the lack of impact. In some instances, we offer recommendations to address the identified issues. These recommendations might be of value to the founda-

tion, the sites, or others that might be contemplating similar large-scale reform efforts in the future.

Incomplete Implementation and Lack of Successful Models

The levers might not have been implemented with sufficient quality, intensity, or duration to achieve their potential full effect. None of the main policy levers—staffing, PD, or compensation and CLs—was implemented fully or as initially envisioned. Implementing the levers required the sites to develop specific policies and practices to translate general ideas from the theory of action into operational realities. For many of the levers, the sites had to do this in the absence of successful models. This practical gap was most apparent in the sites' efforts to implement evaluation-linked PD; they did not find successful models (e.g., from other districts) they could observe, adopt, or adapt. Instead, they had to develop their own systems, and they encountered practical problems in doing so—problems that had still not been fully resolved by the end of the grant period.

The sites' experiences with evaluation-linked PD (and other levers) suggest to us the value of distinguishing between implementation (putting designed practices into place) and innovation (developing new practices). At the beginning of the evaluation, many members of the research team thought that the main challenge of the IP initiative would be effective implementation (i.e., carefully and systematically enacting well-developed ideas). Effective implementation might include such things as explaining new practices to stakeholders, adapting them to the local context, or reallocating resources to support new systems or positions. However, we found that, for many of the levers, the main challenge in enacting the IP theory of action was innovation. The sites had to create new methods or approaches in the absence of functioning models. Our experience suggests that the capacities needed to implement might differ from the capacities needed to innovate. In addition, implementation requires less time than innovation. For example, had there been time, the sites might have allocated staff or hired consultants to analyze existing data to help them understand how well PD efforts were working, revised their PD infrastructure to collect better information about participation and perceived quality, or

tried different ways of linking PD to effectiveness to see which worked best. Alternatively, the foundation might have worked with other sites to develop and test prototypes before incorporating them into a larger reform. This is essentially what the foundation did with the Measures of Effective Teaching project; it conducted a research-and-development effort to determine the best combination of measures to use for measuring TE. The IP sites would have benefited if similar preparatory work had been done on systems to link effectiveness measures to HR policies.[10]

Problems in Making Use of Teacher-Evaluation Measures

Teacher evaluation was at the core of the initiative, and the sites were committed to using the measures to inform key HR decisions. But as we describe in Chapters Three through Eight, the sites encountered two problems related to these intended uses of the TE measures. First, it was difficult for the sites to navigate the underlying tension between using evaluation information for professional improvement and using it for high-stakes decisions. Second, some sites encountered unexpected resistance when they tried to use effectiveness scores for high-stakes personnel decisions; this occurred despite the fact that the main stakeholder groups had endorsed the initiative at the outset.

Tension Between Different Purposes for Measures

Researchers distinguish between measures used for summative purposes (to make overall judgments about quality) and measures used for formative purposes (to improve conditions or practices). For a variety of reasons, designing a measure that is equally good for both purposes is hard. Perhaps most importantly, the characteristics of good summative measures differ from those of good formative measures. For example,

[10] Developing and validating methods to increase the value of observation feedback for teacher improvement could be a particularly generative focus for research and development. In particular, much could be learned by helping districts gather systematic data on the recommendations for PD that flow from teacher evaluations, the PD in which teachers actually participate, and changes in teachers' evaluation ratings associated with participation. Sites might also conduct small-scale experiments randomly assigning teachers to receive different forms of PD based on evaluation results (e.g., more or less intensive coaching).

because summative measures are used to make consequential decisions about individuals or programs, they need to meet high standards of technical quality (i.e., reliability and validity). Formative measures, which are used to guide improvement, do not need to be as technically sound. Because of the need for quality, summative measures are usually longer, and they are administered only occasionally. Formative measures can be shorter, more frequent, embedded in curriculum, and so forth.

The IP sites had hoped that their TE measures could be used both for improvement and for accountability purposes. But these two goals often conflicted. For example, because sites planned to use effectiveness in tenure and dismissal decisions, effectiveness measures had to meet high standards for reliability and validity. Thus, all classroom observers had to be trained and had to pass certification tests to ensure that they scored lessons accurately. Furthermore, each teacher had to be observed multiple times during the year for full lessons. On the other hand, sites found that shorter, more-frequent observations could be more useful for improvement purposes because they allowed for more-immediate feedback to teachers. Many sites changed the structure of the observations to reduce the time burden and to better support improvement purposes. Sites also found greater support for effectiveness-based improvement policies that involved low or no stakes (e.g., coaching or mentoring informed by measured effectiveness) and greater resistance to policies that threatened compensation or employment.

Resistance to Using Teacher Evaluation for High-Stakes Decisions

During the initiative, the sites adapted their evaluation systems, either formally or informally, to avoid having to dismiss many teachers. The IP initiative was launched with great fanfare and with the endorsement of each site's board, administration, teachers, and local community. During the first two years, most stakeholders had positive attitudes toward it. A change occurred when new policies threatened some teachers with loss of employment or reduction in salary. Teacher organizations, notably in PPS, began to object and mount public campaigns against the effectiveness measures when high stakes were due to be attached and larger numbers of teachers were threatened. Over time,

in most of the sites, fewer and fewer teachers were identified as lower performing, possibly because local norms and expectations influenced how ratings were assigned or where performance levels were set. The lesson we draw from this experience is that reformers should not underestimate the resistance that could arise if changes have major negative consequences for staff employment.

Changes in the Local or State Context

Changes in the local or state context might also have interfered with the sites' ability to implement the policy levers fully. In Chapter Two, we provide an overview of the context in each site, and, in Chapters Three through Eight, we describe external conditions that hampered implementation of the initiative. The most-unexpected and most-problematic shifts came from political decisions relating to governance and testing and from changes in local leadership. With respect to testing, each of the four states changed its statewide test during the course of the initiative, which necessitated adjustments to the sites' achievement and growth measures and caused some concerns about comparability of scores across years. Other significant but unexpected changes included the merger of legacy MCS with legacy SCS; the Pennsylvania budget crisis; the Hillsborough County school board's abrupt decision to remove the superintendent; and the California education budget cuts that significantly reduced funding to the CMOs.

In addition, every site except Green Dot and PUC had a turnover in top leadership during the IP initiative. In some places, the new superintendent or director maintained the focus on TE, but, in other sites, the new leaders eliminated parts of the IP reform, slowed implementation, or established other priorities.

Insufficient Attention to Other Factors

Finally, we should mention that the IP initiative might have fallen short of its dramatic goals because improvement on that scale requires attention to a broader set of factors. The initiative was appealing, in part, because of its tight focus on TE. Research that informed the design of the initiative suggests that attention to TE is necessary to the improvement of student outcomes. Although teachers remain the most salient

in-school factor in determining student outcomes, and thus improving teaching is a plausible lever for improvement, differences among teachers explain only a small percentage of the variation in student achievement.[11] Perhaps a near-exclusive focus on TE is insufficient to dramatically improve student outcomes. Many other factors might need to be addressed, ranging from early childhood education, to students' social and emotional competencies, to the school learning environment, to family support.

The Importance of Measuring Implementation

Apart from suggesting lessons about the reform of teacher human capital systems, our work has a few implications for the evaluation of large-scale initiatives. The IP initiative was a multifaceted reform that addressed a range of interrelated policies and procedures in each site. In such change efforts as this, it is important to measure the extent to which each of the new policies and procedures is implemented in order to understand how the specific elements of the reform relate to outcomes. We think that it is particularly important to measure *dosage* (i.e., how much particular activities occur, over what period of time, and with what intensity). In the present study, we tried to document the extent to which each site enacted various levers. We also obtained the evaluation ratings that teachers received, and we were able to link teachers' ratings to their attitudes, compensation, and retention. However, we were not able to measure other aspects of dosage (e.g., the extent of feedback teachers received, such as duration of meetings with observers; the number and types of PD activities in which teachers participated, or teachers' placement on improvement status). Even with the data we had, we could not disentangle the effects of levers that were enacted simultaneously on the same set of teachers (e.g., the impact of effectiveness-based compensation separate from CLs' impact on teachers' retention decisions). The initiative does not seem to have placed sufficient emphasis from the start on developing the data systems that

[11] Estimates of the percentage of the variation in student achievement growth that is attributable to variation in teachers range from 1 to 14 percent (American Statistical Association, 2014).

would have been required to track implementation systematically. In particular, most of the IP sites did not generally collect information about access to, participation in, or quality of PD, and we could not gather that information on our own. Knowing more about PD dosage would have increased our ability, as well as the sites' ability, to judge the effectiveness of the sites' emerging effectiveness-linked PD efforts. We think that this is a lesson worth considering in future reforms.

Final Thought

A favorite saying in the educational measurement community is that one does not fatten a hog by weighing it. The IP initiative might have failed to achieve its goals because the sites were better at implementing measures of effectiveness than at using them to improve student outcomes. Contrary to the developers' expectations, and for a variety of reasons described in the report, the sites were not able to use the information to improve the effectiveness of their existing teachers through individualized PD, CLs, or coaching and mentoring.

Acknowledgments

We are grateful to a substantial number of school district and charter management organization (CMO) staff, including superintendents and chief executive officers, central-office personnel, school leaders, and teachers, who gave generously of their time to share their insights and experiences with the Intensive Partnership (IP) initiative. Space does not permit us to name them all, but we would like to acknowledge the contributions of people with whom we worked repeatedly during the course of the study.

The top leadership at each site had the vision to participate in the IP initiative and cooperated with us throughout the evaluation. We appreciate the support of the following superintendents: MaryEllen Elia and Jeff Eakins of Hillsborough County Public Schools (HCPS); Mark Roosevelt, Linda Lane, and Anthony Hamlet of Pittsburgh Public Schools (PPS); and Kriner Cash, John Aitken, and Dorsey E. Hopson II of Memphis City Schools and Shelby County Schools (SCS); and of these CMO chief executive officers (CEOs): Judy Ivie Burton and Dan Katzir of Alliance College-Ready Public Schools; James Wilcox and Carolyn Hack of Aspire Public Schools; Cristina de Jesus, CEO and president of Green Dot Public Schools California, and Marco Petruzzi, CEO and president of Green Dot Public Schools National; and Jacqueline Elliot, CEO and president of Partnerships to Uplift Communities National.

In addition, we appreciate the efforts of central-office staff members who provided information about each site's implementation efforts, helped us to coordinate our communication with staff, facilitated our

access to site data, and reviewed our tables and charts for accuracy. These staff members include Anna Brown, Dena Collins, Jessica Doherty, Ted Dwyer, Jamalya Jackson, Justin R. Rogers, David Russell, and Marie Whelan of HCPS; Sam Franklin, Lorraine Marnet, Tara Tucci, and Ashley Varrato of PPS; Tequilla Banks, John Barker, Tangie Ray Davis, Anasa Franklin, Leah Halteman, Rorie Harris, Monica Jordan, Jessica Lotz, Jeffery Shive, and Kristin Walker of SCS; Laura Alvarez, Vireak Chheng, Harris Luu, David Medina, Liam Mennis, and Howard Morales of Alliance; Veronica Chew, Benjamin Crosby, Elise Darwish, James Gallagher, Cathleen Li, Alexander Orsini, and David Roth of Aspire; Julia Fisher, Annette Gonzalez, Kelly Hurley, Kevin Keelen, Samantha Mita, Quang Nguyen, and Sam Regnier of Green Dot; Adriana Abich, Allegra Brown, Leslie Chang, Matt Goodlaw, George Lewis, Johana Olaya, Nik Orlando, Malena Orozco-Otero, Jaime Serrano, and Jonathan Stewart of Partnerships to Uplift Communities Schools; and Erin Abshere of the College-Ready Promise.[1]

Teacher organizations helped us understand the teachers' concerns and explain the evaluation to teachers; we want to acknowledge the assistance of Jean Clements, president of the Hillsborough Classroom Teachers Association Executive Board; Stephanie Baxter-Jenkins, executive director of the Hillsborough Classroom Teachers Association; John Tarka, former president of the Pittsburgh Federation of Teachers American Federation of Teachers Local 400 (PFT); Nina Esposito-Visgitis, president of PFT; William Hileman, vice president and staff representative of PFT; Ken Foster, former executive director of the Memphis–Shelby County Education Association; and Salina Joiner, president of Asociacion de Maestros Unidos.

We received assistance in obtaining and analyzing state testing data from John Nau and Milad Elhadri at the Pennsylvania Department of Education, Jonathon M. Attridge from the Tennessee Depart-

[1] The College-Ready Promise was the organization formed originally to administer the foundation grant to the CMOs; midway through the initiative, the foundation began making grants to the individual CMOs, and the College-Ready Promise continued to perform selected functions.

ment of Education, and Sonja L. Bridges from the Florida Department of Education.

We also appreciate the willingness of staff at the Bill & Melinda Gates Foundation to engage with us; we are especially grateful to our program officers, David Silver, Eli Pristoop, and Andrew Sokatch for their advice and responsiveness throughout the project. In addition, we received helpful feedback from other foundation staff, including Sara Allan, Nate Brown, Sarah Buhayar, Steve Cantrell, Josh Edelman, Irvin Scott, Greg Sommers, Ky Vu, and Noah Wepman.

We also want to acknowledge the contributions of Allan Odden of the University of Wisconsin, who was a consultant to the evaluation team throughout the study; we benefited from his extensive knowledge of the research on strategic human capital management in education and prior efforts to implement similar performance-based teacher management systems.

We thank Catherine H. Augustine, Cathleen Stasz, and Kata Mihaly of RAND; Andrew Wayne of the American Institutes for Research; and Katharine O. Strunk of Michigan State University for reviewing the document and providing constructive feedback. Finally, we acknowledge other members of the RAND team, including Jan M. Hanley, Courtney Ann Kase, Beth Katz, Emma Lee, Stephanie Lonsinger, Paco Martorell, Jennifer Sloan McCombs, Nate Orr, Andrea Phillips, Mollie Rudnick, Ethan Scherer, Jennifer L. Steele, and Stephanie Williamson; and of the American Institutes for Research team, including Melissa Arellanes, Charles Blankenship, Linda Choi, Jennifer Ford, Gur Hoshen, Kathleen Jones, Ben Kalina, Jesse D. Levin, Jeanette Moses, Jeffrey Poirier, Bradley Salvato, and Daniel R. Sherman.

Abbreviations

AIR	American Institutes for Research
ALLI	Alliance Learning and Leadership Initiative
AP	assistant principal
ASD	Achievement School District
AYP	adequate yearly progress
BTSA	Beginning Teacher Support and Assessment Program
CAHSEE	California High School Exit Examination
CAO	chief academic officer
CBA	curriculum-based assessment
CEO	chief executive officer
CL	career ladder
CMO	charter management organization
CRI	clinical resident instructor
CST	California Standards Test
CSUDH	California State University, Dominguez Hills
CT	consulting teacher
D	distinguished

DiD	difference in differences
DTTE	Department of Teacher Talent and Effectiveness
E	effective
EET	Empowering Effective Teachers
ELA	English language arts
ELL	English language learner
EOC	end of course
F	failing
FCAT	Florida Comprehensive Assessment Test
FEAP	Florida Educator Accomplished Practice
FFT	Framework for Teaching
FRPL	free or reduced-price lunch
FSA	Florida Standards Assessments
FY	fiscal year
HC	human capital
HCPS	Hillsborough County Public Schools
HCTA	Hillsborough Classroom Teachers Association
HE	highly effective
HR	human resource
HS	high school
ILT	instructional leadership team
IP	Intensive Partnerships for Effective Teaching
IT	information technology
ITL2	instructional teacher leader 2

iZone	Innovation Zone
LAUSD	Los Angeles Unified School District
LES	learning environment specialist
LIM	low-income minority
MAP	Merit Award Program
MCS	Memphis City Schools
MET	Measures of Effective Teaching
MS	middle school
M-SCEA	Memphis–Shelby County Education Association
NAEP	National Assessment of Educational Progress
NBPTS	National Board for Professional Teaching Standards
NCES	National Center for Education Statistics
NCLB	No Child Left Behind
NGSS	Next Generation Science Standards
NI	needs improvement
N/S	not significant
P	proficient
PAR	peer assistance and review
PD	professional development
PFT	Pittsburgh Federation of Teachers
PIT	performance improvement team
PLC	professional learning community
POWER	Performance Outcomes with Effective Rewards
PPS	Pittsburgh Public Schools

PRC	Promise-Readiness Corps
PSSA	Pennsylvania System of School Assessment
PUC	Partnerships to Uplift Communities or Partnerships to Uplift Communities Schools
RFP	request for proposals
RISE	Research-Based Inclusive System of Evaluation
RTT	Race to the Top
S	satisfactory
SCS	Shelby County Schools
SEED	Supporting Effective Educator Development
SGP	student growth percentile
SIG	School Improvement Grant
SL	school leader
SLO	student learning objective
STAR	Students and Teachers Achieving Results
TAP	Teacher Advancement Program
TCAP	Tennessee Comprehensive Assessment Program
TCRP	the College-Ready Promise
TE	teaching effectiveness
TEM	Teacher Effectiveness Measure
TFA	Teach for America
TIF	Teacher Incentive Fund
TIP	Teacher Induction Program
TLE	Teacher and Leader Effectiveness

TNTP	the New Teacher Project
TPS	Tulsa Public Schools
TVAAS	Tennessee Value-Added Assessment System
U	unsatisfactory
UC	University of California
VAM	value-added modeling
VARC	Value-Added Research Center at the University of Wisconsin—Madison

CHAPTER ONE

Introduction

> A good personnel system encourages employees and managers to work together to set clear, achievable goals. Annual reviews are a diagnostic tool to help employees reflect on their performance, get honest feedback and create a plan for improvement. Many other businesses and public-sector employers embrace this approach, and that's where the focus should be in education: school leaders and teachers working together to get better.
>
> —Bill Gates, 2012

In Brief

What was the rationale for the Intensive Partnerships for Effective Teaching (IP) initiative, and how did we assess its implementation and impact?

RAND and the American Institutes for Research (AIR) conducted a six-year evaluation of the IP initiative, documenting the policies and practices (levers) *that were enacted and their effects. The study included annual surveys of teachers and school leaders (SLs), interviews of central-office administrators, case studies of schools, analyses of site records and student achievement test scores, and reviews of relevant documentation. To assess the impact of the initiative, we estimated changes in the levels and distribution of teaching effectiveness (TE) and the reform's impact on student outcomes.*

In this chapter, we provide an overview of the IP initiative and of the theory of action that underlies the initiative. We then briefly lay

out our objectives and approach, before discussing the structure of the remainder of this report.

Overview of the IP Initiative

The IP initiative was a large-scale, multiyear effort funded by the Bill & Melinda Gates Foundation to work in partnership with seven school districts and charter management organizations (CMOs) to bring about dramatic gains in the achievement of low-income minority (LIM) students by improving teacher workforces. The IP initiative began in 2010 and was designed to continue through the 2015–2016 school year. The foundation contracted with a team from RAND and AIR to evaluate the initiative's implementation and impact, and this is the final report of that evaluation. The following sections briefly describe the features of the initiative.

Rationale for the IP Initiative

In its invitation-only request for proposals (RFP) for the IP initiative, the foundation described the core belief motivating this effort:

> All young people, regardless of their background, deserve a rewarding education that results in the ability to earn a degree or credential with real value in the labor market. The foundation and its partners work together to ensure that all students, especially low-income and minority students, attend schools from kindergarten through college that make this vision a reality. (Bill & Melinda Gates Foundation, 2009, p. 1)

The RFP further explains that the foundation's thinking about empowering effective teachers was shaped by ideas contained in three publications: (1) suggestions for changes to recruitment, tenure, and compensation policies designed to increase the proportion of effective teachers where traditional certification and tenure policies have failed (Staiger, Gordon, and Kane, 2006); (2) descriptions of policy changes to increase the supply of effective teachers and distribute them more equitably (Murnane and Steele, 2007); and (3) suggestions for evaluat-

ing teachers using a combination of direct measures of teaching and measures of student achievement (Toch and Rothman, 2008).

At the time of the RFP, the foundation was engaged in a parallel research effort—the Measures of Effective Teaching (MET) study—to identify the best combination of measures to use to judge TE (Kane and Staiger, 2012). The IP initiative was designed to leverage such robust TE measures to support improved human-resource (HR) policies—such as more-tailored recruitment; more-thoughtful placement of teachers to best serve LIM students; effectiveness-based promotion and tenure; evaluation-linked professional development (PD); and effectiveness-based compensation—that might raise the overall level of TE and give LIM students greater access to effective teaching.

Goals for the IP Initiative

The initiative's primary goal was to dramatically improve outcomes (including achievement, graduation, and college-going) for LIM students in each of the participating sites. This improvement was to be accomplished by increasing TE overall and by ensuring that LIM students had increased access to effective teaching—the intermediate goals of the initiative. The IP theory of action describes the mechanisms for accomplishing these primary and intermediate goals.

Participants in the IP Initiative

After a thorough screening process, the foundation selected three intermediate-size school districts and four CMOs to participate in the IP initiative. The three school districts were Hillsborough County Public Schools (HCPS) in Florida; Memphis City Schools (MCS) in Tennessee (which merged with Shelby County Schools, or SCS, during the initiative); and Pittsburgh Public Schools (PPS) in Pennsylvania. These three districts were chosen from the population of districts enrolling at least 25,000 students, with at least 40 percent of students eligible for free or reduced-price lunch (FRPL), and were located in states that had relatively rigorous standards for granting tenure to teachers. The four CMOs were Alliance College-Ready Public Schools, Aspire Public Schools, Green Dot Public Schools, and Partnerships to Uplift Communities (PUC) Schools, all based in California; these CMOs are all

nonprofit organizations targeted at LIM students in areas with underperforming schools.[1]

The IP sites were not randomly selected, and findings from this evaluation do not necessarily generalize to any other district or CMO. The foundation selected sites that it thought offered "fertile soil" for cultivating the reforms. In particular, the selected sites were places where a confluence of support was assembled, including support from the district or CMO leadership, the school board, teacher professional organizations, and community political and philanthropic groups. Nevertheless, we believe that the chosen IP sites provide important lessons about the HR levers and the implementation of large-scale HR reform from which other sites can learn.

Timeline for the IP Initiative

Grants for the initiative sites were awarded midway through the 2009–2010 school year, and funding and support continued through the end of the 2015–2016 school year.[2]

The IP Theory of Action

Underlying the IP initiative was a theory of action that explains how the initiative was intended to achieve its goals. Specifically, the foundation staff members who designed the initiative identified a set of HR policies and practices—referred to as *levers*—that they believed would boost overall TE and improve student outcomes. Figure 1.1 illustrates

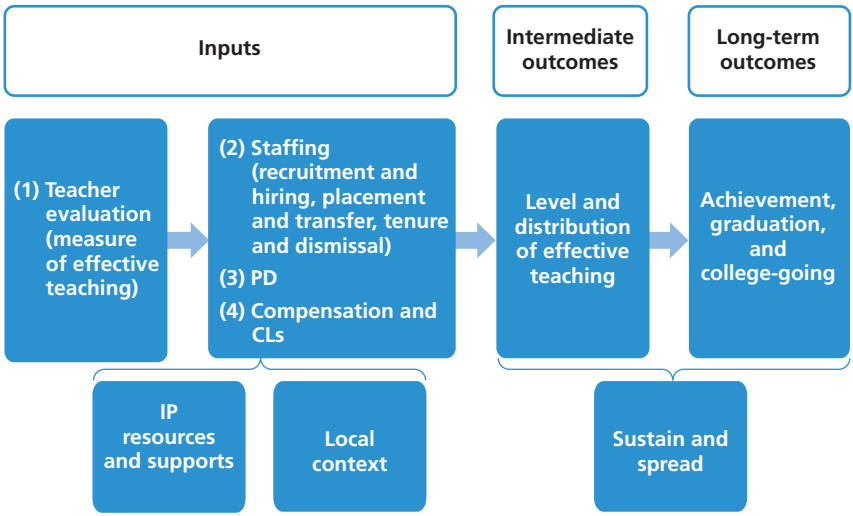
---

[1] Originally, there were five CMOs, but the Inner City Education Foundation Public Schools did not participate after 2010–2011. Initially, the CMOs operated jointly for the purposes of the IP initiative as the College-Ready Promise (TCRP), but, beginning in 2011–2012, the four remaining organizations were given separate grants from the foundation and operated independently, although a central TCRP office continued to coordinate modifications of some assessment measures and facilitate collaboration among staff from all four CMOs.

The foundation also supported similar but less intense efforts in a few other districts, which were referred to as the Accelerator Partnership sites. The Accelerator Partnership sites were not part of this evaluation.

[2] PPS and SCS received no-cost extensions to extend the grant through 2016–2017.

Figure 1.1
The IP Theory of Action



NOTE: This diagram is a distillation of various foundation documents.
RAND *RR2242-1.1*

the underlying theory of action, which incorporated changes to four major levers: (1) measures of effective teaching; (2) staffing policies (including recruitment and hiring, placement and transfer, and tenure and dismissal); (3) evaluation-linked PD; and (4) compensation and career ladders (CLs).

The central element of the initiative was a change to each site's teacher-evaluation policies. Specifically, administrators and staff in each site were expected to develop a shared vision of TE embodied in a robust measure of effective teaching that combined at least two elements: a direct measure of teaching practice and a measure of a teacher's contribution to student achievement growth. The TE measure might also include other elements, such as feedback from students and parents.

The TE measure was to be used as part of a larger reform of staffing, PD, and compensation and CL policies. Suggested staffing-related changes included improved recruitment and hiring practices, more-thoughtful teacher placement, and reformed tenure and dismissal poli-

cies. Evaluation-linked PD was seen as a way to meet teacher needs that the TE measure identified. Effectiveness-based compensation and the creation of new career pathways for effective teachers were additional ways in which the TE measure could be used to improve HR policies and enhance the teacher workforce. The developers of the IP initiative believed that, working together, these policies would lead to improvements in the teacher workforce—specifically, an overall improvement in effectiveness and greater access to effective teaching for LIM students. This, in turn, would boost student success. Each of the levers and component parts is described in greater detail in the following paragraphs. As we report in subsequent chapters, the sites differed in the extent to which they implemented these ideas in practice.

Lever 1: Teacher Evaluation, or Measure of Effective Teaching

Creating a robust measure of effective teaching was the starting point for each site's effort to improve its teacher-evaluation system. That process involved both developing a shared vision for TE and creating a composite measure to assess it. The foundation believed that it was possible to create a robust measure of effective teaching based on measures of growth in student achievement, systematic classroom observations, and other factors. When the IP initiative was being developed, the foundation was also undertaking the MET project, which was investigating relationships among such measures of effectiveness and would soon make recommendations about how to create a reliable and valid composite (Kane and Staiger, 2012). Although the MET study had not produced results when the IP initiative was launched, the MET recommendations were influential in the sites' eventual decisions about their composite effectiveness measures.

Lever 2: Staffing

The IP initiative envisioned changes to staffing policies that would ultimately boost student outcomes, including (1) improved procedures for recruitment and hiring to ensure a timely flow of good teacher candidates, (2) revised policies governing the placement of new teachers and the transfer of experienced teachers to ensure that schools serving high

proportions of LIM students had early access to strong candidates, and (3) effectiveness-based tenure and dismissal policies.

Lever 3: PD

The initiative's developers believed that one of the long-term advantages of having a rigorous measure of TE would be the ability to identify areas for teachers' improvement, so that individualized feedback (from observations) and targeted PD could be provided to boost the effectiveness of all teachers. Although evaluation-linked PD was not emphasized at the outset of the initiative, it grew in importance after the first year or two. The foundation and the sites realized that changes to dismissal and retention policies were likely to affect only a small proportion of teachers, while evaluation-linked PD could be a powerful way to improve the performance of the vast majority of teachers in the middle of the effectiveness distribution.

Lever 4: Compensation and CLs

The initiative assumed that a rigorous TE measure could be used as the basis for additional compensation and as part of the qualifications for specialized roles that would give teachers additional leadership opportunities (CL positions). Effectiveness-based compensation might come in the form of one-time bonuses or permanent salary increases tied to TE. CL positions would allow teachers to take on new responsibilities while maintaining instructional roles; for example, highly effective (HE) teachers might be assigned to spend part of their time mentoring new teachers, leading a data team in a collaborative teacher group, or coaching a group of teachers in a particular subject or grade. Both performance-based compensation and CL positions would be designed to reward more-effective teachers, retain them for longer periods of time, encourage the best teachers to work with the highest-need students, and motivate all teachers to improve. CL positions would also afford teachers leadership opportunities and encourage them to share their expertise with other teachers.

IP Resources and Supports

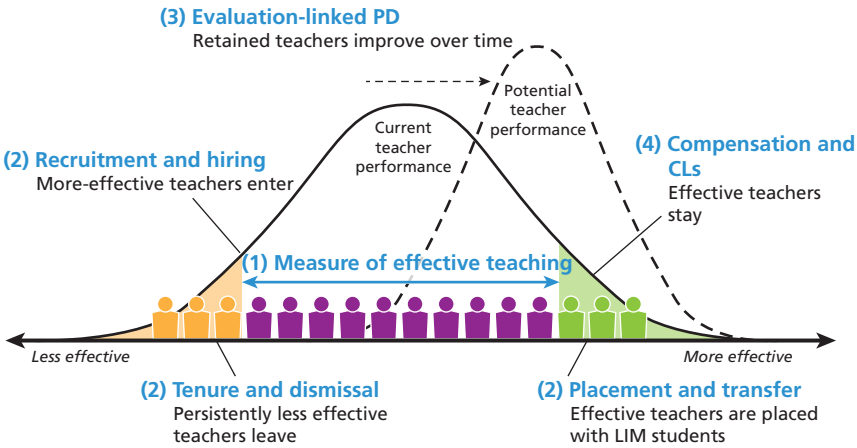
The sites received financial and professional resources from the foundation to support their change efforts. The foundation pledged $290 million to support the initiative over the seven-year period from 2009 through 2016, and the funds were awarded to the sites roughly proportionally to their numbers of students. Financial resources were used to support staff time for such purposes as developing a vision of effective teaching, adopting measures of effectiveness, informing teachers and other stakeholders about the reform, training observers, and engaging local communities. Sites also invested some of the resources to improve technology and infrastructure to support the reform (e.g., data systems, communication methods). Also, local philanthropic organizations and agencies pledged matching funds to increase the impact of the changes and ensure that the reforms could be sustained after the foundation's grant expired.

The foundation also provided expert consulting and support for planning and knowledge-sharing. It assigned each site a foundation program officer who spent up to one week per month on-site assisting with implementation planning and providing outside professional support where needed. The foundation also held semiannual "convenings" of all sites to provide training and encourage collaboration and shared problem-solving. In addition, the foundation provided sites with access to professional consulting organizations to assist with strategic planning and communications.

Intermediate Outcomes: Levels and Distribution of Effective Teaching

According to the theory of action, the four levers would operate singly and in combination to improve the overall distribution of TE in each site by raising the average effectiveness of teachers and narrowing the range of variation in that effectiveness. Figure 1.2 illustrates five mechanisms through which this change was expected to occur, all utilizing the robust measure of effective teaching each site was to adopt. First, revised tenure and dismissal policies would accelerate the removal of persistently ineffective teachers at the bottom of the distribution. Second, improved recruitment and hiring practices would

Figure 1.2
How Levers Are Expected to Improve the Distribution of Effective Teaching



SOURCE: Attributed to the New Teacher Project (TNTP).
RAND *RR2242-1.2*

bring in new teachers who were more effective when they began their careers and more open to improvement. Third, evaluation-linked PD would help all teachers improve their practice. Fourth, modified compensation policies and the development of CL positions would reward effectiveness, motivate teachers to improve, increase the retention of effective teachers, and allow more-effective teachers to mentor or coach others to improve their performance. Finally, changes in placement policies would increase LIM students' access to effective teaching. The net results would be a shift to the right in the distribution of TE in each site.

Long-Term Outcomes: Achievement, Graduation, and College-Going

The goal of the initiative was to improve outcomes for students, particularly for LIM students. First, the foundation and the sites expected to produce gains in student achievement, as measured by state tests in reading and mathematics. Second, the foundation hoped that these

gains would translate into dramatically higher graduation and college-going rates among high school (HS) students.[3]

Sustain and Spread

Although the main purpose of the initiative was to produce widespread improvement in outcomes for students in the IP sites by 2015–2016, the foundation also hoped that the sites and their communities would find ways to sustain the reforms after the grant period was over, and it expected the sites to share their experiences so that others could replicate the successful features of the initiative. The foundation asked the sites and their communities to develop plans to sustain the reforms after the grants ended. The reforms might be continued by sites transforming how district resources were used or by sites finding other sources of funds that could be used to support the reformed processes. The foundation also hoped that lessons that the sites learned would be relevant to other school systems. It encouraged the sites to make efforts to share their experiences and evolving insights with local districts in their geographic areas and with districts across the country that had similar characteristics and were facing similar challenges.

Influence of Local Context

Of course, the IP initiative did not operate in a vacuum, and the foundation, the sites, and the evaluators recognized that external factors were likely to play a role in its performance. These influences included both other initiatives undertaken by the sites and changes in the state and local context that occurred during the initiative. Although participation in the IP initiative required considerable focused effort by central-office administrators and educators, it did not preclude the sites from continuing old initiatives or undertaking new ones. For example, most sites were involved in curriculum reform efforts during this period, such as the implementation of the Common Core State Standards. Two of the sites (MCS and HCPS) participated in the MET study, which overlapped with the beginning of the IP initiative. Some sites initiated

---

[3] We did not assess college-going rates as part of the study because these data are difficult to obtain and because it was too soon to expect to find significant changes.

efforts to reform their principal-evaluation systems so they were more consistent with the new teacher-evaluation system. Such local reform efforts might have interacted with the IP initiative. In addition, most of the sites experienced one or more changes in leadership during the IP initiative. PPS had two superintendents during this period and a third shortly after the grant ended; SCS had three, and HCPS had two. Similarly, Alliance and Aspire each had two chief executive officers (CEOs) during the initiative. Some of these leadership changes were unexpected, while others were planned; in some cases, the new leader's priorities did not align with the goals of the IP initiative as well as the previous leader's had, creating uncertainty about the ongoing implementation of the initiative, as well as its sustainability. Similarly, state and regional policy changes might have influenced the success of the IP initiative. A prominent example was Florida's decision to eliminate teacher tenure statewide, which occurred early in the IP initiative. As a result, HCPS had to adapt its teacher-retention policies given this statewide change. Other external "shocks" occurred in all the sites: The Tennessee legislature passed a bill supporting the merger of MCS with SCS; the Pennsylvania Department of Education changed its statewide achievement tests; and the California Department of Education cut school funding dramatically and suspended the state testing program for two years.

Research shows that changes like these in the local, regional, or state context affect the implementation of new practices (Cohen and Spillane, 1992; Spillane, Reiser, and Reimer, 2002). We attempted to monitor a range of contextual factors, including changes in state and local HR policies, resource constraints, the involvement and influence of professional unions or associations, the national policy environment, and other factors that affected sites' efforts to implement the IP theory of action. In Chapter Fourteen, we revisit these external factors and their impact on the implementation of the IP initiative.

Overview of the IP Evaluation

The foundation hired RAND and AIR to conduct a thorough evaluation of the IP initiative. The evaluation had three major substudies: (1) an implementation evaluation, (2) an outcome study, and (3) an investigation of the replication of the reforms in other districts.[4] In this section, we briefly review the evaluation questions, data-collection activities, and analytic methods used in the outcome studies. Three interim reports published in 2016 provide more details about our methods (Stecher, Garet, Hamilton, et al., 2016; Gutierrez, Weinberger, and Engberg, 2016; Baird et al., 2016); much of the text about methods in this report has been adapted from those reports.

Evaluation Questions

The theory of action (see Figure 1.1) drove the evaluation questions. We designed the implementation study to provide information to help us understand the mechanisms through which the IP initiative might have achieved its long-term outcomes or, if it did not achieve these outcomes, why not. It focused on the left-hand side of the theory of action and was designed to answer five primary questions:

- Which HR policies and practices were implemented in each site as part of the IP initiative, when were they enacted, how many teachers did they affect, and were they consistent with best practices derived from research?
- How did teachers and SLs respond to the IP initiative reforms?
- What other local factors might have moderated the effects of the IP initiative?
- How did sites use the foundation's resources (and their own resources) to support the implementation of the initiative?

[4] The replication strand was completed with a report to the foundation in 2015. We do not include its findings in this report.

- How will the sites sustain the HR policies and practices after the grant periods end?[5]

The answers to these questions are interesting in their own right; they are also valuable because they provide a detailed description of the IP "treatment" that we use in interpreting changes in outcomes and in investigating the relative importance of the different levers.

The outcome study focused on the right-hand side of the theory of action; it was designed to answer three specific questions:

- How much better did students in the IP sites perform (in terms of achievement and graduation) than they would have without the initiative?
- Did TE improve overall, and did students, particularly LIM students, have greater access to HE teachers as a result of the IP initiative?[6]
- Which of the HR policy levers that characterize the IP initiative had the greatest impact on improved student performance or access to effective teaching?

Data Collection

Appendix A describes the data-collection procedures for our surveys, interviews, and archival achievement data in detail, and Appendix L describes the data-collection procedures for resource data; we provide a brief overview here.[7] Each year, we interviewed key central-office administrators in each IP site (approximately six to ten in each district

---

[5] We cannot fully answer this question about the future, but, in Chapter Fourteen, we discuss our limited information about the sites' intentions with respect to continuing HR policies.

[6] The first two questions are related but not causally linked. We assessed the overall impact by comparing student test performance (and other outcomes) with the scores of students from schools outside the IP sites. We assessed TE, in part, on the basis of value-added modeling (VAM) scores derived from the same student test results. However, the analyses differ, and they could yield seemingly contradictory results (e.g., TE in the IP sites could have improved even if the overall impact were null, if students in other sites were also improving).

[7] All appendixes for this report are available online at this report's product page (www.rand.org/t/RR2242).

and three to five in each CMO) who were involved in making, implementing, or reviewing HR policy;[8] we also interviewed selected local stakeholders. These interviews focused on the development and implementation of HR policies, the use of TE ratings, challenges encountered, time commitments related to HR administration, and interactions with the foundation and with other districts.

Each year, we also obtained site fiscal records, staff administrative records, student administrative records, and student achievement data. We also collected additional information about plans for implementation and about implementation itself by reviewing reports that each site prepared as part of its annual "stocktake" with the foundation—a one- to two-day meeting during which leaders at the site and representatives from the foundation reviewed progress made during the past year and worked together to plan for the next year. In addition, we collected publicly available school-level test results for every school in the state in which each IP site was located.

We administered web-based surveys to SLs and teachers each spring, beginning in 2011. (We did not survey teachers in 2012.) The surveys asked about the respondent's role and responsibilities, their opinions about the teacher-evaluation process, their use of TE ratings and evaluation data, their experiences with and opinions about their site's other HR policies (e.g., policies related to tenure, dismissal, compensation, and CLs), their PD, and their time allocation among various activities. We designed the surveys to help us understand teachers' and SLs' reactions to the initiative along dimensions that are likely to be related to implementation and long-term sustainability; these dimensions include awareness, endorsement, sense of fairness, and perceived effects. We surveyed all SLs and a sample of teachers from every school within each of the seven IP sites. We used a stratified random sampling procedure to select the teachers, taking into account subject areas taught and years of teaching experience.

---

8 The actual numbers varied from site to site and year to year.

Finally, we selected seven schools from each site that agreed to serve as case-study schools, and we visited these schools biennially.[9] During each half-day to day-and-a-half visit, we interviewed the principal, selected teachers, and teacher coaches and mentors, and we conducted focus groups with teachers.[10] The case studies focused on implementation of the levers and how they affected the school, the staff, and classroom instruction.

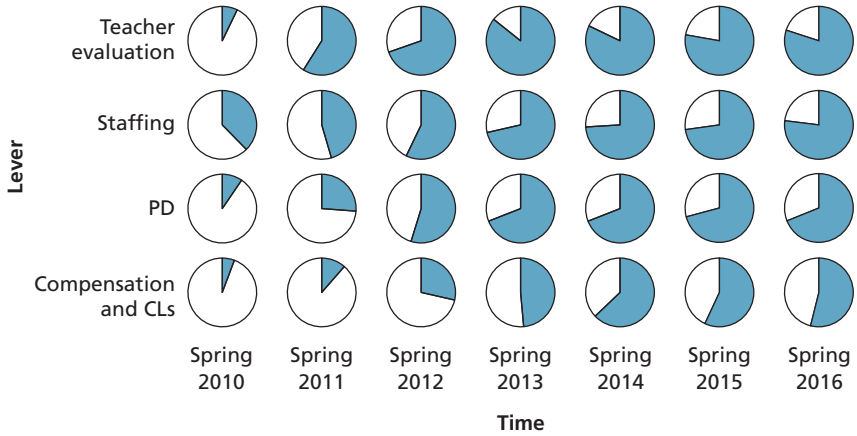
Analysis

Appendix A describes the analysis procedures for surveys, interviews, and achievement effects; Appendix L does the same for resource data. We describe supplemental analytic procedures related to intermediate and long-term outcomes in appendixes associated with individual chapters (Appendix I for Chapter Seven, Appendix O for Chapter Eleven, Appendix R for Chapter Twelve, Appendix T for Chapter Thirteen); we provide a brief overview of the analytic methods here. We used two approaches to examine the extent to which each site implemented the levers—one was more quantitative and the other more qualitative. The quantitative approach involved identifying specific policies and practices that were consistent with the foundation's conceptualization of the key elements of the reforms. Each site planned to implement many but not necessarily all of these individual practices, but we used the complete set of potential practices as our comparison set for tracking progress. We classified each site's status with respect to each practice annually based on administrative data, central-office interviews, stocktake reports, and supplementary phone calls. This allowed us to track the percentage of policies or practices in effect in each site each year. It provided a general overview of the progress of implementation. Figure 1.3 shows the average proportion of the practices we

[9] We visited all the case-study schools in 2010–2011; we then divided the sample in half and visited each half every other year. In each school's "off" years, we interviewed the SL by telephone.

[10] Again, the numbers of teachers who were interviewed or participated in focus groups varied from site to site and year to year.

Figure 1.3
Average Proportion of IP Levers Implemented, Springs 2010–2016



RAND *RR2242-1.3*

tracked that were implemented each year.[11] The figure illustrates two general points. First, sites implemented each lever gradually, with some changes occurring in the first or second year of the initiative and other changes occurring later. Second, sites implemented some of the levers more quickly than others; for example, implementation of new teacher-evaluation policies typically preceded implementation of the other levers. Although we do not show it here, each site followed a slightly different trajectory, in terms of both the specific practices it chose to implement and the speed of implementation (see Stecher, Garet, Hamilton, et al., 2016, for further detail).

Yet, this quantitative approach to characterizing implementation had some important limitations; most notably, it did not convey any information about the scope, duration, and coverage of the policies and practices that sites implemented. To try to identify connections between HR policies and student outcomes, we needed to know about quality as well as quantity. For that reason, we also engaged in a more

[11] Because each site adopted only a subset of these practices, we did not expect to see all practices implemented in every site. Consequently, a site might have fully implemented its planned policies without producing a fully shaded circle.

qualitative summary of implementation that captured the conditions or policies that sites changed, how they changed, how long changes lasted, what proportion of teachers were affected, and so forth. We compiled the data into prose tables that were more useful for understanding the implementation process, although they were not as visually engaging. The individual lever chapters contain these detailed prose tables (lever 1 in Chapter Three, lever 2 in Chapters Four and Five, lever 3 in Chapter Six, and lever 4 in Chapters Seven and Eight).

We analyzed the survey results annually by site and across sites, focusing on frequency distributions of individual survey items. In addition, we selectively compared responses from groups of teachers based on secondary factors, such as experience, TE rating, and the school's proportion of students who were LIM. Teacher survey response rates ranged from 61 percent to 86 percent across years and sites, and SL response rates ranged from 41 percent to 83 percent.[12] We applied sampling and nonresponse weights to the final survey responses so the results would reflect each site as a whole.

We examined trends in the distribution of effective teachers between LIM students and other students. This analysis involved two steps. The first step was to estimate each teacher's contributions to his or her students' achievement—that is, that teacher's value added. We used a common value-added model with teacher-linked data on the mathematics and reading performance of students in grades 3 through 8 in the 2006–2007 through 2014–2015 school years. The second step was to examine the sorting of teachers by their value added between LIM students and other students for each site. We repeated the analysis separately for elementary and middle school (MS) grades.

We also estimated the impact of the IP initiative in each site on two main outcomes: student test results and HS dropout rates. We assessed progress toward these goals by comparing the performance of students in the IP sites with an estimate of their performance absent the initiative. We estimated this "counterfactual" result based on the per-

[12] In the spring 2016 surveys, the leadership at Alliance severely restricted our access to teachers and SLs; as a result, our 2016 response rates for Alliance were lower (16 percent for teachers and 15 percent for SLs).

formance of comparable students in the same state who were not participating in the IP initiative. We based our estimate of this impact on a difference-in-differences (DiD) methodology. This method compared outcomes before and after the initiative's implementation between the schools in the IP sites and the rest of the schools in the same state.

Although we had originally hoped to estimate individual levers' unique effects on specific outcomes, this proved more difficult than anticipated for a variety of reasons. As we noted earlier, sites implemented practices gradually over two or more years and sometimes made modifications to those practices. In addition, the sites tended to implement multiple levers at the same time (e.g., a site would implement a new TE measure and new recruitment and hiring policies in the same year), and multiple levers affected the same outcomes (e.g., effectiveness-based compensation and CLs both influenced teacher retention). As a result, rather than examining the effects of individual levers, we examined the combined effects of sets of levers on two broad outcomes: the effectiveness of newly hired teachers and the retention of effective teachers. To examine effects on new teachers, we tested whether the enactment of recruitment and hiring policies over time improved the effectiveness of recently hired teachers relative to that of experienced teachers. To examine the effects on retention, we tested whether, over time, the enactment of policies designed to retain effective teachers improved the percentage of effective teachers who returned to teaching in the same district in the subsequent year or increased the percentage of less effective teachers who did not return.

Cross-Site Versus Site-Specific Findings

One of the major challenges we faced in presenting the results of the evaluation was the tension between drawing overall findings that were valid for the initiative as a whole and drawing narrower findings that reflected the distinct features of each of the sites. On the one hand, the IP initiative was conceived as an HR improvement strategy that could be applied in any district; the developers, and educators more generally, want to know whether it would likely be effective if adopted in other sites. On the other hand, each site's implementation differed from those at the other sites, and each site's conditions were unique.

An educator considering similar reforms would benefit from knowing which contextual factors and implementation options influenced effectiveness. We attempt to address the tension between broader and narrower results by presenting site-specific information in each discussion while also trying to draw overall conclusions when they are warranted.

Reporting

We provided formative feedback to the foundation and the sites throughout the initiative, including annual results of our surveys and periodic progress reports and technical reports on measuring impact. As noted above, in 2016, we issued three interim reports focused on the implementation of the initiative through 2013–2014, changes in LIM students' access to effective teaching, and overall impact on student achievement and graduation rates. These reports provide details about research methods and procedures to which we only allude in the present report.

Remainder of the Report

We present the remainder of this report in two parts and 13 chapters. Chapter Two describes each of the seven IP sites and offers some background and context that might facilitate interpretation of the implementation and outcomes described in the remainder of the report. Chapters Three through Nine describe the results of the evaluation that focus on the *implementation of the reform*, focusing on sites' efforts to implement each of the levers and the sites' use of the foundation's resources. These chapters attempt to answer the question, "When and how thoroughly did sites implement each of the individual elements of the reform?" In most cases, we document how quickly each of the levers was enacted; educators' reactions to the policies in each site, including perceived effects; and, where applicable, the extent to which thresholds for performance identified many or few teachers for rewards or sanctions. Chapter Three describes the sites' measurement of TE, a key aspect of teacher evaluation. Chapter Four describes staffing reforms related to recruitment and hiring procedures and changes to placement

and transfer policies. Chapter Five describes staffing policies related to tenure and dismissal. Chapter Six presents information about PD and how it was linked to teacher evaluation. Chapters Seven and Eight cover compensation and CL policies, respectively. In Chapter Nine, we examine the cost of the initiative, with a focus on how the sites used the foundation's resources and the resources they raised locally.

Chapters Ten through Thirteen present the results of the evaluation of *the reform's effect on intermediate and long-term outcomes*. In these chapters, we address the question, "Did the IP initiative achieve its primary goals?" First, we present evidence on the intermediate goals of hiring more-effective teachers (Chapter Ten), increasing the retention of effective teachers (Chapter Eleven), and improving the overall distribution of TE and LIM students' access to effective teachers (Chapter Twelve). Then, in Chapter Thirteen, we examine the initiative's overall impact on student outcomes. In Chapter Fourteen, we draw conclusions about the effectiveness of the initiative and the factors that influenced its effectiveness. We also summarize the information we have about the sites' plans for sustaining the reforms after the end of the grants.

Appendixes A through U provide supplementary information:[13]

- Appendix A describes our survey, interview, and archival academic data collection and analysis.
- Appendix B provides supplementary material for Chapter Three on TE measures in the sites.
- Appendix C provides additional exhibits for Chapter Three.
- Appendix D provides supplementary material for Chapter Four on the sites' recruiting, hiring, placement, and transfer policies.
- Appendix E provides supplementary material for Chapter Five on tenure and dismissal policies in the sites.
- Appendix F provides supplementary material for Chapter Six on the sites' PD policies.
- Appendix G provides additional exhibits for Chapter Six.

[13] All of the appendixes are available online only, at this report's product page (www.rand.org/t/RR2242).

- Appendix H provides supplementary material for Chapter Seven on compensation policies in the sites.
- Appendix I explains our methods for Chapter Seven for analyzing the relationships between the three levers.
- Appendix J provides supplementary material for Chapter Eight on CL policies in the sites.
- Appendix K provides additional exhibits for Chapter Eight.
- Appendix L explains our methods for analyzing the resources invested in the initiative.
- Appendix M provides additional exhibits for Chapter Nine.
- Appendix N provides additional exhibits for Chapter Ten.
- Appendix O explains how we estimated the relationship between TE and teacher retention.
- Appendix P provides additional exhibits for Chapter Eleven.
- Appendix Q provides additional exhibits for Chapter Twelve.
- Appendix R explains how we determined the initiative's effects on TE and LIM students' access to effective teaching.
- Appendix S provides additional exhibits for Chapter Thirteen.
- Appendix T explains how we estimated the initiatives effects on student outcomes.
- Appendix U provides additional impact estimates for Chapter Thirteen.

Next Steps

Although our overall evaluation of the IP initiative has concluded, we will continue to monitor teacher and SL attitudes and student outcomes through 2017–2018 to see whether the reforms' impact changes as the policies mature. We are also investigating variation in outcomes within the sites to try to understand what features are associated with better or worse outcomes at individual schools.

CHAPTER TWO

The IP Sites

In Brief

The foundation selected seven sites to participate in the IP initiative; these specific sites were chosen to provide a test of the efficacy of the IP theory of action in a diverse set of districts and CMOs. The sites varied with respect to enrollment, geography, and structure, but LIM students made up a moderate to high proportion of each site's student body. Most of the sites had implemented or experimented with some IP-like reforms prior to the initiative and were subject to federal or state policies that intersected with the IP initiative after the grants were awarded. The sites differed in the types of local contextual issues that shaped their reforms, such as the presence and role of teacher organizations and turnover of key site leaders.

Introduction

In this chapter, we provide background information on each of the sites to help the reader understand the sites' specific characteristics, as well as the sites' prior experience with the policies and practices that were key to the initiative. The sites included three districts (HCPS in Florida, MCS [now SCS] in Tennessee, and PPS in Pennsylvania) and four CMOs (Alliance, Aspire, Green Dot, and PUC). We describe the demographic characteristics of each site, the status of the IP levers prior to the initiative, key provisions of state and federal policy that affected the site during the period under study, local issues that were salient to the site, the status of the local teacher labor market, and any related ini-

tiatives the site was undertaking that might have influenced the implementation of the IP initiative. We have organized the descriptions of the CMOs slightly differently from those of the three districts, with common background information provided first, followed by specific details for each of the four CMOs.

Demographic Characteristics of the IP Sites

Table 2.1 summarizes key demographic features of the seven sites at the beginning (2009–2010) and toward the end (2014–2015) of the IP initiative. For the sake of consistency, we drew the data from the Common Core of Data prepared by the National Center for Education Statistics (NCES) (NCES, undated [b]), and they might not exactly match counts, generally derived from site-provided data, presented in other places in this report.

At the beginning of the initiative, the sites varied in size from ten schools (with 124 teachers and 2,667 students) to 228 schools (with 12,840 teachers and 185,258 students), but every site had at least 55 percent of students being low income and at least 51 percent minority. The demographic characteristics of students did not change markedly during the initiative, but the CMOs all grew in size by opening or taking over additional schools.

The Districts

HCPS

Background

HCPS, serving the Tampa, Florida, area, is the third-largest public school district in Florida and among the ten largest nationwide. It is the largest of the IP sites (see Table 2.1 for demographic information), and, appropriately, it received the largest Gates Foundation grant, totaling $81 million. HCPS named its IP initiative Empowering Effective Teachers (EET). The Common Core of Data categorizes HCPS as

Table 2.1
Characteristics of the IP Sites, 2009–2010 and 2014–2015

	2009–2010					2014–2015				
			Percentage of Students Who Are					Percentage of Students Who Are		
Site	Schools	Enrollees	Low Income	Minorities	Teachers	Schools	Enrollees	Low Income	Minorities	Teachers
HCPS	228	185,258	55	51	12,840	229	189,830	63	61	13,603
PPS	65	27,341	74	58	1,958	54	23,979	68	63	1,690
SCS	184	104,346	78	91	6,385	168	104,396	73	89	5,780
Alliance	16	5,541	93	99	273	26	12,927	95	99	615
Aspire	25	7,695	72	82	384	38	15,488	81	89	598
Green Dot	13	7,095	93	99	423	19	11,093	94	99	511
PUC	10	2,667	77	98	124	15	4,719	79	98	224

SOURCE: NCES, undated (b).

NOTE: Data in all columns reflect only operational schools having nonmissing enrollment and serving at least one grade level from kindergarten through grade 12 (i.e., we exclude early childhood centers and adult education centers). In the three districts, we exclude charter schools. SCS information reflects MCS for 2009–2010 and SCS for 2014–2015. *Low-income* is defined as eligible for FRPL. *Minority* consists of black, Hispanic, American Indian, or (2014–2015 only) multiracial. *Teachers* reflects full-time–equivalent classroom teachers, rounded to nearest integer. Data might not be completely consistent with numbers reported elsewhere in this report because of different sources of data and different restrictions made.

a large suburban school district, but the district contains schools that are located in urban, suburban, and rural areas.

IP Levers Prior to the Initiative

HCPS had three related programs partially implemented before the start of the initiative: a teacher-evaluation system, school-based hiring practices, and a merit pay program. The district also offered a robust suite of PD programs, although these were not tied to teacher evaluation. HCPS also participated in the Gates Foundation–funded MET project, which studied various measures of TE and the relationships among them.

Teacher evaluation was based entirely on classroom observations, typically one observation conducted by a school administrator every three years for experienced teachers. Observations used a rubric based on state guidelines, called the Florida Educator Accomplished Practices (FEAP). The rating system had four tiers: outstanding, satisfactory (S), needs improvement (NI), and unsatisfactory (U). These ratings were used primarily to inform tenure and dismissal decisions. In addition, at the time the IP grant was awarded, HCPS was already in the process of developing a new rubric based on the 22 components of professional practice from Charlotte Danielson's Framework for Teaching (FFT) (C. Danielson, 2013). Also prior to the initiative, HCPS had developed or identified assessments to measure student performance in all subjects and grade levels. As HCPS implemented the EET initiative, this universal student testing allowed HCPS to calculate some form of student growth measure for every teacher.

Another practice already in place at HCPS was school-level control over hiring. In every HCPS school, school administrators had, and still have, final authority over hiring, retention, and dismissal decisions in that school. The district also already offered financial incentives to work at high-need schools, and, in the highest-need schools, called Renaissance schools, transferring teachers did not have preference over new hires.

Florida statute 1012.01(2)(a)–(d), codified in 2005–2006, mandated merit pay for public school teachers in Florida. In response, HCPS adopted a Merit Award Program (MAP), beginning in 2006–

2007, available to any teacher who had been with the district for at least four years. The top 25 percent of teachers, as measured by student scores on state tests, local standardized tests, and a crude measure of value added, received a modest salary adjustment. The district did not refer to this award as a bonus because a teacher was eligible only if he or she continued teaching in the district the following year. Starting in 2007–2008, the district also offered special bonuses to teachers who taught in certain high-need schools, referred to as Performance Outcomes with Effective Rewards (POWER) schools.

State and Federal Policy Influences

Two major state policy changes affected HCPS during the time of the initiative. First, in July 2011, as part of Senate Bill 736 (Florida Senate, 2011), the state of Florida abolished the practice of granting tenure. Teachers who already had tenure were allowed to retain their status, but tenure could not be granted to any new teachers. HCPS responded by offering every newly hired teacher "nonprobationary" status after three years of satisfactory performance as a probationary teacher and a fourth-year appointment to a teaching position. (See Chapter Five for more details.) The other big state policy change was Florida's adoption, in 2014, of a modified version of the Common Core State Standards, referred to as the Florida standards. These new standards led to a change in the state achievement test, from the Florida Comprehensive Assessment Test (FCAT), which had been used since 1999, to the Florida Standards Assessment (FSA), in spring 2015.

HCPS received the first of three federal Teacher Incentive Fund (TIF) grants during the 2007–2008 school year; these grants, which provided financial incentives to HE teachers in selected high-need schools, continued during the initiative. The POWER1 grant, affecting 116 high-need schools, ran from 2007–2008 through 2011–2012; POWER2, involving 35 schools, ran from 2010–2011 through 2014–2015; and POWER3, including 30 schools, extended from 2012–2013 through 2016–2017. Each grant provided every HE teacher (defined as top quartile in POWER1 and with a TE rating of 4 or 5 in POWER2 and POWER3) with a lump-sum bonus. For more details about these bonuses, see Chapter Seven.

HCPS has been active in applying for other federal grants and programs. In addition to TIF, HCPS drew on multiple federal funding sources during the EET period, including funds from Florida's Race to the Top (RTT) grant, as well as the U.S. Department of Education's School Improvement Grant (SIG) and Supporting Effective Educator Development (SEED) grant programs.

Local Policy Influences and Role of the Teachers' Organization

At the start of the initiative, HCPS was led by superintendent MaryEllen Elia, who had been the superintendent since 2005 and was instrumental in promoting and shaping the district's IP initiative. In January 2015, the HCPS school board dismissed Elia, after a period of tension related largely to fiscal issues and communication and following a school board election. In her place, the board appointed the deputy superintendent, Jeff Eakins.

HCPS renegotiated its contract with the local teachers' union, the Hillsborough Classroom Teachers Association (HCTA) in 2010–2011. HCTA was deeply involved in the process of winning the IP grant—a union representative was a key member of the team that wrote the proposal, and teams of HCTA members routinely reviewed its provisions—and the union's support was instrumental in getting the TE measure into the new contract. The teachers approved this contract with 94 percent of the vote. Relations between HCTA and the district were generally positive throughout the initiative.

Teacher Labor Market

Florida has been experiencing a teacher shortage, which has affected HCPS, as well as other districts. Central-office staff reported particular difficulty hiring special-education teachers and teachers for English language learners (ELLs). HCPS actively recruits from other regions of the United States, including Puerto Rico, and offers hiring incentives for teachers who elsewhere have been rated as HE. Florida is a full-reciprocity state, meaning that teachers with valid credentials from other states can teach in Florida without meeting new coursework or assessment requirements.

PPS

Background

PPS is the second-largest school district in Pennsylvania but is the smallest of the three IP districts and has a total enrollment and number of schools closer to those of Aspire than to those of the other two districts (see Table 2.1 for demographic information). PPS decreased in size over the course of the initiative due in large part to regional population decline. PPS had two different superintendents from 2009–2010 through 2015–2016, which is the period covered by this report, and a third immediately thereafter. The first left to take another position in 2010, the second retired in the spring of 2016, and the third has served since the summer of 2016. PPS received $40 million from the Gates Foundation in 2009 for its Empowering Effective Teachers initiative.

PPS reorganized its central-office staff at least three times during the initiative. Central-office staff members told us in interviews that each reorganization was intended to support a different phase of the initiative: designing the reforms, implementing the reforms, and sustaining the reforms. In the early years of the initiative, the district created an Office of Teacher Effectiveness and hired a project manager and a team of staff who were tasked with developing and managing the project timeline, working across the district's various departments to develop the specific reforms, planning for implementation, and reporting regularly to the board and the foundation.

Once implementation of the reforms had begun, many Office of Teacher Effectiveness staff were reassigned to specific departments (e.g., HR, information technology [IT], PD) to better integrate the IP reforms throughout the district and to acculturate staff. The district also reorganized its HR department into four teams—one focused on recruiting, hiring, and onboarding new teachers (talent management); the second focused on teacher growth and evaluation systems and performance-based compensation systems (performance management); the third focused on benefits, employee relations, and compliance (workforce management); and the fourth focused on providing IT, data, and communication support across the department (shared services). One objective of the HR reorganization, according to staff we interviewed, was to integrate positions funded by the IP grant into

the district's operating budget. The district also experienced budget shortfalls during this time (discussed in more detail later), and many central-office positions (including, but not limited to IP-related positions) were eliminated.

PPS reorganized a third time in the final year of the grant, further integrating IP-specific staff into relevant departments, moving the final grant-funded positions into the operating budget, and eliminating positions that had become redundant as existing staff took on aspects of the IP work.

IP Levers Prior to the Initiative

PPS was not implementing many of the IP reforms prior to the initiative. Until 2008, PPS's teacher-evaluation measure consisted of an observation and included two performance categories: S and U. Although the rating was supposed to be based on the principal's observation of teacher practice, many veteran teachers we interviewed reported that their principals had never observed them. Throughout the initiative, and as of the writing of this report, PPS recruitment focused largely on recruiting teachers from Pennsylvania, and the district was obligated to find positions for all internal candidates (i.e., teachers transferring schools) before making offers to external candidates. No special supports (e.g., early hiring, exemption from internal transfer requirements) were given to the highest-need schools. As set by state policy, tenure was awarded after six semesters of satisfactory performance, and dismissal was based on two consecutive U ratings for a tenured teacher or one U rating for a pretenure teacher. PD was provided through district-wide in-service days or school-based sessions offered at a principal's discretion. The district offered content-specific coaching in math and literacy, but few other individualized resources were available to teachers, and principals were not expected to provide feedback after observing teachers. PPS offered roles (e.g., athletic coach, department chair) with more responsibility and additional pay in the form of a stipend, which is consistent with those generally offered in most districts, but PPS did not have CL roles, any form of performance-based pay for teachers, or policies or incentives designed to equitably distribute effective teachers.

State and Federal Policy Influences

Pennsylvania was awarded funding under the federal RTT grant program in 2011, and, in 2012, the state enacted Act 82 as part of its RTT commitment (Pennsylvania General Assembly, 2012). The law required new multiple-measure rating systems to evaluate teachers and principals, specified policies for teacher dismissal, and mandated participation in PD for all teachers rated as low performing. PPS has twice been granted a three-year approval to use its own rating system, which, although similar to the state system, has some small differences (e.g., using the Tripod student perception survey), to meet the Act 82 requirements; this approval will expire at the end of the 2019–2020 school year. With the implementation of Act 82, award of tenure and grounds for dismissal based on performance were determined by effectiveness as defined in the new teacher-evaluation system. Although Act 82 did not change the requirements for tenure or dismissal, it did introduce the four performance levels that translate to the S/U ratings on which tenure and dismissal are based. The district's contract with the union, together with Pennsylvania state law, governs teacher furlough policies, and both were based on seniority in 2015–2016 and in preceding years.

Pennsylvania has used its state assessment system (the Pennsylvania System of School Assessment, or PSSA) in grades 3 through 8, as well as grade 11, since before 2000. The PSSA has undergone some changes and expansion to new grade levels over the years in response to changes in state and federal law. Pennsylvania adopted the Common Core State Standards in 2010 and joined both consortia convened to develop Common Core–aligned assessments: the Partnership for the Assessment of Readiness for College and Career and the Smarter Balanced Assessment Consortium. Pennsylvania withdrew from both consortia in 2013 and adopted state-specific standards, known as the Pennsylvania Core Standards and based on the Common Core. In 2012–2013, Pennsylvania implemented the Keystone Exams, which were aligned to the Pennsylvania Core Standards, in algebra I, literature, and biology; these replaced the grade 11 PSSA exam. Pennsylvania retained the PSSA exams in grades 3 through 8 but revised the test content to be aligned with the Pennsylvania Core Standards; the newly

aligned version was administered starting in 2014–2015. In addition to revised content, the new version included a revised scoring system that increased the score required for students to attain proficiency.

In 2010, PPS received funding under the federal TIF grant program to support new performance-based compensation policies. PPS used the funds to implement CL stipends and effectiveness-based bonuses, along with performance-based bonuses and stipends for CL roles, as well as some operations costs for these programs. When the TIF grant ended, the programs were absorbed into the district's budget.

Local Policy Influences and Role of the Teachers' Organization

In 2008, PPS administrators, Pittsburgh Federation of Teachers (PFT) leaders, and teachers collaboratively developed the Research-Based Inclusive System of Evaluation (RISE), a new system for observing and evaluating teacher practice and for guiding teacher professional growth. Union membership is near universal among teachers in PPS. In interviews, district and PFT staff reported that the strong district–union partnership, along with teachers' support for RISE, played a large role in their receipt of IP funding, and RISE was incorporated into the TE measure developed for the initiative. The strong partnership continued for several years. For example, in 2010, PPS and PFT jointly developed and passed a collective bargaining agreement that codified new CL roles associated with additional compensation and a merit-based salary schedule for teachers hired after the agreement was passed. This agreement expired in 2015 but was extended for two years, through 2017. PPS and PFT are still negotiating the next contract as of the writing of this report.

However, unlike in the other two IP districts, the relationship between the union and the district became strained over time. One disagreement involved the issue of seniority-based furloughs. At the beginning of the IP initiative, PPS planned to improve the quality of its teacher workforce by hiring high-quality candidates and attracting some of the district's best teachers to the highest-need schools. This plan included the creation of two teacher academies, in which new hires would participate in a one-year residency during which they would receive on-the-job training and coaching from some of the dis-

trict's most-effective teachers. However, shrinking enrollment and budget shortfalls forced the district to furlough about 180 teachers. Per the district's contract with the union, as well as state law, new teachers hired into the academies' residency program would be the first to be furloughed. PPS and the union were not able to negotiate an exception, and, ultimately, the academies were not implemented.

Another disagreement involved setting cut points for the combined TE measure. Although the district–union relationship continued to be productive in many areas (e.g., CL roles), the relationship again became stressed in the 2012–2013 school year, when the first-year scores were combined to form a composite TE measure. PPS consulted experts, as well as the union and teachers, for input and proposed cut points based on pilot data, but the union objected because the district-proposed cut points classified relatively high percentages of teachers as ineffective. This disagreement became contentious enough to receive coverage in the local and national media. According to our interviews with central-office staff, to accommodate union concerns, the district adjusted the initial cut points so fewer teachers would be at risk of receiving low ratings. The resulting cut points produced a distribution in which nearly 15 percent of teachers were rated as NI or failing (F) in 2012–2013. In our interviews with them, union officials and many teachers said that they thought that this result was unfair and objected to the district's emphasis on high-stakes teacher evaluation.

Teacher Labor Market

According to central-office staff, the teacher labor market in Pennsylvania has historically been strong in many domains, such as early childhood, elementary education, and special education; staff reported that PPS generally has enough applicants in these areas. However, there are shortages of teacher candidates in mathematics and the sciences, particularly in MS and HS grades, and PPS has struggled to recruit qualified candidates to fill these types of positions. In the later years of the initiative, PPS began to expand its recruitment strategy to include teachers outside of Pennsylvania, with a focus on increasing the diversity of the teaching workforce. PPS SLs have some input about which teachers are hired to, placed in, or dismissed from their schools. To hire teachers,

SLs interview each candidate with a committee of other school administrators and teachers, including the union representative, and make recommendations to central-office HR, although some placement decisions are made centrally. Through the teacher-evaluation system, SLs have input into which teachers are dismissed from their schools, 50 to 100 percent of whom, depending on the teacher, the principal determines. In addition, teacher furloughs are based on seniority. PPS tenure and dismissal policies, which the state sets, have not changed during the initiative.

Related Initiatives

The Pittsburgh Promise, a scholarship program founded in 2006 as a partnership among the city, the district, and the community to help qualified Pittsburgh students attend in-state two- and four-year colleges and workforce certification programs, began awarding scholarships in 2008. Since 2008, the district has publicized the eligibility criteria, which include city residency, along with achievement and attendance requirements, and has encouraged teachers to help students meet them. One such effort is the Promise-Readiness Corps (PRC), a CL role implemented as part of the IP initiative.

One element of PPS's proposal to the Gates Foundation was to focus on improving the teaching and learning environment; this remained an area of focus throughout the grant period. Efforts in this area involved implementing a measure of teachers' perceptions of their school environment (the New Teacher Center's Teacher Working Conditions Survey), using those data to inform school improvement plans, providing teachers with more training in classroom management techniques, and implementing a CL role, the learning environment specialist (LES), focused on coaching teachers to improve the learning environment in their classrooms.

PPS also focused on improving equity and reducing racial disparities in educational outcomes throughout the IP grant period. Starting in 2008, all district staff participated in Courageous Conversation training, and teachers and principals were particularly encouraged to apply these skills in their classrooms and schools; this training was

ongoing for staff as of the writing of this report.[1] PPS's RISE teacher-observation rubric was adapted from Charlotte Danielson's FFT. PPS revised the rubric specifically to include a component to capture equitable teaching practices (e.g., engaging all students in classroom discussions and activities) as part of its definition of distinguished practice. The district established an Equity Office to oversee these efforts.

SCS

Background

The SCS district in Tennessee encompasses the city of Memphis and the surrounding county. It is the largest district in the state and the second-largest of the IP sites (see Table 2.1 for demographic information). The Gates Foundation awarded an IP grant of $90 million to MCS in 2009. In 2013, MCS merged with the surrounding county district in 2013, and the merged district (called SCS) and the foundation agreed to continue the grant. MCS called its program the Teacher Effectiveness Initiative; the name was later changed to the Teacher and Leader Effectiveness (TLE) initiative. In this chapter, we use the term *legacy MCS* to refer to the city district as it existed prior to the merger, *legacy SCS* to refer to the county district prior to the merger, and *SCS* to refer to the district after the merger.[2]

The size of the district has remained relatively stable over the course of the initiative, although enrollment has fluctuated. Enrollment increased because of the merger but then decreased in the fall of 2014, when several municipalities in the county left SCS to form their own school districts. In addition, starting in 2012, the state-run Achievement School District (ASD) took over some low-performing city schools, and SCS has closed several other schools in response to regional economic decline.

---

[1] Courageous Conversation is Pacific Educational Group's "protocol for effectively engaging, sustaining and deepening interracial dialogue" (Courageous Conversation and Pacific Educational Group, undated).

[2] In other chapters, for the sake of simplicity, we refer to the site as *SCS* even in reference to years prior to the merger, except for relatively rare occasions when it is important to focus on MCS prior to the merger, when we do use the separate terms specified here.

The district had three different superintendents from 2009–2010 to 2015–2016, (i.e., the period covered by this report): The first left at the end of his contract in 2013, at the start of the merger; the second oversaw the period of the merger but was not appointed to a longer contract; and the third, who was appointed to lead the new postmerger SCS in 2014, is still in the position as of the writing of this report.

The district reorganized its central-office staff at least twice over the course of the initiative. When the grant was awarded, the district established a Department of Teacher Talent and Effectiveness (DTTE) to lead the design of the reforms, monitor the timeline, work across the district's various departments to implement the reforms, and report regularly to the board and the foundation. The DTTE was led by a project manager and staffed by coordinators, each of whom was responsible for an individual lever (e.g., teacher evaluation, compensation, PD, CLs). The DTTE was disbanded during the merger, when many staff left and the central office cut positions because of the merger. The work of implementing the reforms was then integrated into the specific departments, many of which were under the supervision of the chief talent officer and, later, the chief academic officer (CAO). According to central-office staff, turnover in the positions tasked with supporting the IP reforms was an ongoing problem, particularly in some HR positions and in the project manager overseeing the CL and compensation reforms.

IP Levers Prior to the Initiative

The district had not implemented many effectiveness-related reforms on a large scale prior to the initiative but was experimenting with some policies that were consistent with the initiative's goals. For example, legacy MCS, like HCPS, participated in the MET project and piloted classroom-observation rubrics as a measure of TE. Although legacy MCS did not link HR decisions to teacher performance, the Tennessee Value-Added Assessment System (TVAAS), the state's system for assessing teachers' value added, had been in place for more than 20 years, and the data had been shared privately with teachers.

At the time of the district's IP proposal, few of legacy MCS's teacher staffing policies were in alignment with the goals of the IP ini-

tiative. Although Tennessee is a right-to-work state and the influence of the teacher's association was not strong, internal transfers were placed, and teachers were furloughed, in order of seniority, and teacher-hiring decisions were made at the school level. However, the district was cultivating a partnership with TNTP, which was working to expand recruitment to more states and alternative certification programs. In accordance with state policy, tenure was awarded after three years of satisfactory performance. District documents state that, although principals could dismiss teachers for poor performance, they rarely did; a poor-performing teacher was more likely to be transferred to another school.

Teacher PD prior to the initiative was largely online, although a principal could choose to organize training for the teachers in his or her school. Prior to the IP initiative, legacy MCS did provide induction, coaching, and mentoring for new teachers but did not otherwise use information about TE to target PD to teachers' individual needs. Legacy MCS relied primarily on a traditional, step-based salary schedule, but a small federally funded program provided group-based bonuses to teachers and administrators in schools with high achievement gains. There were no CL roles.

State and Federal Policy Influences

Tennessee was one of the first states to win an RTT grant, and the state quickly revised its teacher-evaluation policies and measures, as well as its hiring, placement, and tenure policies; all these changes were consistent with the goals of the IP initiative. The state's new teacher-evaluation policy required the use of multiple measures to evaluate teachers and principals, linked tenure and dismissal to performance, and required that teachers who received low ratings participate in PD. To ensure compliance with the state's requirements, legacy MCS delayed finalizing its own TE measure until the state requirements were published in July 2011. Legacy MCS implemented its TE measures in July 2011 in accordance with the state mandate. Two years into the IP initiative, in the spring of 2012, legacy MCS combined these measures using weights to create a measure of TE (the Teacher Effectiveness Measure, or TEM), which was used throughout the initiative. The measure met

state requirements for the number and type of data sources and the weights for each one but differed in some of the specific data sources used (e.g., the district's observation rubric differed from the one that the state adopted).

In July 2011, the state changed its tenure policy to link tenure more closely to teacher performance, as required by RTT. Teachers hired starting in July 2011 were required to complete a 45-month (five-year) probationary period and to achieve TE scores of 4 or 5 in the last two years of that period to receive tenure. Once tenure was granted, a TE rating of 3, 4, or 5 was required to maintain tenure. There was no change for teachers who had achieved tenure prior to July 2011. RTT also motivated the state to revise its teacher-hiring and teacher-placement practices. In 2013, state law changed to eliminate seniority-based interviewing, hiring, and furloughing. All teacher candidates were considered at the same time; districts, including SCS, no longer had an obligation to provide positions to internal candidates or assign teachers to positions.

Tennessee adopted new academic standards to be implemented in the 2009–2010 school year, the year the IP grant was awarded, and redesigned the state test (the Tennessee Comprehensive Assessment Program [TCAP]) to meet the higher standards. Scores dropped as a result. Tennessee began implementing the Common Core State Standards in 2011 and implemented Common Core–aligned state tests (called TNReady) in 2015–2016. According to news media reports from 2015–2016, the first administration of the TNReady tests was challenging because the online tests did not function as expected. The problems were so extensive that the state decided not to administer the tests in grades 3 through 8 and suspended use of TVAAS data in teacher evaluations unless doing so was to a teacher's benefit.

Local Policy Influences and the Role of the Teachers' Organization

When the county and city districts merged in 2013, the merged district largely adopted MCS's IP reforms, with a few modifications. The district retained legacy SCS's Tiered Coaching teacher support system, which included CL roles, into TLE, as the primary source of support for struggling teachers. SCS retained legacy MCS's version of the TE

measure but, in accordance with guidance from the state, incorporated portfolio-based measures of student growth for some subjects (e.g., world languages, fine arts) not tested by the state. Legacy MCS's measure of teacher content knowledge was replaced with a rubric-based measure of professionalism.

The teachers' organization (Memphis Educators' Association and, postmerger, the Memphis–Shelby County Education Association, or M-SCEA) partnered with the district to create the original IP proposal and, according to central-office staff, was initially on board with the proposed reforms. According to media reports and interviews with central-office and union staff, the union did not play a strong role in implementation of the reforms. M-SCEA membership was not mandatory for SCS teachers and was relatively low compared with similar organizational affiliations at other IP sites (M-SCEA membership was not mandatory for SCS teachers; about 40 to 60 percent belonged, compared with 100 percent of PPS teachers being members of PFT). Moreover, SCS was not required to collectively bargain with M-SCEA. Central-office staff reported that the district made an effort to give M-SCEA "a seat at the table" throughout the initiative and delayed implementation of compensation and CL reforms based in part on M-SCEA's concerns, despite the district's ability to create its own salary schedule as part of the RTT legislation.

Teacher Labor Market

According to central-office staff, the teacher labor market in SCS has historically been strong but local; that is, most candidates come from local teacher-preparation programs. TNTP, which took over teacher recruitment and staffing as part of the IP initiative, worked to expand the geographic diversity and size of the teacher applicant pool. The district also contracted with alternative certification programs, such as Teach for America (TFA) and the Memphis Teacher Residency, as a source for teacher candidates.

In 2012, Tennessee implemented the ASD, a state-led nongeographic district consisting of the state's lowest-performing schools, most of which are in Memphis. The ASD has worked with charter organizations to take over several schools each year and reorganize

them in the hope of improving student outcomes. In interviews, SCS central-office staff reported that the district was competing with the ASD for the best teacher candidates and worried that the district's lack of a competitive, performance-based salary schedule would contribute to the loss of high-quality candidates to the ASD.

Related Initiatives

Another element of SCS's proposal to the foundation was to improve the teaching and learning environment. This effort involved administering the Teaching, Empowering, Leading and Learning Tennessee Survey to gather data on teachers' perceptions of their school environment and encouraging principals to use these data in their school improvement plans. The district also implemented the Student Envoy program in selected schools; this program trains students to be leaders among their peers and advocate for improved educational experiences and trains teachers to implement the district's positive behavior intervention system, which focuses on solutions to discipline problems aimed at keeping students in school.

CMOs

Overview

All four of the CMOs participating in the IP initiative—Alliance College-Ready Public Schools, Aspire Public Schools, Green Dot Public Schools, and PUC—are nonprofit organizations targeted at underperforming LIM students in high-need communities. All of these CMOs started in California, and most of their schools are located in California. Students typically enter the schools (which, except in Aspire, are mostly middle and high schools) years behind their grade level, but the schools' culture is built on high expectations. Green Dot's website tells teacher applicants, "If you're ready to work as hard as you can to help others, join us" (Green Dot Public Schools, undated [b]). The CMOs' goal is to graduate all students prepared for success in college. Schools are small, and collaboration and teamwork are key aspects of the CMOs' culture. Green Dot is the only one of the CMOs with a

teachers' union. The teachers attracted to these organizations tend to be young and often new to teaching. Teacher turnover averages about 15 to 20 percent a year.[3]

TCRP

When the Gates Foundation invited the CMOs to submit a proposal for a program to improve student performance and promote college-going, the prospect fit many of the goals already held by the CMOs, and they considered that it would support much of the work in which they were already engaged. Initially, the Gates grant was awarded to TCRP, a consortium of the four CMOs plus one other CMO that withdrew in 2011. As differences among the CMOs emerged (e.g., the need for union approval of all measures in Green Dot, the advanced use of technology in Aspire), the CMOs found the TCRP consortium arrangement restrictive and asked the Gates Foundation to consider restructuring it. The foundation agreed, and, in the summer of 2011, each CMO submitted its own proposal and received a separate grant from the foundation. The individual proposals generally maintained the core components of the original consortium proposal, including the multiple evaluation instruments, and TCRP's central office continued to play a role in facilitating development of the initiative, especially in its early years.

IP Levers Prior to the Initiative

Evaluation in the CMOs prior to the IP initiative rested primarily on principals' observations of teachers, although each CMO also had other measures, such as student feedback, family feedback, and student achievement, that influenced teacher evaluation. Some of the CMOs used a state-developed rubric, and some developed their own rubrics. CMOs are not bound by the types of workforce policies typically found in traditional school districts, such as central hiring and teacher tenure. Even prior to the IP initiative, each principal had hiring authority for his or her own school, and none of the CMOs offered tenure. Employment is at will, and teachers' contracts are renewed (or not) annually.

---

[3] As reported in interviews with CMO HR directors, 2013.

Recruitment and hiring policies were a central concern of the CMOs because, besides the turnover rate mentioned above, they were all expanding their operations and opening new schools, so they regularly required additional staff. Typically, candidates applied online, and each CMO had its own recruitment and hiring process.

All four of the CMOs offered teachers a few days of CMO-wide or regional PD sessions, but PD was primarily the province of each school principal, with most schools holding weekly PD sessions. Generally, the content of these sessions varied by school, including operational or instructional matters at the discretion of the principal. Typically, however, some sessions each month were organized by department and some by grade level to foster collaboration. Few professional leadership positions were available to teachers in the CMOs. Prior to the initiative, each CMO used a step-and-column salary schedule.

State and Federal Policy Influences

Several state factors influenced the CMOs' implementation of the IP initiative, including the adoption of the Common Core State Standards, the transition to a new state assessment, and a budget crisis. After adopting the Common Core standards in 2010, the state of California embarked on a multiyear implementation process, including the adoption of Common Core–aligned mathematics programs in 2014 and aligned English language arts (ELA) and English language development programs in 2015. As part of the shift to the Common Core, in 2013–2014, California ceased the administration of the mathematics and ELA California Standards Tests (CSTs), which had been administered since 2002, and field-tested the Smarter Balanced assessments, which had been specifically developed by a multistate consortium to assess student mastery of the Common Core. The Smarter Balanced assessments were administered operationally starting in the following year, in the spring of 2015.

The change in test created a challenge for the CMOs. From 2011–2012 through 2012–2013, each CMO calculated its student achievement growth measure, a student growth percentile (SGP), based on the CST results, but, in 2014 (for the spring 2014 test administration), the state did not report test scores because the 2013–2014 school year's

administration of the Smarter Balanced assessments was considered a field test only.[4] Smarter Balanced scores were first reported in 2015. Because the CMOs needed a minimum of two years of data to calculate teachers' growth scores, the CMOs would not have had enough information to calculate SGPs based on the state test until the spring of 2016. Moreover, the CMOs had some concerns about the reliability of the first cycles of data from the Smarter Balanced measure. Thus, all the CMOs except Aspire elected not to use it as the student achievement component of their TE measure in 2016. From 2014 on, each CMO developed its own method for calculating a student growth measure to be used in teacher evaluation.

The great recession severely affected California's state school budget, with cuts of more than $20 billion (California Department of Education, 2012). As a result, no raises were available for CMO teachers from 2008–2009 through 2010–2011. The uncertain financial climate in the early years of the initiative made the CMOs reluctant to implement an effectiveness-based salary scale out of concern that they might not be able to sustain it. It was not until 2014 and 2015 that the state education budget began to revive ("Spending," undated) and the CMOs could expand their supports for teachers. However, the improved financial climate created its own problem. In the early years of the initiative, the depressed economy created a bounty of teacher candidates, despite the lack of raises. But, as the economy improved, the CMOs found it increasingly difficult to recruit skilled candidates.

Three of the CMOs applied for and received federal TIF grants in 2012. These grants focused on supporting evaluation systems that reward effective teachers, provide more professional opportunities for teachers, and drive HR decisions. They were designed to provide financial incentives for teachers to improve student achievement in high-need schools, and grant recipients were required to tie pay to teacher observations and student achievement. Alliance's and Aspire's grants continued for five years, while Green Dot's grant ended after three years. When Green Dot stopped including student growth as part of

[4] Test administration takes place in the spring of each school year, but the calculation and reporting of results typically occur in the summer or fall following that school year.

its teacher-evaluation measure in 2014, it could not tie compensation to student achievement.

Despite their commonalities, each of the CMOs has its own culture and began the IP reforms with its own perspective. In the rest of this section, we describe each CMO's inception and participation in the IP initiative, organizational structure, implementation of the IP levers prior to the initiative, and any local influences that affected their implementation of the levers.

Alliance

Inception and Participation in the IP Initiative

Alliance was first incorporated in 1991 as Los Angeles Educational Alliance for Reform Now. It teamed up with the Los Angeles Annenberg Metropolitan Project, a private organization dedicated to improving schools, and began to develop a network of charter schools. The first Alliance school opened in 2004, and it is now the largest charter school network in Los Angeles (see Table 2.1). As stated on the CMO's website,

> The mission of Alliance College-Ready Public Schools, a nonprofit charter management organization, is to open and operate a network of small high-performing 9–12 and 6–8 public schools in historically underachieving, low income, communities in California that will annually demonstrate student academic achievement growth and graduate students ready for success in college. (Alliance College-Ready Public Schools, undated)

Alliance staff saw their goal of college readiness as wholly consistent with the IP initiative and viewed the addition of Gates Foundation funds as a way to add resources to some of their strategies.

Organizational Structure

The Alliance organizational structure is highly decentralized. Principals are expected to achieve the core principles for the Alliance school model, but, as an administrator put it, "there's a lot of flexibility for how you get there." The central office changed organizational structure every year during the period of the initiative, adding and dropping

the CAO function, changing the reporting structure and the responsibilities of various departments, and reflecting the resignation and retirement of employees. Implementing TCRP was a separate function throughout the initiative, perceived primarily as supporting SLs in the evaluation process, with little integration with the academic departments. Leadership of the CMO changed in 2015, and among the goals of the new CEO was a more intentional focus on teacher and SL recruitment, selection, development, and retention, moving away from a "culture of evaluation" in order to build a "culture of coaching, feedback and support," and a new focus on ensuring college completion for Alliance graduates.[5]

IP Levers Prior to the Initiative

Prior to the IP initiative, principals evaluated teachers primarily through observations using the California Standards for the Teaching Profession rubric. Alliance also conducted family and student surveys, and considered overall school performance on the California state assessment at the school level, as part of teachers' evaluations. Principals were responsible for recruiting, screening, and hiring staff, and they conducted teacher evaluations and enjoyed a great deal of autonomy in providing PD for teachers. The content of weekly PD sessions was at the principal's discretion. Alliance viewed its principals as the primary resource for PD and did not begin hiring instructional coaches until 2013. Compensation was based on the traditional step-and-column salary structure and included a school-level bonus based on student achievement targets. Prior to the initiative, Alliance had one instructionally targeted teacher leadership position: HE mathematics teachers received a stipend to train teachers in other schools while continuing to teach in their own schools.

Federal, State, and Local Policy Influences

In 2012, Alliance received a federal TIF grant, which was used to support teacher-coach CL positions, effectiveness-based compensation, and a counselor position.

---

[5] Alliance CEO email communication, 2017.

Alliance began implementing Common Core online curricula in math and English for all students in 2013–2014. The shift involved new curriculum, new benchmark exams, and providing PD for teachers. The CEO thought that the TCRP rubric was compatible with the Common Core shifts. She said, "The way the rubric was designed, it did not need to change with Common Core." Generally, the teachers and SLs with whom we talked concurred.

Aspire

Inception and Participation in the IP Initiative

Aspire was founded in 1998 by Don Shalvey, the "father of California charter schools" under whose leadership the first California charter school was sponsored; he was also instrumental in the passage of California's Charter School Expansion Act of 1998. Aspire was the first nonprofit organization to develop a network of charter schools in underserved communities, and its motto was "College for certain." Shalvey said, "If we can create a long-lasting institution that focuses on underserved kids and gets them to graduation, we will have had an impact" (James Irvine Foundation, undated). Aspire perceived the IP initiative as a way to improve its college-readiness efforts, as well as its teacher-readiness and teacher-retention efforts.

Aspire is the largest of the four CMOs, with schools in southern, central, and northern California and, since 2013–2014, in Memphis, Tennessee, as part of that state's ASD. Over the course of the initiative, Aspire grew substantially (see Table 2.1).

Organizational Structure

Most of the Aspire senior leadership involved in the IP initiative remained involved throughout the implementation. There was a good deal of collaboration across departments, and central-office leaders shared common goals. In 2015, the CEO retired and was replaced that fall by a new CEO who had been an administrator with a group of East Coast CMOs. One change that occurred in 2015–2016 was the movement of the TE team from the Education Department to the Human Capital (HC) Department, where its focus shifted to more support for HC activities, such as teacher surveys.

IP Levers Prior to the Initiative

Prior to the initiative, Aspire had an evaluation system in place that included a teacher-observation rubric similar to the rubric that TCRP developed; the similarity eased the adoption of the new process. However, the previous rubric was not based on any research, nor was evidence gathered through multiple teacher observations calibrated across observers, so Aspire was enthusiastic about being able to calibrate observations and train principals on capturing objective evidence. Also included prior to the initiative were student and family surveys, benchmark assessments, and California state test data. The recruiting and hiring process in place in 2009–2010 continued throughout the initiative. PD practices, including weekly school sessions and regionally based coaches, also continued throughout. Only one teacher leadership position was available prior to the initiative: lead teacher, which was similar to a department chairperson. Aspire has always had a merit pay system integrated with a step-and-column compensation structure. However, as mentioned earlier, in the years just prior to the initiative, there were no raises.

Federal, State, and Local Policy Influences

In 2009, Aspire launched a major initiative focused on technology research and development that influenced all aspects of the IP initiative. The technology team, referred to as Team Godzilla, began developing a teacher data portal and a teacher resource library, Purple Planet, with long-term goals of creating a fully integrated student, teacher, and HR data platform. Purple Planet contained videos and lesson plans directly linked to TCRP's rubric and served as an online PD resource. Eventually, student and teacher data were integrated into the system.

In 2013, Aspire expanded into Tennessee. The first school opened in Memphis using a curriculum and assessments aligned to the Common Core. Aspire expected to use lessons learned from Memphis to roll out the Common Core in California; however, students in Memphis entered at such a low reading level that priorities shifted to emphasize reading instruction. In California, Aspire began training teachers for Common Core implementation in 2012–2013 and explicitly linked its instructional guides and videos of effective teaching to the indica-

tors in TCRP's rubric. However, a central-office administrator noted that the shift to the Common Core drew the PD emphasis away from TCRP instructional practices: "When momentum was building for TCRP, [coaches] had to be pulled to go deeply into Common Core content and curriculum."

Aspire received a TIF grant in 2012 and used the money to fund three roving observers and to make a second observer available to schools to help reduce principals' observation burden.

Green Dot

Inception and Participation in the IP Initiative

Green Dot was created in 1999 in response to the poor performance of the Los Angeles schools. As its website explains, "Green Dot had a vision of creating small, successful charter schools and in doing so, demonstrating to the school district and the public-at-large that there was a more effective way to provide public education to low-income, high-risk youth" (Green Dot Public Schools, undated [a]). Its first school opened in 2000; in 2008, it began a parallel effort to transform the worst-performing schools in the Los Angeles school system. Green Dot's first transformation school was a chronically low-performing HS in Watts with about 1,500 students. The organization has continued to grow, expanding to Memphis, Tennessee, in 2014 and Tacoma, Washington, in 2015 (see Table 2.1).

The CEO explained that the IP initiative complemented Green Dot's interest in TE and supplied the resources to pursue the improvement of instructional practices. As he said in an interview in 2010, "This is core work, but I'm not sure we would be tackling this [without the grant] in this current economic crisis, because it demands too much additional expenditure."

Organizational Structure

Senior leadership staff in the central office involved in the initiative were very stable throughout the period of the initiative, and, because the central office plays a fairly strong role in directing school programs, SLs and teachers received a unified vision of the IP reforms. The CAO who directed TCRP implementation for Green Dot eventually became

the president and CEO of Green Dot Public Schools California. Green Dot is the only nondistrict public school operator in California that has unionized teachers. Green Dot's teachers have organized as the Asociación de Maestros Unidos, a California Teachers Association and National Educational Association affiliate (the former being the California state affiliate of the latter).

IP Levers Prior to the Initiative

Prior to the initiative, Green Dot's evaluation system included semiannual teacher ratings for first-, second-, and third-year teachers based on the California Standards for the Teaching Profession rubric (and annual ratings thereafter); in addition, veteran teachers were also expected to conduct a project, while new teachers submitted a unit portfolio. The CMO also conducted family and student surveys, although they did not contribute to the teacher evaluation. Because of the depressed economy, Green Dot had thousands of job candidates. Candidates applied online, and Green Dot used an intensive screening process of phone interviews, the Haberman Educational Foundation's Star Teacher Pre-Screener, and personal interviews to narrow down the group of applicants and create a pool from which principals could select. To coordinate PD, Green Dot used a train-the-trainer model, with central-office staff training the principal, who then trained the teacher. PD was led by principals and a teacher team at each school. Principals were guided by monthly meetings with the director of teacher support and focused on data from benchmark assessments. The coaching staff was minimal. Prior to the initiative, Green Dot had two teacher leadership positions: the instructional leadership team, a school team of department chairs who assisted with instructional planning, and new-teacher mentors. Green Dot used a step-and-column salary structure, but, as with the other CMOs, teacher salaries were frozen in the years prior to and early in the IP initiative.

Federal, State, and Local Policy Influences

Green Dot is unique among the four CMOs in having a teacher union. The union has an explicit voice in school policy and the pace of reform. Key terms embodied in the Asociación de Maestros Unidos contract include teachers having an explicit say in school policy and curriculum;

no tenure or seniority preference; a professional workday rather than defined minutes; and flexibility to adjust the contract in critical areas over time. The union's ongoing impact on the initiative was to slow down implementation. At the union's urging, Green Dot agreed to pilot the TE measures for a year before implementing them operationally. Like all of the other CMOs, Green Dot described the initiative to teachers as a structure that would lead to increased pay for effective teachers. Given this understanding, the teachers' union was cautiously welcoming of the initiative and ratified the new evaluation system in the spring of 2012.

Common Core implementation at Green Dot occurred over several years, during which PD was dominated by what Green Dot called the *six shifts*—instructional shifts that the CMO considered to be key to implementing the Common Core.[6] The shifts had to do with such things as depth instead of breadth, pacing, and reading and writing across the curriculum. Although they were not directly related to TCRP's rubric, Green Dot considered them to be embedded in the rubric and developed a crosswalk between the Common Core State Standards and the rubric.

In 2012, Green Dot received a TIF grant, which it used to hire staff to develop teacher supports, such as instructional guides describing implementation of the observation rubric, and to support teacher-leader facilitators who developed training sessions around specific rubric indicators.

---

[6] The Common Core State Standards require 12 shifts—six in English language arts and literacy and six in mathematics—for full alignment. The six for English language arts and literacy are balancing informational and literary text, building knowledge in the disciplines, ascending a staircase of complexity, giving text-based answers, writing from sources, and building an academic vocabulary. The six for math are focus on the concepts prioritized in the standards, coherence from one grade to the next, fluency in core functions, deep understanding of a concept before moving on, application of concepts, and dual intensity (practicing and understanding). See EngageNY, undated.

PUC

Inception and Participation in the IP Initiative

PUC was founded on the belief that communities could be strengthened through high-quality public schools. Its stated mission was "to develop and manage high quality charter schools in densely populated urban communities with overcrowded and low achieving schools. We create school programs and cultures that result in college graduation for all students" (PUC, undated [b]). PUC is the smallest of the four CMOs. Its first school opened in 1999; it was the first charter MS in Los Angeles. It now has one charter school in Rochester, New York, as well as more than a dozen in northeast Los Angeles and the northeast San Fernando Valley.

PUC had been doing some work on TE and saw the Gates Foundation grant as a way to "get more money for what we're already doing, which would allow us to enhance that," in the words of one central-office staff member we interviewed.

Organizational Structure

PUC is a very close-knit organization that is highly responsive to SL and teacher feedback. For example, all principals meet weekly to discuss operations and initiatives. As the CMO's website states, "[s]hared leadership and collective decision-making" is one of its organizational practices (PUC, undated [a]). Although PUC had one CEO throughout the IP initiative, there was considerable turnover of senior leadership involved with the initiative, and central-office organization shifted to a more horizontal structure. PUC's CAO left in the spring of 2012, and the TCRP implementation lead, regional superintendents, and the PUC CEO assumed her duties. The initial TCRP implementation lead (who had subsequently become the director of teacher development) left in the spring of 2015, and his role was split among other staff, with responsibility for the implementation of the initiative eventually shifting to the three regional superintendents. Despite changing roles, most of the senior leadership remained at PUC throughout the initiative in one position or another, providing institutional continuity.

IP Levers Prior to the Initiative

Prior to the IP initiative, SLs conducted approximately 20 observations of each teacher using an in-house–developed rubric, leading to quarterly, midyear, and end-of-year evaluations. Quarterly benchmarks and California state assessment data were loosely included in teacher evaluation. Although principals had full hiring authority, recruitment and screening were done centrally, and applicant interviews were centrally organized. PUC had a strong culture around performance management. SLs developed growth goals and target areas at the school level and conducted weekly PD sessions. SLs were also expected to provide coaching because there was no central-office coaching staff. Prior to the initiative, PUC offered two teacher leadership opportunities: teaching demonstration classes for new teachers during a summer orientation learning lab and assisting new teachers in completing their California credential requirements. Although PUC utilized a step-and-column salary structure, prior to the IP initiative, it had already begun talking about tying compensation to teacher performance.

State and Local Policy Influences

Unlike the other CMOs, PUC did not apply for federal grants. Its approach was to not rely on soft money to fund ongoing programs. For example, it did not intend to rely on temporary grant money to fund salary increases, but rather planned to rearrange how money was distributed once it could model TE ratings and revenue.

The implementation of the Common Core State Standards shaped PD for several years in both CMO-wide sessions and school-level sessions. PUC began training teachers for the Common Core in 2013–2014. The Common Core had a minor effect on the TCRP observation rubric. PUC revised its instructional guides for the rubric indicators to explain how proficiency on an indicator would equate to student success on the Common Core standards. It also made a few adjustments to the rubric (e.g., expanding descriptions to reflect the rigor of the Common Core and adding one indicator on planning for student self-monitoring).

CHAPTER THREE

Measures of Effective Teaching

> The evaluation system is only as good as the people implementing it.
>
> —teacher, written comment in survey

In Brief

What measures did sites adopt to assess TE, and what did educators think of their sites' teacher-evaluation systems?

All the IP sites designed multiple-measure teacher-evaluation systems. Measures of student achievement growth and measures based on observations of teachers' practices had the greatest weights in each teacher's overall rating, and sites took different approaches to measuring these and other components. When we analyzed the distributions of teachers' evaluation scores, we found that, over time, an increasing number of teachers were assigned to the highest categories and fewer teachers were assigned to the lowest. Sites faced challenges in implementing the teacher-evaluation measures; these included perceptions of high burden on principals' time as a result of the classroom-observation component and incomplete test-score data that limited the number of teachers for whom achievement growth measures could be calculated. In surveys, majorities of teachers indicated that the evaluation measures were valid for the intended purposes, although teachers' opinions about the observation component were more positive than their opinions about the other components. Teachers expressed

some concerns about fairness, but most of them reported that the evaluation system had helped them improve their teaching.

Introduction

High-quality, useful measures of teaching are an essential element of the IP theory of action. To be useful for informing decisions about PD, hiring, or compensation, measures need to support accurate inferences about specific aspects of teachers' effectiveness. Ideally, they should provide information that is sufficiently detailed to allow central-office leaders to determine what course of action will be most likely to improve the quality of teaching in the district or CMO. Each of the IP sites invested significant time and resources to develop multiple-measure systems that would support decisionmaking.

In this chapter, we describe what systems the sites developed to evaluate teachers and determine their effectiveness—the linchpin of the IP initiative. We start by briefly summarizing literature describing the best practices in measuring TE as part of teacher-evaluation systems. Then, we describe the individual components that the IP sites adopted, and we discuss how sites combined these components to create composite scores. We then present distributions of teacher performance on the composite measure for each site to illustrate the proportion of teachers judged to be performing at each level, and we discuss some implementation processes and challenges. The final section discusses teachers' perceptions of the evaluation system, including their opinions about the quality of the measures and the system's effects on their teaching.

Best Practices for Measuring TE as Part of a Teacher-Evaluation System

The IP initiative was designed largely to address the lack of high-quality, informative teacher-evaluation systems in most districts and CMOs. Many systems in the pre-IP era relied on principals' ratings of teachers

and failed to meaningfully differentiate levels of effectiveness among teachers (Weisberg et al., 2009). This lack of variability diminishes those systems' usefulness in informing district decisionmaking and in helping teachers identify their strengths and address their weaknesses. At the same time, research showed the important role teachers play in influencing student learning and the wide variation in the extent to which different teachers contribute to student achievement and other outcomes (Chetty, Friedman, and Rockoff, 2014). In response to the growing awareness of the importance of teaching quality, and supported by federal programs, such as RTT, as of 2015, large majorities of states had adopted teacher-evaluation systems that included measures of student achievement (Doherty and Jacobs, 2015), as did many of the largest school districts in the United States (Steinberg and Donaldson, 2016).

Given the complexity of teaching, combined with the high-stakes nature of many evaluation systems, most scholars and assessment experts have recommended basing teacher evaluation on multiple measures that provide complementary perspectives on various aspects of teachers' work. In particular, even though research on achievement-based indicators of effectiveness, such as measures obtained through VAM, prompted much of the interest in improving teacher-evaluation systems, experts have argued that systems should draw on information from multiple sources when making high-stakes decisions (Glazerman et al., 2010; E. Baker et al., 2010).

One widely used measure that states and districts have incorporated into their teacher-evaluation systems is the direct observation of how teachers teach, which is typically gathered by principals or teachers' peers, multiple times during the year for novice teachers, using a structured rubric (see, e.g., C. Danielson, 2013). These rubrics typically ask observers to rate teachers' instruction on multiple dimensions, and many of them provide composite scores and subscores that provide indicators of performance on specific aspects of instruction. Observation scores are common in teacher evaluation, but researchers have raised some concerns about their quality (Grissom and Loeb, 2017).

The information from observations is generally combined with information from an achievement-based measure. In most cases, this

measure is designed to reflect changes in achievement that are attributable to the teacher rather than simply capturing information about student performance at a single point in time. Two methods are commonly used to estimate student achievement growth attributable to a teacher: VAM (Braun, 2005; Koedel, Mihaly, and Rockoff, 2015) and calculation of SGPs (Betebenner, 2010). The relative strengths and weaknesses of these two approaches have been the subject of debate (Castellano and Ho, 2013; Guarino, Reckase, et al., 2015), but both are intended to provide evidence of a teacher's or a school's contribution to student learning while minimizing the influence of extraneous factors that might influence student achievement but that are not under teachers' direct control. Research acknowledges that neither approach is perfect—that these measures can sometimes produce misleading information because of such factors as peer effects, nonrandom assignment of students to teachers, and limitations in the underlying test scores (Reardon and Raudenbush, 2009). Moreover, estimates can be sensitive to the specific achievement measure used, which can result in an individual teacher being rated as effective (E) based on one test or set of items and as ineffective based on a different test or items (Lockwood, McCaffrey, Hamilton, et al., 2007). However, combining VAMs or SGPs with information from classroom observations can partially mitigate the flaws in both measures.

Observations and achievement growth measures are often supplemented with other indicators, such as survey data gathered from students or parents. The MET project that the Gates Foundation supported showed that a measure of student achievement growth, direct observations of how teachers teach, and student feedback could be combined to create a composite measure with scores that showed reasonable levels of reliability and validity in assessing teachers' future performance on a VAM measure (Kane and Staiger, 2012). The foundation used the MET findings to inform the guidance it provided to the IP sites, and many state and district evaluation systems also adopted combined measures that incorporated student achievement data, observation ratings, and other information.

Research conducted in states and districts that use multiple measures of TE suggest that these measures, particularly the classroom-

observation component, can be useful in promoting teacher reflection and guiding PD (Marsh et al., 2017; Stecher, Garet, Hamilton, et al., 2016; Strunk, Weinsten, and Makkonen, 2014; Taylor and Tyler, 2012). However, the observation component can impose a significant time burden on principals and others who are responsible for conducting the observations (Donaldson and Cobb, 2015); thus, it might be necessary for schools to adjust schedules or staffing to accommodate this extra responsibility. Principals and other observers also need training to enable them to use the observation rubrics consistently and in a way that differentiates among teachers at different levels of effectiveness (Bell et al., 2012). An additional consideration is the need to engage teachers and other stakeholders in designing the evaluation system. The extent to which teachers have confidence in the validity and utility of the measures and the appropriateness of the consequences attached to them will likely determine how sustainable the system is (Goldrick, 2002).

IP Sites' TE Measures

The IP initiative is multifaceted, but all the elements of the reform depend first and foremost on a high-quality measure of TE—something reflected in the theory of action model shown in Figure 1.1 in Chapter One. The other levers shown in that figure can be implemented effectively only if the TE measure provides accurate information about teaching quality and is broadly implemented across the district or CMO. The guidance that sites received from the Gates Foundation indicated that they should develop a measure that included "growth in student learning over time, teachers' knowledge and skill, observed teaching practices, and student perceptions and levels of effort in the classroom" (Bill & Melinda Gates Foundation, 2009, p. 3).

Each IP site adopted a composite measure that combined several data sources. Tables 3.1 and 3.2 summarize the features of the TE

Table 3.1
TE Measures in the Districts Through 2015–2016

Policy Area	HCPS	PPS	SCS
Evaluation system prior to initiative			
Teacher practice measure	Formal school administrator observations every three years, with informal classroom visits annually. Rubric based on the FEAP. Four rating levels (outstanding, S, NI, and U) used for tenure and dismissal decisions	Annual principal rating of S or U, in use until the fall of 2010	Principal rating on Tennessee Comprehensive Assessment Form, in use until July 2011. Apprentice teachers (three years or less teaching) evaluated yearly; all others evaluated every five years
Student achievement	State and local test data used to calculate MAP bonuses (provided to top 25% of teachers by this metric)	None	None
Student feedback	None	None	None
Weight	None	None	None

Table 3.1—Continued

Policy Area	HCPS	PPS	SCS
Component weight during initiative			
Teachers in tested grades and subjects	**2010–2011 through 2011–2012:** 40% student performance as measured by VAM; 60% teacher practice (30% school administrator observations; 30% peer evaluator or swap mentor observations)[a]	**2009–2010 through 2012–2013:** 100% teacher practice	**2011–2012 through 2012–2013:** 40% teacher practice; 35% student achievement growth; 15% student achievement on teacher-selected measure; 5% student feedback; 5% teacher content knowledge
	2012–2013 through 2015–2016: 40% student performance as measured by VAM; 60% teacher practice (35.1% school administrator observations; 24.9% peer evaluator or swap mentor observations)	**2013–2014 and ongoing:** 50% teacher practice; 30% individual student achievement growth; 15% student feedback; 5% school student achievement growth	**2013–2014 and ongoing:** 40% teacher practice; 35% student achievement growth; 15% student achievement on teacher selected measure; 5% student feedback; 5% professionalism

Table 3.1—Continued

Policy Area	HCPS	PPS	SCS
Teachers in nontested grades or subjects	N/A (all grades and subjects are tested in HCPS)	**2009–2010 through 2012–2013:** 100% teacher practice	**2011–2012 through 2012–2013:** 60% teacher practice; 20% teacher-level student achievement growth; 10% school-level student achievement growth; 5% student feedback; 5% teacher content knowledge
		2013–2014: 50% teacher practice; 30% student growth as measured by rubric component 3f;[b] 15% student feedback; 5% school-level student achievement growth	**2013–2014 and ongoing:** 60% teacher practice; 20% teacher-level student achievement growth; 10% school-level student achievement growth; 5% student feedback; 5% professionalism
		2014–2015 and ongoing: 50% teacher practice; 30% SLOs; 15% student feedback; 5% school-level student achievement growth	
Student achievement measure			
Teachers in tested grades and subjects	**Achievement growth, 2010–2016:** VAM based on Florida state test and local test results. Calculated based on a three-year average	**Student achievement growth, 2013–2014 and ongoing:** Individual: three-year VAM; school: two-year VAM using a variety of state and local tests	**Student achievement growth, 2011–2012 and ongoing:** Teacher-level VAM (TVAAS) or yearly teacher portfolio
			Student achievement, 2011–2012 and ongoing: Teachers select yearly from a menu of state-approved options.

Table 3.1—Continued

Policy Area	HCPS	PPS	SCS
Teachers in nontested grades and subjects	N/A (all grades and subjects are tested in HCPS)	**Student achievement growth, 2013–2014:** RISE component 3f; school: two-year VAM	**Student achievement growth, 2011–2012 and ongoing:** School-level one-year VAM (TVAAS)
		2014–2015 and ongoing: Teacher-level annual student learning objectives; school: two-year VAM	**Student achievement, 2011–2012 and ongoing:** Teachers select annually from a menu of state-approved options.
Teacher practice measure	Rubric with 22 weighted indicators based on the Danielson FFT and aligned to the FEAP. Teachers rated on a 1–4 scale on each indicator	RISE rubric, based on the Danielson FFT	TEM rubric (based on Washington, D.C., IMPACT)
When	2010 and ongoing	Piloted 2009–2010, sole TE measure 2010–2013, part of composite measure 2013 and ongoing	2011 and ongoing
Who	All teachers	All teachers	All teachers

Table 3.1—Continued

Policy Area	HCPS	PPS	SCS
Frequency	Two to five formal observations and one to five informal observations, depending on seniority and prior year's observation score	P or D pretenure teachers: Five to seven formal and informal observations per year; at least one announced formal per semester. P or D tenured teachers: Five to seven formal and informal observations per year, including at least one announced formal; observations every two to three years, depending on principal preference. Teachers at F or NI level: Up to 15 formal and informal observations; minimum of two formals per semester (pretenure teachers) or two formals per year (tenured teachers)	New teachers: One announced and three unannounced observations per year. Teachers with prior-year TE scores of 1 or 2: One announced and two unannounced per year. Teachers with prior-year TE scores of 3, 4, or 5: One unannounced, one announced per year
Observers	School administrators, peer evaluators, and swap mentors	Principals, APs, ITL2s, and some central-office staff	Principals, APs, some building coaches (e.g., instructional facilitators), some central-office staff
Certification and calibration of observers in most recent year	All observers certified annually by an external calibrator	Principals participate in periodic calibration conversations and individual activities with their supervisors and peers.	Every observer must rate practice videos within 1 point of the master rater on an annual basis.

Table 3.1—Continued

Policy Area	HCPS	PPS	SCS
Student feedback measure	Not used in HCPS	Tripod survey	Tripod survey
When		Piloted 2011–2013, part of composite 2013 and ongoing	2011–2015 full survey, 2015 and ongoing shorter survey
Frequency		Twice per year	Twice per year
Other measures	None	None	
What			Teacher content knowledge
When			2011–2013
Frequency			Yearly
What			SCS-developed professionalism rubric
When			2013 and ongoing
Frequency			Yearly
Year composite score results first reported to teachers	September 2011 (2010–2011 data)	August 2013 (2012–2013 data)	Fall 2012 (2011–2012 data)

Table 3.1—Continued

Policy Area	HCPS	PPS	SCS
Year composite score results first used as part of teacher evaluation	2010–2011 (applied retroactively)	June 2014 (2013–2014 data)	Fall 2012 (2011–2012 data)
Levels of effectiveness	**Five levels:** U, NI, E, HE 4, HE 5	**Four levels:** F, NI, P, D	**Five levels:** Significantly below expectations, below expectations, meeting expectations, above expectations, significantly above expectations

SOURCES: Annual interviews with central-office staff in each of the IP districts.

NOTE: P = proficient. D = distinguished. SLO = student learning objective. ITL2 = instructor teacher-leader 2 (a CL role; see Chapter Eight for more information). AP = assistant principal.

[a] Peer evaluators performed classroom observations and provided ratings and feedback to veteran teachers, while swap mentors provided the same service for novice teachers.

[b] *Component 3f* refers to RISE component 3f, which PPS added. In 2013–2014, PPS removed this component from its RISE rubric and used it as a measure of student growth in its composite TE measure.

Table 3.2
TE Measures in the CMOs Through 2015–2016

Policy Area	Alliance	Aspire	Green Dot	PUC
Evaluation system prior to initiative				
Practice measure	Annual teacher rating on the California Teaching Standards rubric, conducted by an SL	Annually, multiple informal observations conducted by an SL lead to teacher rating on Aspire-developed rubric (not research-based).	Semiannual teacher rating on the California Teaching Standards rubric, conducted by an SL, plus, for new teachers, submission of a unit portfolio; for veteran teachers, conduct a project	Annually, multiple informal observations conducted by an SL lead to teacher rating on PUC–developed rubric (not research-based); includes student work analysis
Stakeholder feedback	Family and student surveys	Family and student surveys	Annual student survey; semiannual family survey	No information
Student achievement measure	Overall school performance—based on the California state test	Aspire-wide benchmark assessments, California state test data	No	Quarterly benchmark assessments, California state test data, CAHSEE (HS only)
Weight	Information not available	Information not available	Information not available	Information not available

Table 3.2—Continued

Policy Area	Alliance	Aspire	Green Dot	PUC
Component weight during initiative				
Teachers in tested grades and subjects	**2011–2012 through 2012–2013:** 40% achievement growth (30% individual teacher and 10% school-wide); 40% teaching practice; 10% student survey; 10% family survey	**2011–2012 through 2012–2013:** 40% achievement growth (30% individual teacher, 10% school-wide); 40% teaching practice; 10% student survey; 5% parent survey; 5% peer survey	**2011–2012 through 2012–2013:** 40% achievement growth (30% individual teacher, 10% school-wide); 40% teaching practice; 10% student survey; 5% parent survey; 5% peer survey	**2011–2012 through 2012–2013:** 40% achievement growth (30% individual teacher, 10% school-wide); 44% teaching practice; 10% student survey; 3% family survey; 3% peer survey
	2015–2016: 25% achievement growth; 55% teaching practice; 10% student survey; 10% family survey	**2015–2016:** 30% achievement growth (20% individual teacher, 10% school-wide); 50% teaching practice; 10% student feedback; 5% family survey; 5% peer survey	**2014–2016:** 65% teaching practice; 15% student feedback; 5% parent survey; 15% peer survey	**2014 and ongoing:** No composite measure

Table 3.2—Continued

Policy Area	Alliance	Aspire	Green Dot	PUC
Teachers in nontested grades or subjects	**2011–2012 through 2012–2013:** 30% achievement growth (25% ELA SGP and 5% math SGP of teacher's students); 50% practice; 10% student survey; 10% family survey	**2011–2012 through 2012–2013:** 40% achievement growth (school-wide); 40% observation; 10% student survey; 5% family survey; 5% peer survey	**2012–2013 through 2013–2014:** 25% achievement growth (school-wide); 55% practice; 10% student survey; 5% parent survey; 5% peer survey	**2011–2012 through 2013–2014:** 40% achievement growth (school-wide); 44% practice; 10% student survey; 3% family survey; 3% peer survey
	2013–2014 through 2015–2016: [all students take Achieve3000 ELA test] 25% Achieve3000 ELA achievement growth (Lexile score) of teacher's students; 55% practice; 10% student survey; 10% family survey	**2013–2014 through 2015–2016:** 30% achievement (school-wide); 50% practice; 10% student survey; 5% family survey; 5% peer survey	**2014–2016:** 65%, practice; 15% student survey; 5% parent survey; 15% peer survey	**2014 and ongoing:** No composite measure
Student achievement growth measure				
	Piloted 2010–2011; full implementation 2011–2012 through 2012–2013: California Standards Test SGP relative to that of LAUSD	**Piloted 2010–2011; full implementation 2011–2012 through 2012–2013:** California Standards Test SGP relative to that of LAUSD	**Piloted 2010–2011; full implementation 2011–2012 through 2012–2013:** California Standards Test SGP relative to that of LAUSD	**Piloted 2010–2011; full implementation 2011–2012 through 2012–2013:** California Standards Test SGP relative to that of LAUSD

Table 3.2—Continued

Policy Area	Alliance	Aspire	Green Dot	PUC
	2013–2014 and ongoing: Achieve3000 ELA achievement growth (Lexile score)	**2013–2014:** California Standards Test (2012–2013 version)	**2013–2014 and ongoing:** None	**2013–2014 and ongoing:** Lexile growth score
		2014–2015: Grades K–5: Star Renaissance test (Renaissance Learning); grades 6–11: ACT Aspire		
	2015–2016: Teachers can elect to use pilot of ACT Aspire for 10% of student achievement score.	**2015–2016:** Grades 3–8, 11: Smarter Balanced assessment; grades K–2: Star Renaissance test; grades 9–10: ACT Aspire		
Teacher practice measure	TCRP teaching framework (based on Danielson's Framework for Teaching)	TCRP teaching framework (based on Danielson's Framework for Teaching)	TCRP teaching framework (based on Danielson's Framework for Teaching)	TCRP teaching framework (based on Danielson's Framework for Teaching)
When	Piloted 2010–2011; full implementation 2011 and ongoing	Piloted 2010–2011; full implementation 2011 and ongoing	Piloted 2010–2011; full implementation 2011 and ongoing	Piloted 2010–2011; full implementation 2011 and ongoing
Who	All teachers	All teachers	All teachers	All teachers

Table 3.2—Continued

Policy Area	Alliance	Aspire	Green Dot	PUC
Frequency	Two formal and two to four informal observations per year	Two formal observations and three to four informal	One formal and two informal observations per semester	**2011–2012 through 2013–2014:** One formal observation and minimum of four informal; **2014–2015 and ongoing:** Two formal and at least two informal observations focusing on three to five teacher goals from the framework
Observers	Principal or AP	Principal, AP, area superintendent, or director of secondary schools: **2013–2014 through 2014–2015:** One formal, three mini-observations. **2015–2016:** Option of no formal and six mini-observations (adopted by 46% of schools)	Principal or AP: **2015–2016:** One formal and two informal observations per school year	Principals
Certification and calibration of observers[a]	Annually	Annually	If certified, biennially, otherwise annually	Annually
Who	All evaluation observers (about 10% or less of SLs are not certified each year)	All evaluation observers (only one or two SLs are not certified)	All evaluation observers (about 10% or less of SLs are not certified each year)	All evaluation observers (about 10% or less of SLs are not certified each year)

Table 3.2—Continued

Policy Area	Alliance	Aspire	Green Dot	PUC
Student feedback measure	Modified Tripod survey	Modified Tripod survey	Modified Tripod survey	Modified Tripod survey
When	**2011–2012:** Original modified version; **2012–2016:** Shortened and aligned with observation rubric	**2011–2012:** Original modified version; **2012–2013 and ongoing:** Shortened and extensively revised to use language aligned to the observation rubric	**2011–2012:** Original modified version; **2013–2014:** Shortened and revised	**2011–2012 and ongoing:** Original modified version
Frequency	Annual	Annual	**2011–2012 through 2012–2013:** Semiannual; **2013–2014 and ongoing:** Annual	Annual
Other measures				
What	Tripod family survey, aggregated to school level; response rate about 80%	Tripod family survey; teacher level for K–5, school level for 6–12; response rate about 40%	Tripod family survey, aggregated to school level; response rate 30–40%	Tripod family survey, aggregated to school level; response rate about 50%
When	**2011–2012:** Original modified version; **2012–2013 and ongoing:** Shortened and aligned with rubric	**2011–2012:** Original modified version; **2012–2013 through 2013–2014:** Shortened and aligned with rubric. **2014–2016:** Aspire-developed survey	**2011–2012**: Original modified version; **2012–2013 and ongoing:** Revised and shortened	**2011–2012:** Original modified version; **2012–2013 and ongoing:** Revised and shortened

Table 3.2—Continued

Policy Area	Alliance	Aspire	Green Dot	PUC
Frequency	Annual	Annual	Annual	Annual
What		360 peer survey	360 peer survey	360 peer survey
When		**2011–2012:** Original version; revised 2012–2016	**2013–2014 and ongoing:** Based on observation rubric	Survey 2011–2013
Frequency		Annual	Annual	Annual
Year evaluation results first reported to teachers	2010–2011 (pilot results)	2011–2012	2011–2012	2011–2012
Year evaluation results first used for teacher evaluation	2011–2012	2011–2012	2011–2012 (pilot year)	2011–2012
Levels of effectiveness	Five levels: entering, achieving, E, HE, master	Five levels: entering, emerging, E, HE, master	Five levels: entry, emerging, E, HE, HE II	Five levels: emerging, progressing, progressing+, HE, exemplary

SOURCES: Annual interviews with central-office staff in each of the IP CMOs.

NOTE: LAUSD = Los Angeles Unified School District. CAHSEE = California High School Exit Examination.

[a] Administrators who are not fully certified cannot conduct evaluation observations. They receive additional training until they become fully certified. Administrators who are partially certified conduct observations in the company of fully certified observers.

[b] After 2012, Alliance and Aspire changed the labels of the bottom two levels of effectiveness. The table reports the labels used beginning in 2013.

measures developed by the districts and CMOs, respectively.[1] In this section, we discuss some of the key themes that we observed across the seven sites' TE measures. Appendix B presents additional details about each site's effectiveness measure.

The development of the TE measures helped the sites create a common language for discussing effective teaching. One benefit the new measures had was the creation of a common language to describe effective teaching. We heard repeatedly from teachers and SLs during interviews that the development of the observation rubrics, in particular, helped sites build a shared vocabulary for talking about effective teaching. For example, in November 2016, a central-office administrator for Aspire said that one of the benefits was "the adoption of a singular instructional rubric which led to a common vocabulary . . . when it came to effective practice." We heard similar comments from teachers and SLs in all the sites.

Prior to launching the IP initiative, all but two sites already had multiple-measure evaluation systems. As shown in the first row of Table 3.1, PPS and SCS relied solely on principal ratings to evaluate teachers in the years before the IP initiative began. The other sites used multiple measures to assess TE, with HCPS and three of the CMOs including a measure of student achievement. Three CMOs included student and family feedback surveys in their preinitiative evaluation systems. These cross-site differences in prior evaluation systems might have influenced sites' approaches to measuring TE under the IP reform.

As part of the IP initiative, all the sites developed composite measures that assigned the most weight to student achievement growth and classroom observations; additional components varied across sites. Every site included a measure of student achievement growth that counted for 25 to 50 percent of the total evaluation score. The districts used VAM and the CMOs used SGPs to calculate a teacher's contribution to student achievement growth. Weights

[1] In this chapter and the several subsequent chapters that focus on the IP levers, we provide separate implementation tables for these two groups—the districts and the CMOs—because there are often systematic differences between the two groups in how they implemented aspects of the reform and because having one large table for all seven sites would be unwieldy.

assigned to observation-based measures of teacher practice were generally in the range of 40 to 60 percent. One challenge was that not all teachers were in grades or subjects that had student achievement test data that could be used to calculate a growth measure. HCPS developed assessments for all subjects and grades that the state did not test that could be used to calculate a VAM score for all teachers, but all the other sites had at least some teachers (and often a majority) for whom it was impossible to calculate individual test-based achievement growth measures. In those cases, sites sometimes assigned school-level average scores to those teachers. Some sites adopted alternative approaches to measuring student growth, such as SLOs, which consisted of specific learning goals set by the teacher or principal and an approach to measuring students' progress toward those goals (Reform Support Network, undated).

Although sites designed and implemented their observation systems differently, most used Danielson's FFT as a starting point. All but one of the sites developed rubrics based on the Danielson framework, which meant that these sites emphasized a constructivist approach to pedagogy that involves high levels of student engagement and communication (Danielson Group, undated). The exception was SCS, which based its rubric on the Washington, D.C., IMPACT rubric (District of Columbia Public Schools, undated). The Danielson Framework is content neutral, while the DC IMPACT rubric is subject-specific. The sites modified the original rubrics to meet local needs and priorities. Teachers typically received one or two formal observations per year, but, in some cases, they received up to five. In all the sites, principals or other school administrators, such as APs, conducted the observations; in some sites—most notably, HCPS (see next paragraph)—some or all teachers had other observers as well. Additional informal observations occurred in some sites; typically, these informal observations were not scheduled in advance, were shorter in duration, focused on a subset of indicators, and did not include pre- and postobservation conferences. The number of observations sometimes differed depending on whether a teacher was pretenure or tenured.

HCPS was the only site that used peer observers. In HCPS, teachers were observed not only by school administrators but also by a

special cadre of staff called *peer observers*.[2] Peer observer was a full-time, nonteaching, non–school-based position created as part of the IP initiative in 2010–2011. There were two types of peer observers: peer evaluators and swap mentors. Peer evaluators performed classroom observations and provided ratings and feedback to veteran teachers, while swap mentors provided the same service for novice teachers (defined in HCPS as teachers in their first two years of teaching).[3] Novice teachers received three school administrator observations and three swap mentor observations. For more-senior teachers, the number of observations was based on the prior year's classroom practice score, ranging from two school administrator observations and one peer evaluator observation to three school administrator observations and five peer evaluator observations (including both formal and informal observations). Peer evaluators and swap mentors were originally assigned to two-year appointments; however, HCPS consistently extended the appointment for all interested peer evaluators and mentors throughout the IP grant.

Most sites used feedback from student surveys as an additional component, but it was not weighted heavily. Observations and measures of student achievement growth made up the bulk of the weight assigned to components of the overall effectiveness measure. Most sites incorporated additional components, but these tended to receive much less weight. A common additional component was student feedback surveys, which were included in the evaluation systems of six of the seven sites (with HCPS the one exception). These six sites all used the Tripod survey, sometimes with modifications, and assigned weights ranging from 5 to 15 percent to this measure.

---

[2] The peer evaluator component was discontinued in 2016–2017.

[3] Each novice teacher was assigned a mentor who provided advice and support. The mentors also served as observers for novice teachers who were not their mentees; in the observer role, the mentors were called *swap mentors*. One teacher's mentor served as a swap mentor for a different novice teacher (hence, the term *swap*); in other words, the same cadre of staff served as both swap mentors and mentors but played each role for a different set of novice teachers. For a more detailed description of the mentor position, see Chapter Eight.

Distribution of TE Ratings over Time

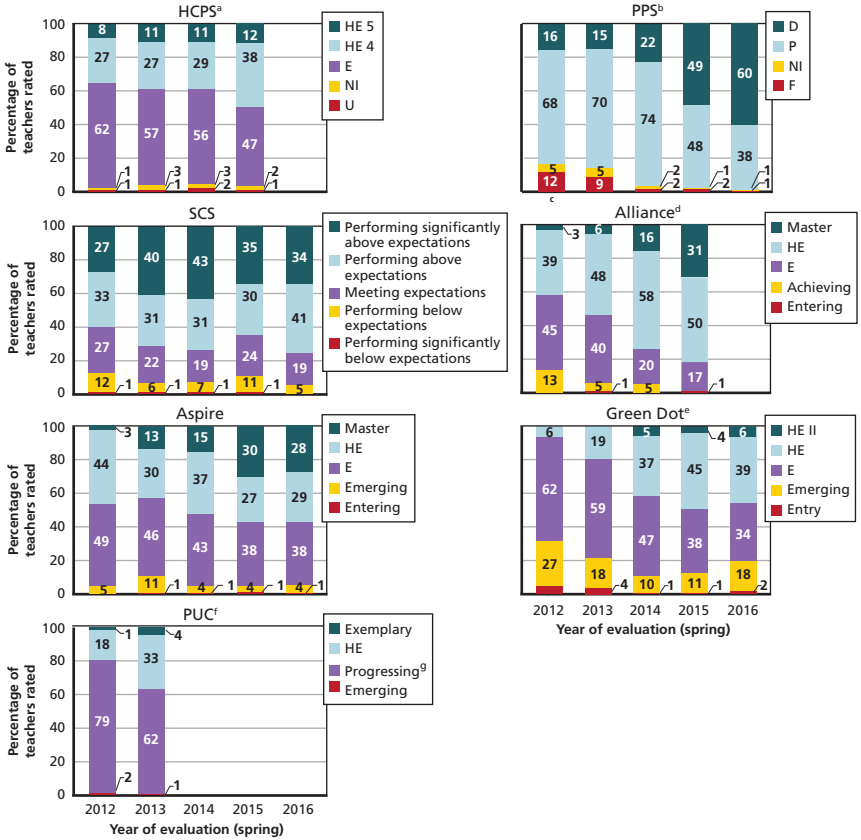
Given the features of the sites' TE measures, we now turn to assessing the distributions of the ratings based on the TE measures. The utility of the TE measures depends largely on the extent to which they accurately differentiate among teachers who are performing at different levels. Even the most thoughtfully designed system will fail to be useful if all teachers receive the same scores. Figure 3.1 shows the distributions of scores on the composite TE measure for each site between 2012 and 2016.

In every site except PPS in 2012 and 2013, no more than 5 percent of teachers received the lowest TE rating, and typically the percentage receiving the lowest rating was closer to 0, 1, or 2 percent (see Figure 3.1). The proportion of teachers in each site's second-to-lowest category also tended to be very small. Of course, we do not know how many teachers *should* have been identified as low performing. The Gates Foundation RFP did not specify a desired distribution, but it did say that the grant recipients should ensure that "only those teachers who meet a defined threshold of effectiveness" should earn tenure or be retained. We reviewed the proposals submitted by the seven sites and found that only HCPS specified the expected percentage of low-performing teachers: 15 percent, which was far higher than the percentage actually identified (as shown in Figure 3.1).

With the exception of PPS, there was little controversy about the cut points sites set for determining performance levels on the TE metric. As part of the implementation process, most of the sites set the cut points (as well as the component weights) that determined how many teachers would fall into each rating category (Nuttall, 2013).[4] To our knowledge, the process for setting the cut points—and, therefore, for determining how many teachers would be classified as low performing and therefore be at risk for nonrenewal or dismissal—was controversial only in PPS, which, as previously noted and as shown in Figure 3.1, is also the only site that classified a relatively large proportion of teachers in the lowest rating category.

[4] SCS was the exception; the district adopted the cut points set by the state.

Figure 3.1
Effectiveness Rating Distributions, 2012–2016, by Site



[a] 2016 data were not available for HCPS.
[b] PPS teachers in their first three years of teaching receive RISE scores only.
[c] The distribution of ratings shown for PPS teachers in 2011–2012 is based on a pilot version of PPS's composite measure that was never shared with teachers or school leaders. The distribution reflects the cut points and ranges that PPS used in subsequent years.
[d] 2016 data were not available for Alliance.
[e] Percentages for Green Dot are based on teachers in California schools only.
[f] Data beyond 2013 were not available for PUC.
[g] The PUC progressing category combines progressing and progressing+; the site did not distinguish these two ratings in the TE data it provided to us.

RAND *RR2242-3.1*

PPS set its cut points for its four performance levels during the 2012–2013 school year, the first year in which scores were combined to form a composite TE measure. PPS took a data-based approach to setting the cut points, using pilot data from prior years; the district consulted outside experts, as well as the union and teachers, for input. The union objected to the specific approach favored by the district because it classified relatively high percentages of teachers as ineffective. The disagreement became contentious enough to receive coverage in the local media. According to central-office staff, the district adjusted the proposed performance ranges (i.e., lowered the ranges so fewer teachers would be at risk of receiving a low rating) at least once during the negotiations to accommodate union concerns. The resulting cut points produced the distributions shown in Figure 3.1—including, uniquely among the sites, nearly 15 percent of teachers being rated as NI or F in 2012–2013. But importantly, no stakes were attached to the ratings that year. When stakes kicked in the following year, far fewer teachers—only about 3 percent—received the two lowest ratings. PPS central-office staff we interviewed cited multiple possible reasons for this change. Although many staff believed that, in 2013–2014, principals adjusted their ratings of the components over which they had control to put fewer teachers at risk of receiving a low overall rating, some said that the change in distribution might reflect true improvements in teacher performance or changes in the composition of the teacher workforce.

The distributions of TE ratings shifted over time, with more teachers scoring in the highest categories and fewer being assigned to the lowest. Figure 3.1 shows a gradual increase over time in the proportion of teachers rated highly. If these trends reflected true improvement in teachers' performance, this would be encouraging news. However, these trends can, at least in some cases, be attributed to changes in the way sites measured TE or in how they created composite ratings. In PPS, for instance, the introduction of SLOs in the 2014–2015 school year was accompanied by a decision to transform teachers' VAM scores so that they had the same distribution across effectiveness categories as the SLO scores. Because the latter tended to be higher, this decision resulted in an improvement in ratings even when VAM scores

did not improve. In addition, following the process set by the state, teachers developed their own SLOs in the fall and were permitted to adjust them in the middle of the year, with principal approval. PPS staff reported that most teachers received high ratings on the SLOs, a fact most staff we interviewed believed accounted in large part for the large proportion of teachers in the D category in that year. Green Dot's ratings improved substantially in 2014, when the CMO no longer included an SGP score for teachers. Also, particularly as stakes were attached to teachers' ratings, observers might have become less inclined to give teachers low ratings—or more inclined to give teachers the benefit of the doubt when the observer was unsure which of two ratings to assign—therefore inflating the overall ratings and changing the distribution. Principals' reluctance to assign low ratings when high stakes are attached to those ratings has been documented in other studies (Grissom and Loeb, 2017; Kraft and Gilmour, 2017). Consequently, the changes in ratings across sites could reflect some combination of real changes in teaching and other factors. Unfortunately, we do not have the separate component scores to test whether changes in the weighting might explain changes in the overall scores.

Implementation Challenges

As noted earlier, site leaders worked with key stakeholder groups to develop measures that would provide accurate information about TE and that could be implemented feasibly. Sites had to address several challenges as they developed and rolled out their revised teacher-evaluation systems; in this section, we discuss the primary ones.

The sites faced trade-offs between minimizing observation burden and providing adequate information on classroom instruction. Because all sites required principals to observe most, if not all, teachers, principals often struggled to find time not only to conduct the observations but also to meet with teachers afterward, fill out the observation rubrics with ratings, document the evidence, and so forth. In interviews, both principals and teachers noted that sometimes these responsibilities led principals to shift time away from other tasks. Some

teachers we interviewed reported that principals sometimes shortened their planned observation time or were not available to address other concerns and needs that arose. Over time, some of the sites responded to these concerns by allowing other school administrators, such as APs, to conduct some observations; HCPS, PPS, and SCS all implemented this change. Another strategy some sites employed was to alter observation time or frequency. All four CMOs, for instance, adopted (or, in the case of Alliance, were piloting) shorter but more-frequent observations that focused on fewer rubric indicators; PPS and SCS reduced both the time and frequency of formal observations. PPS also reduced the number of RISE components on which teachers were rated and eliminated ratings from informal observations. Although these changes did mitigate the burden of conducting observations, several teachers we interviewed in PPS and SCS indicated that they had concerns about this reduction, believing that the shortened observation time failed to provide a sufficiently comprehensive impression of their teaching. These responses illustrate the difficult trade-offs that sites face when trying to balance a desire for extensive information on teaching with a need to keep the burden manageable.

Test-score data were not available for all teachers, preventing sites from adopting a common achievement or growth measure for all teachers. In all sites but HCPS, the standardized test scores needed to generate a VAM score or SGP existed for only a minority of teachers (primarily mathematics and reading teachers in late elementary and MS grade levels). As discussed above, sites took different approaches to creating measures for teachers who lacked these scores. Regardless of the specific approach, the different ways of measuring achievement growth within sites raised questions about equity. For example, in PPS, teachers, following the state process, were permitted to develop their own SLOs and revise them midyear, with principal approval. Although the district provided teachers with guidelines and examples for developing SLOs, the process was largely teacher-driven, which was not the case for teachers who were evaluated based on VAM. PPS addressed this concern by scaling the VAM scores so that the VAM distribution matched that of the SLO scores. At the beginning of the initiative, PPS had added district-developed tests to increase the number of teach-

ers for whom a VAM score could be calculated. However, the district dropped these new tests after a few years because of concerns expressed by teachers.

The CMOs faced an additional challenge when California changed its state test because that change resulted in an inability to calculate SGPs using state test scores. The CMOs responded by modifying the approach to calculating the measure. All the CMOs except Green Dot continued to include a student growth measure in their composite scores. Aspire administered the previous year's state test again and used the results to calculate an SGP and then switched to several other tests (see Appendix B) to calculate an SGP. Alliance and PUC used Lexile scores to create a fall-to-spring growth measure. Green Dot eliminated the achievement growth component from its composite TE score.

HCPS's experiences in this area were somewhat different from those in the other sites because of the more-widespread testing in that district. Having standardized tests for every classroom addressed the equity issue discussed earlier because all teachers were evaluated based on the test scores of their students. At the same time, multiple central-office interviewees reported that the change from the FCAT to the FSA in 2014–2015, combined with curriculum changes related to the Common Core, led to complexities in calculating VAM and in explaining the change to teachers. As one interviewee noted, "[T]here's this big push to 'how could we calculate VAM [when] we have a new test?' Well, you can—it's a comparative analysis—but that's really difficult to explain to teachers."

The sites' systems for developing and maintaining observer accuracy were limited, threatening validity and teacher buy-in. To ensure accurate observation scores, observers must be trained on the rubrics, not only when they are first introduced but also periodically thereafter. Ongoing training is necessary to ensure that raters continue to apply the rubric accurately and to provide new raters with the necessary experience. Such rigorous training and recalibration did not always occur in the IP sites, and this threatened teacher buy-in and might have reduced the validity of the observation scores. For example, central-office interviewees in PPS and SCS acknowledged that their processes for certifying principals as observers changed during the initiative and

were inconsistently implemented both at the central-office level and by principal supervisors. In addition, although principals who did not perform well on certification tests received additional support, principals who did not meet the certification criterion were still allowed to conduct observations.

In HCPS, teachers' mistrust of the peer evaluators created a challenge for the district. In HCPS, about half of the observations were conducted by teachers who were assigned full time as peer evaluators. Many teachers formed negative opinions about the peer evaluators (see "Teacher and SL Reactions" after this section), and district administrators were well aware of this. In interviews, central-office staff described several possible reasons for the rather poor reception by teachers, including a tendency for veteran teachers to reject critical feedback from peers and a lack of a clear mechanism to remove poorly performing peer evaluators. An additional challenge was that, in the first year of the initiative, there was often a mismatch between the grade level and subject of the peer evaluator and the observed teacher. HCPS revised assignments in subsequent years to improve the match, but negative perceptions lingered. Many central-office staff also expressed a belief that, although the number of poorly performing peer evaluators was small, these few peer evaluators were having an overly negative influence on teachers' opinions about the peer evaluator approach.

Ultimately (after the 2015–2016 year and the end of the initiative), HCPS discontinued the peer evaluator and mentor positions, giving all classroom-observation duties back to school administrators. Teachers who had served as peer evaluators or mentors returned to the classroom full time, took other district-level positions, or became half-time instructional mentors. Instructional mentor is a district-level position supporting both novice and experienced teachers. The primary motivation behind discontinuing the peer evaluators was financial. HCPS decided that it was neither cost-efficient nor in students' best interests to have some of the best teachers working outside the classroom full time after the end of the grant. The unpopularity of peer observations was also a factor; as one central-office administrator put it, "[we knew that] one or two rude people or inappropriate people or people [who]

don't have the right social skills or whatever will kill the whole program, and that happened."

Sites found it challenging to communicate effectively about the teacher-evaluation system and to provide the necessary training to teachers to understand and respond to it. The sites adopted different approaches to helping teachers and other school staff understand the goals and components of the teacher-evaluation system. Some used train-the-trainer models in which one representative from each school was given the information and was responsible for sharing it with the rest of the staff at that school. However, the sites did not maintain these models consistently over the initiative. PPS's train-the-trainer model involved having a small number of teachers from each school (called RISE teams) receive training from the district and then train the other teachers at their schools. Central-office staff indicated that, although this process was an efficient way to share information and promoted buy-in and acculturation into the initiative, it sometimes resulted in miscommunication. CMO principals received training in the observation process, and it was generally the school's responsibility to provide training for teachers through the weekly school-site PD sessions. Such sessions for administrators and teachers focused heavily on the evaluation system in the first few years, and then the focus shifted to the Common Core State Standards. HCPS interviewees reported that the district was successful in getting buy-in to the system from new teachers, and they suggested that the mentor program was responsible for this success. Mentors (as opposed to the peer evaluators who observed experienced teachers) were seen as supportive, not just as evaluators, and the mentoring program was well liked by the teachers who participated in it (i.e., novice teachers).

Nevertheless, multiple central-office interviewees at HCPS expressed a wish that the district had rolled out the new TE system more slowly, doing a small-scale pilot or creating scores without assigning any stakes for a year or so before going live, rather than bringing all teachers into an active evaluation system in a single year. They thought that a pilot or test rollout would have allowed more time for training teachers in the new rubric and preparing them to receive feedback and for getting buy-in from teachers and administrators. Remarks from

personnel at the CMOs, each of which conducted a one-year pilot, support this view. As one teacher remarked, "It was nice to be a pilot school because I think we had more say, and voice, and then we also were able to experience it before it became something that was our reality."

Teacher and SL Reactions to Sites' Evaluation Systems

To examine teachers' perceptions of the sites' evaluation systems, we explore their input, from surveys and interviews, on four broad topics:

- awareness: Did teachers know about the evaluation systems?
- endorsement: Did teachers approve of the evaluation systems?
- fairness: Did teachers think that the evaluation systems were fair?
- perceived effects: Did teachers report that the evaluation systems helped improve their teaching?

Awareness

Teachers were aware of their sites' evaluation systems, including what components were part of their sites' systems. When (and where) teachers were being evaluated, they were definitely aware of it. In nearly every site and nearly every year from 2011 through 2016, nearly 100 percent of teachers reported that their performance as a teacher was being evaluated. The exceptions were consistent with sites' teacher-evaluation policies: In 2011, fewer teachers reported being evaluated in sites where teachers were not necessarily evaluated every year prior to the initiative (e.g., SCS).[5] Another exception is PPS, where, depending on the year, 15 to 30 percent of teachers reported that they were *not* being evaluated that year; this is consistent with PPS's policy

[5] Accordingly, survey results for the 2010–2011 year are omitted from the remainder of this chapter. There was no teacher survey in 2012, so, where we present multiple years of survey results, they usually begin in 2013. In many cases, not only in this chapter but also in several others, for the sake of simplicity, we present results only from 2016 (or, in some cases, 2015) and provide the earlier years' results in an appendix. Surveys were administered in the spring, so (for example) responses from the spring 2016 survey pertain to the 2015–2016 school year.

of not observing every teacher every year (given that some teachers likely conflated being evaluated with being observed). Finally, percentages of teachers reporting being evaluated were somewhat lower in PUC in 2015 and 2016, which is consistent with PUC's discontinuation of an overall evaluation rating after 2014. (However, all PUC teachers were still technically being evaluated, so this indicates some conflation among PUC teachers of being evaluated with receiving an overall rating.)

Similarly, teachers were aware of having received overall evaluation ratings when applicable. Nearly every teacher who reported having been evaluated in the previous year (typically 80 to 95 percent of teachers in every site except PPS in every year from 2013 through 2016) reported having received an overall rating as part of the evaluation.[6] The exceptions were Alliance, Aspire, and PUC in 2013 and PUC in 2015 and 2016.[7]

Moreover, teachers were generally knowledgeable about what components were included in their evaluations. Among teachers who reported being evaluated, nearly 100 percent (in every site and every year) indicated that their evaluation included observations of their teaching. In sites where student feedback was a component (i.e., every site except HCPS), typically about 80 to 90 percent of teachers reported the inclusion of this component (and most teachers in HCPS correctly did *not*). In the CMOs, in which parent feedback was an evaluation component, 80 to 90 percent of teachers indicated awareness of this,[8] while teachers in the three districts correctly did not. In Aspire and

---

[6] Teachers in their first year of teaching in their site were not asked whether they had been evaluated the previous year, so that does not explain why the percentage of teachers reporting being evaluated the previous year was not 100 percent.

[7] The percentages in PUC in 2015 and 2016 were 88 and 79 percent, respectively. Although far from 100 percent, these percentages are high considering that the PUC teachers did not receive overall numeric ratings in those years. However, PUC teachers were told whether they met their growth goals, and they might have been thinking of that as an "overall rating" in answering the survey question.

[8] Green Dot in 2016 (73 percent) and PUC in 2015 and 2016 (59 percent and 68 percent, respectively) were exceptions. In these sites, the parent survey results were aggregated to the school level, so that might explain the lower percentages.

Green Dot,[9] in which feedback from other teachers at the same school was a component, 80 to 90 percent of teachers reported this (and teachers elsewhere correctly did not).

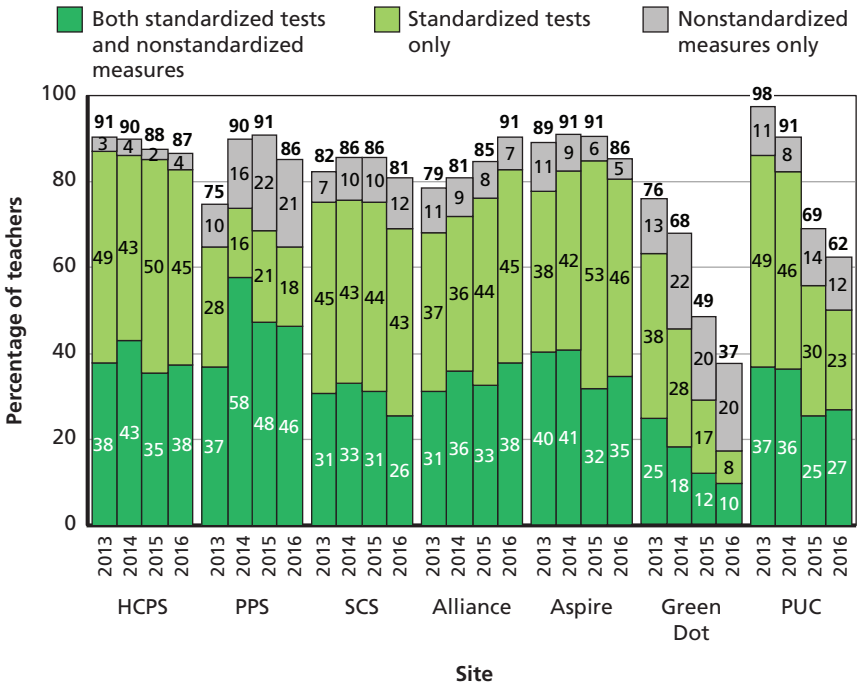
There might have been some confusion about the use of student achievement as an effectiveness measure. The one possible source of confusion about evaluation components, which might have been an artifact of the survey item wording, was about the inclusion of student achievement as an evaluation component. Even though all of the sites incorporated student achievement into all teachers' evaluations in most years, Figure 3.2 shows that, in most sites and years, less than 80 percent of teachers reported that student achievement or growth on standardized tests (the bottom two bar segments) was a component of their evaluation. The percentages were higher in HCPS but still did not exceed 90 percent, which is perhaps surprising given that the district prides itself on having developed (or adopted) a standardized achievement measure for every subject in every grade. The only low percentages that are easily explainable are Green Dot's from 2014 forward, when there were no California state test results and Green Dot elected not to identify a replacement.

In the other sites and years, a teacher of a nontested subject area or grade might not have reported student achievement or growth as a component of his or her evaluation, even if a schoolwide achievement measure was used (as in PPS, SCS, and the CMOs). In addition, some teachers might have been unsure of the distinction between standardized tests and nonstandardized measures, despite the examples given by the respective survey items.[10] When we add in the percentages of teachers who reported nonstandardized measures but not standardized tests (top bar segment), the percentages come closer to 80 or 90 percent in every site except Green Dot, and the addition of this category especially makes a difference in PPS, which employed SLOs, as well

[9] And, correctly, PUC in 2013 only.

[10] For standardized tests, the survey item's exact wording was "Student achievement or growth on state, local, or other standardized tests." For nonstandardized measures, the exact wording was "Student achievement or growth on non-standardized measures (for example, your own assessment, a student work sample, etc.)."

Figure 3.2
Teachers Reporting That Student Achievement or Growth Was a Component of Their Evaluation, Springs 2013–2016



NOTE: Omitted category is "neither standardized tests nor non-standardized measures." Because of rounding, some percentages do not sum precisely.

RAND *RR2242-3.2*

as curriculum-based assessments that were developed in-house (and which teachers thus might not have considered standardized).

Endorsement

Awareness, of course, is only the beginning. For a teacher-evaluation system to realize its full potential, the affected teachers should buy into it and support it. We explore here three aspects of teachers' support for their sites' evaluation systems: perceptions of the validity of the effectiveness measures, perceptions of the accuracy of the effectiveness ratings, and opinions about each of the main components of the evaluation system.

Perceptions of Validity

Teachers were more likely to consider the observation measure as valid than they were to consider other measures—or the combined measure—as valid. As discussed earlier, every site's evaluation system included observations of teaching, and most sites' systems also included measures of student achievement and student input. Figure 3.3 shows the percentages of teachers reporting in the spring of 2016 that each of these three components was, in their opinion, a valid measure of their effectiveness as a teacher. The figure also shows the percentage reporting the combination of all components as a valid measure.[11] As the figure shows, teachers were more likely to report that observations were valid than that the other two components were or, notably, than that the combination of components was.[12] One interpretation of this is that, in teachers' eyes, the combined measure is only as valid as its least-valid component. There is some evidence for this interpretation in PPS, in which, when asked about the validity of the composite measure, many teachers we interviewed commented that, although they thought that the composite measure overall was valid, they had concerns about the VAM. As one teacher put it, "Yes, I do think the combined measure is a valid measure, even though I have the least amount of faith in the VAM score."

Perceptions of Accuracy

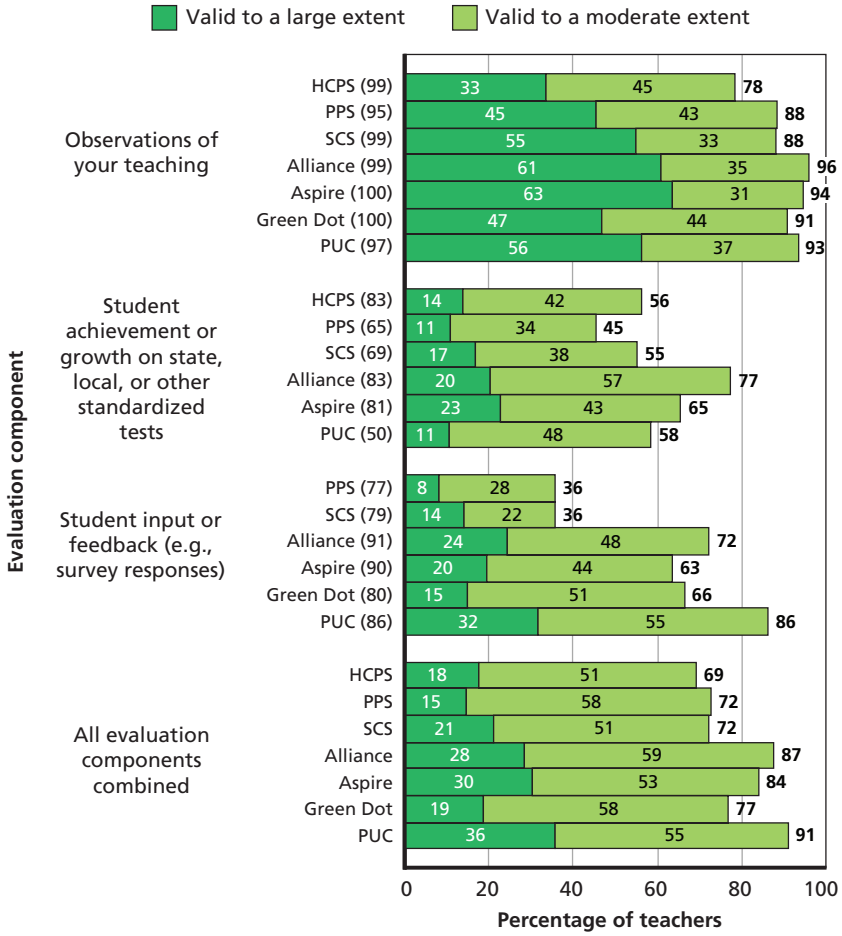
Only about half of teachers thought that the way they were being evaluated accurately reflected their teaching, but more teachers thought that their evaluation ratings were accurate—especially teachers who had received high ratings. Depending on the site and the year, typically about half of teachers agreed with the statement, "The way my teaching is being evaluated accurately reflects the quality of my teaching" (see Figure 3.4). Percentages tended to be lower in HCPS (about 40 percent) and higher in Alliance, Aspire, and PUC (60

---

[11] "All components combined" included all the components that teachers identified as being part of their evaluation, which was not necessarily limited to observations, student achievement or growth, and student input.

[12] Patterns in previous years were similar; see Figure C.1 in Appendix C.

Figure 3.3
Teachers Reporting That Evaluation Components Were Valid Measures of Effectiveness to a Large or Moderate Extent, Spring 2016



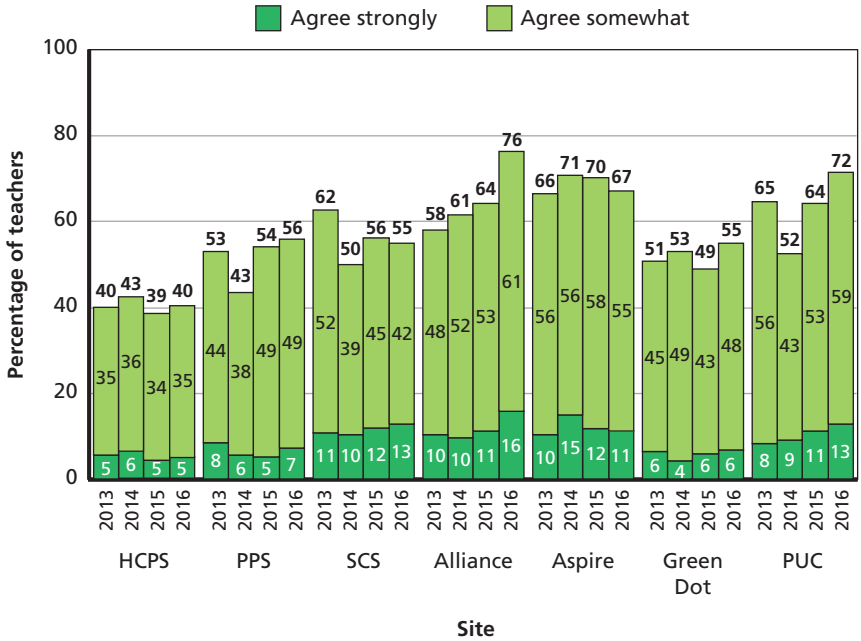
NOTE: Numbers in parentheses next to the site names are the percentages of teachers (among those who reported being evaluated) who indicated that the component was part of their evaluations. Omitted response categories are "valid to a small extent" and "not at all valid." HCPS's TE measure did not include student input, and, in 2016, Green Dot's measure did not include student achievement.

RAND *RR2242-3.3*

to 75 percent in most years). Perhaps not surprisingly, teachers who received higher evaluation ratings were generally more likely to agree with the statement than lower-rated teachers, although the differences were not that dramatic, and even among the higher-rated teachers, less than 70 percent, in any site in 2015 *or* 2016, agreed.

However, when we asked teachers specifically about the accuracy of *their own prior-year TE ratings*, considerably higher percentages—60 to 80 percent—reported that their ratings were at least moderately accurate. On this, the differences by the level of rating received were much starker, with higher-rated teachers being much more likely than lower-rated teachers to say that their ratings were accurate (see Figure 3.5).

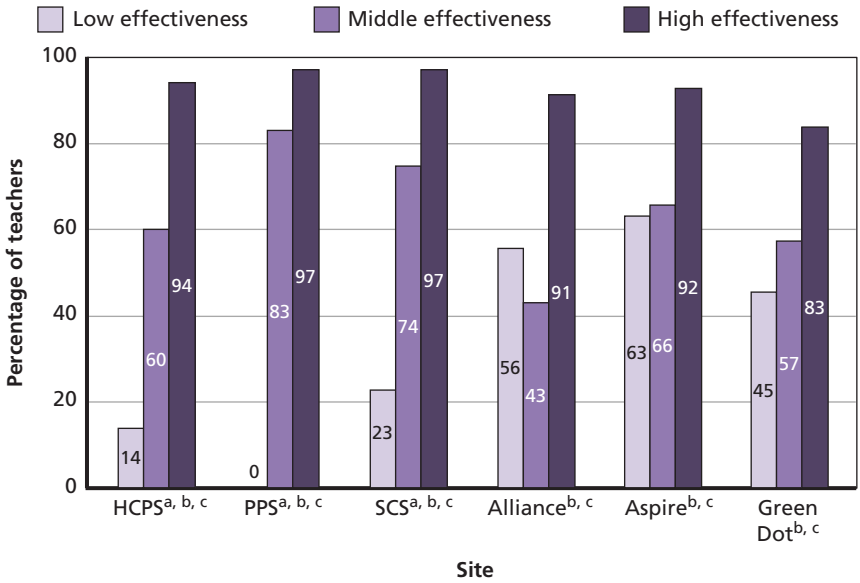
Figure 3.4
Teachers' Agreement That the Way Their Teaching Was Being Evaluated Accurately Reflected the Quality of Their Teaching, Springs 2013–2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly."

RAND *RR2242-3.4*

Figure 3.5
Teachers Reporting That Prior-Year Overall Evaluation Ratings Were Moderately or Highly Accurate, by Prior-Year TE Rating, Spring 2015

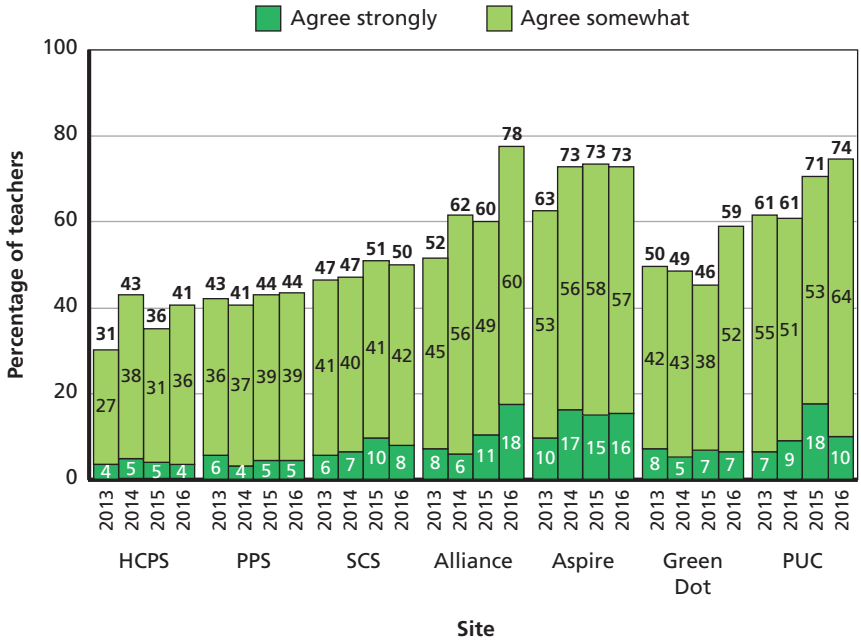


NOTE: We based definitions of *low*, *middle*, and *high* on sites' own categorizations; for details, see Appendix A. Effectiveness ratings were not available for PUC. Results from other years (2013, 2014, and 2016) were similar. The original response categories of "moderately accurate" and "highly accurate" have been combined. Omitted response categories are "slightly accurate" and "not at all accurate."
[a] Low–middle difference is significant at $p < 0.05$.
[b] Low–high difference is significant at $p < 0.05$.
[c] Middle–high difference is significant at $p < 0.05$.

RAND *RR2242-3.5*

A final dimension of accuracy is whether the ratings are viewed as accurately distinguishing among teachers of varying levels of effectiveness. In the three districts, about 40 to 50 percent of teachers agreed that "The teacher evaluation system does a good job distinguishing effective from ineffective teachers." The percentages of CMO teachers agreeing with the statement were somewhat higher (in the 50- to 75-percent range) and tended to increase from 2013 to 2016. (See Figure 3.6).

Figure 3.6
Teachers' Agreement That the Teacher-Evaluation System Did a Good Job Distinguishing Effective from Ineffective Teachers, Springs 2013–2016

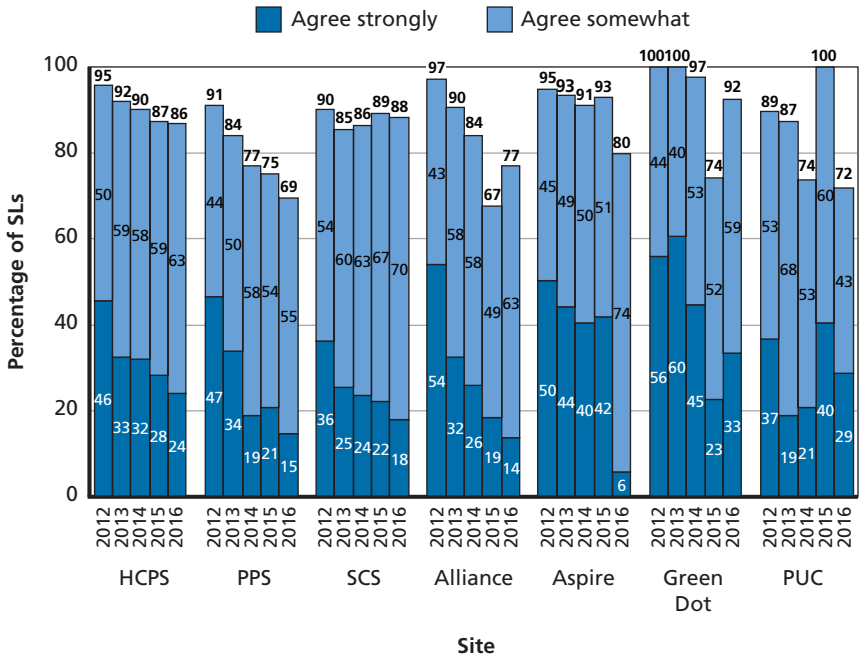


NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly."

RAND *RR2242-3.6*

We also asked SLs, who might be in a better position than individual teachers to judge whether the ratings distinguish appropriately among teachers, whether they agreed with the statement. Although SLs were more likely than teachers to agree with the statement, the percentages of SLs agreeing, particularly agreeing strongly, declined over time in every site (see Figure 3.7). SLs in most sites also became *more* likely over time to agree with the statement, "I don't need teacher evaluations or teacher effectiveness ratings to know who the good and bad teachers are."

Figure 3.7
SLs' Agreement That the Teacher-Evaluation System Did a Good Job Distinguishing Effective from Ineffective Teachers, Springs 2012–2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly."

RAND *RR2242-3.7*

Perceptions of Individual Components of the Evaluation System

As noted earlier, teachers were more likely to indicate that observations were valid than that the student achievement and student input measures were valid. Here, we present some finer-grained reactions to each of these three evaluation components.

Observation Component

Except in HCPS, most teachers held favorable views of the observation rubric, observers, and the number and length of observations. For the observation component, we examine teachers' attitudes on the observation rubric, observer qualifications, the number and length of observations, and observer consistency. We also make

some inferences (based on survey responses) about the "genuineness" of observed lessons.

The Rubric

In most of the sites, a majority of teachers thought that the observation rubric was well-suited for measuring many different forms or styles of good teaching, for measuring instruction in their subject area, and for measuring instruction with the types of students they taught.[13] However, HCPS teachers were less likely than other sites' teachers to agree with these sentiments. (See rows 2, 3, and 4 in Figure C.2 in Appendix C.) In interviews with HCPS teachers, reaction to the rubric was mixed. One teacher shared, "If [the rubric] were more specific as to 'here's what we want in reading,' 'here's what we want in science,' [and] 'here's what we want in math,' I think it could be more helpful." Another teacher had a more favorable view of the rubric:

> When you read the rubric itself, it's very specific as to the difference between this score and this score. . . . It's clear—it's very clear, and there's no debate in it. You can see [that] it specifically says that, if you do this, you'll get that "exemplary." If you do this minus this, you're going to get "accomplished." If you don't follow it at all, you get "requires action." So it's pretty clear and precise.

Observers

In every site except HCPS, more than 75 percent of teachers—in every year from 2013 through 2016—agreed that "The people who observe my teaching are well qualified to evaluate it." In HCPS, about 60 percent of teachers agreed with the statement each year. (See row 5 of Figure C.2 in Appendix C.) About 90 percent of teachers in every site, including HCPS, thought in every year that their principals were at least moderately qualified to evaluate their teaching. But HCPS teach-

[13] Every year from 2013 through 2016, we asked the survey question about the suitability of the observation rubric for measuring many different forms or styles of good teaching. We asked the other two survey questions, about suitability for subject area and types of students, only in 2016.

ers were less likely to say the same about their peer evaluators: Less than 70 percent of HCPS teachers said in any year that their peer evaluators were at least moderately qualified. This suggests that HCPS teachers' relative discontent with observers, and perhaps with the observation component in general, was related to their perceptions of the peer evaluators.[14] Interviews painted a more nuanced picture, one in which teachers and SLs frequently expressed frustration but were not always negative. A common concern raised in these interviews was that peer observers were not school-based and often did not know the school or student population where they were conducting observations. A leader of a high-need school in HCPS commented,

> I believe that some [peer observers] are coming into buildings with a preconceived notion based on the school's outcome as to what those teachers must be. I have seen teachers who were my stellar teachers receiving marks that can, in no way—no way, not even on one instance—be what they actually saw. When I look at what [evaluators] are using to support their markings, it's not even aligned with the rubric. In some cases, it's counter to what the rubric has said.

Number and Length of Observations

Majorities of teachers in every site thought that there were enough observations to provide an accurate view of their teaching. (See row 7 of Figure C.2 in Appendix C.) Percentages agreeing were highest (about 80 to 90 percent) in SCS in all four years and in Aspire in 2015 and 2016 and were lowest in HCPS, Alliance, and Green Dot (about 60 percent agreeing each year). Slightly higher percentages of teachers tended to indicate that observations were long enough to provide an accurate view, but the cross-site patterns of agreement tended to be similar (e.g., SCS with higher percentages and HCPS with lower percentages). (See row 6 of Figure C.2 in Appendix C.) During inter-

[14] In HCPS, novice teachers were more likely than experienced teachers to express favorable views about observer qualifications, likely reflecting the more-positive perceptions in the district of mentors than of peer evaluators.

views, teachers in HCPS expressed skepticism about the accuracy of the observation measure given the limited number and length of observations. As one teacher in 2014 stated, "It's hard for me to think that a 45-minute snapshot . . . depicts accurately my teaching skills." Another HCPS teacher similarly referred to a "snapshot" in a 2016 interview:

> The snapshot of that one hour that you watch me determines how I am for the entire year. . . . What if that day I was really bad or the kids were really bad and the scores were really bad? What about the other 50 days that I did good stuff—nobody saw it, and it didn't count? It's annoying or frustrating that your whole score is based on, like, that one observation or that one hour out of your whole school year.

Observer Consistency

In 2013 and 2015, we asked teachers about the consistency among observers. Among teachers who reported having multiple observers (i.e., nearly everyone in HCPS but fewer in the other sites, especially PPS and the CMOs), more than 70 percent in each site and year thought that the observers were at least somewhat consistent with one another. Although typically less than 50 percent thought that the observers were *very* consistent with one another, "very consistent" was nevertheless almost always a more common response than "not at all consistent." In interviews, however, some teachers mentioned receiving different scores from different observers and expressed concerns about inaccuracy and bias. One PPS teacher, for instance, noted, "I was originally given a lot of 'basics' in this last observation cycle, whereas, in the previous observation cycle, I was all 'proficient' and 'distinguished.'"

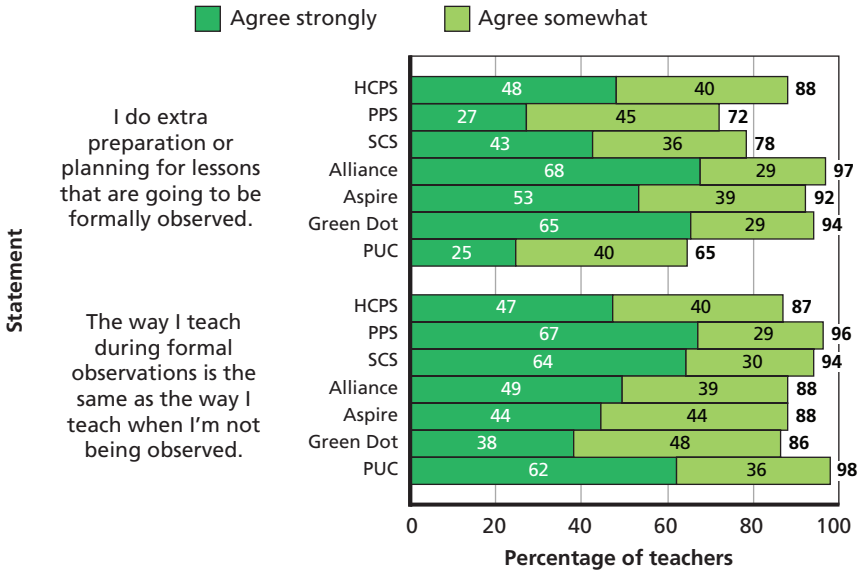
Representativeness of Observed Lessons

In Alliance, Aspire, and Green Dot, majorities of teachers strongly agreed in 2016 that they did "extra preparation or planning" for lessons that were going to be formally observed, while minorities of teachers strongly agreed that the way they taught during formal observations was the same as the way they taught when not being observed. The reverse was true in PPS, SCS, and PUC, with minorities of teachers

agreeing strongly about extra planning for observations and majorities strongly agreeing that their teaching during observations resembled nonobserved teaching. (See Figure 3.8.) This suggests that, for at least some teachers, observed lessons might have been more-genuine reflections of regular practice in PPS, SCS, and PUC than in Alliance, Aspire, and Green Dot. As an Alliance teacher explained,

> The language is about, "If you want a 4, here's what you do." So there's a disconnect between the purpose of . . . doing this, which should be student achievement, and then the compensation, [and] you can't blame people for wanting to get a good score if it's compensation-based. Are we doing this on a regular basis, or are we learning how to do this to get the score that we want on the day of the observation?

Figure 3.8
Teachers' Agreement with Statements About Observations, Spring 2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly."

RAND *RR2242-3.8*

Student Achievement Component

A majority of teachers thought that the student tests used in their evaluations measured important skills and knowledge, but teachers of nontested subjects and grades raised concerns. For the use of student achievement as an evaluation component, we examine teachers' perceptions about test content and alignment, including those held by teachers of nontested subject areas and grade levels. We also touch on teachers' perceptions about how student characteristics relate to the achievement measure.

A majority of teachers (55 to 65 percent in most sites and years) agreed that the student tests used in their evaluations measured important skills and knowledge. There was greater variation across sites and across years in the percentages of teachers agreeing that the student tests were well aligned with their curricula and that the test scores were a good measure of how well students had learned what the teachers had taught. (See rows 2, 3, and 4 of Figure C.3 in Appendix C.) For example, PPS included its locally developed curriculum-based assessments (CBAs) in its VAM; most PPS teachers objected to this because they felt that the tests were inaccurate, poorly designed, and not well aligned to the content that was being taught. As one PPS teacher described it in 2012,

> [U]ntil the CBAs match the curriculum, it won't be worth anything. It's like Test Writing 101. Are you testing what you purport to be teaching? [The CBAs are] not testing what we're supposed to be teaching.

Moreover, this set of survey questions was answered only by teachers who reported teaching a tested subject area and grade level. In interviews, teachers of nontested subjects and grades expressed concerns about how achievement was incorporated into their evaluations. Sample comments include the following:

> For math or science, it isn't really fair to include Lexile as a part of the salary. Some struggling readers do well in math (particularly ELL). (Alliance teacher, 2015)

> It is world history scores in tenth grade being compared to English scores in ninth grade. So, any gains are not based on building history skills or knowledge, necessarily. And the last time students might have taken world history is back in seventh grade. So, I think the growth number is a little arbitrary in terms of how well I teach history. (Aspire teacher, 2014)

Even in HCPS, in which there was nominally a test for every subject in every grade, some teachers expressed concern about how this was operationalized, as exemplified by this comment:

> And even with social studies, they evaluate how they scored. . . . [We] teach world history in sixth grade now. In seventh grade, you have civics. . . . How are you going to compare growth from world history [to] civics when it's two totally different subjects?

Some teachers voiced concerns about whether the achievement measure fully accounted for student characteristics. In most sites and years, only about 40 percent of teachers agreed with the statement that "The ways that student test scores are used to evaluate my performance appropriately adjust for student factors not under my control." In 2015, one Green Dot teacher expressed doubt that the student measures adequately accounted for student characteristics:

> The measures have a role to play, but, if they're going to link that to retention or pay, it has a long way to go, because there's so many variables that go into measuring to what extent a teacher influences students' test scores—outside of school factors. . . . The school is a critical component to student achievement, but it's not the only component.

Many teachers we interviewed in PPS and SCS made similar comments. Other teachers raised questions about the measurement of achievement growth, such as an HCPS teacher who commented in 2014, "It all boils back down to what I have to work with from the beginning of the year." Some teachers expressed concerns about whether the tests could adequately measure growth for higher-achieving students. About 30 to 40 percent of teachers disagreed that the student tests

used in their evaluation had "room at the top for even the [district's or CMO's] highest-achieving students to grow." (See rows 5 and 6 of Figure C.3 in Appendix C.)

Student Input Component

Teachers in the CMOs were more likely than district teachers to voice appreciation for student input. Teachers in the CMOs were more likely than teachers in the two districts that had student input measures to agree that getting input from students is important to assessing TE, that students are good judges of TE, and that students could be trusted to provide honest, accurate feedback. (See Figure 3.9 for 2015; for all years, see Figure C.4 in Appendix C.) One Aspire teacher who was interviewed commented about the student surveys:

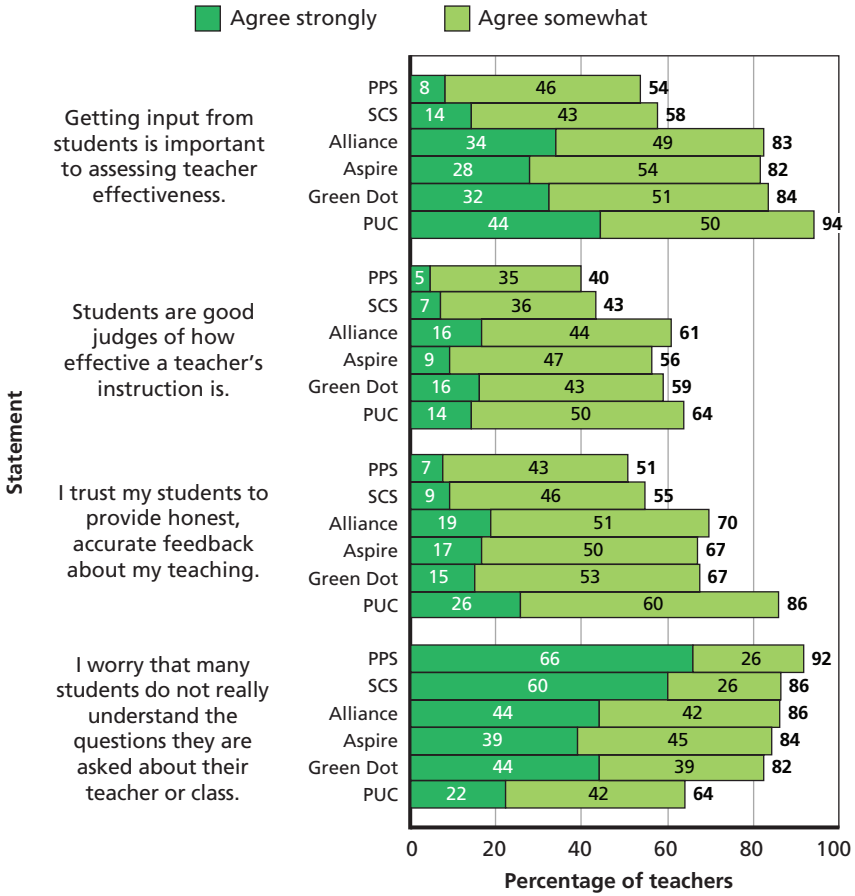
> I love getting that feedback. In some ways, it's the most personal feedback. It reflects how they feel on that day, but it does reflect how students feel and paints a different picture [from what] admin would see in an observation. It's a relevant data point.

Teachers in the two districts, meanwhile, were more likely than teachers in the CMOs to agree, particularly to agree strongly, that "I worry that many students do not really understand the questions they are asked about their teacher or class." One SCS teacher we interviewed expressed the common concern among district teachers that students who struggle with reading might not understand the student survey questions:

> Sometimes, some of the students don't understand [the Tripod questions]. We have students [who] struggle with reading and . . . it affects [teachers'] livelihood if [students] don't understand the questions.

The survey response differences between the districts and the CMOs held even within school levels (elementary, MS, and HS), so they were not a function of a higher proportion of MSs and HSs (and thus older students) in the CMOs. They also were not a function of the CMOs having a higher concentration of newer teachers (who presum-

Figure 3.9
Teachers' Agreement with Statements About the Use of Student Feedback in Teachers' Evaluations, Spring 2015



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly." HCPS's TE measure did not include student input.

RAND *RR2242-3.9*

ably were closer in age to students), although newer teachers in the two districts did tend to respond more favorably than more-experienced teachers about student input.

Fairness

In this section, we examine teachers' perceptions of the fairness of their sites' evaluation systems.

Many teachers did not think that their site's evaluation system was fair to everyone, but most thought that the system had been fair to *them*. In the three districts, less than 40 percent of teachers—in any year from 2013 through 2016—agreed with the statement, "The evaluation system is fair to all teachers, regardless of their personal characteristics or those of the students they teach." Percentages were higher in the CMOs but still did not typically exceed 60 percent. However, teachers were much more likely to agree that the evaluation system had been fair to *them* (as opposed to being fair to all teachers). Not surprisingly, teachers who received higher effectiveness ratings were more likely than lower-rated teachers to agree that the evaluation system had been fair to *them*. Because there were more higher-rated teachers than lower-rated teachers (see Figure 3.1 earlier in this chapter), this pattern of results makes sense: In answering the survey question about fairness to *all* teachers, respondents were likely considering teachers whom they thought had received unfairly low ratings, even if there were relatively few low-rated teachers, and even if they themselves had received a higher rating.

Interview data suggest that many teachers' concerns about fairness stemmed largely from their perceptions about the inaccuracy of the achievement measures. One Alliance teacher noted that the lack of tests in every subject created inequities:

> If there were a test for your specific subject, I think it would be OK to make that a part of the evaluation system. But it is not fair when there is not a test for the subject you are teaching.

Some interviewed teachers also expressed concerns about the other components, including the student surveys (which some teachers thought might elicit biased responses from students) and the observation process. On the latter point, one PPS teacher explained,

> I don't think RISE is very effective. I don't think you can judge a teacher when coming in for 20 minutes to watch him or her.

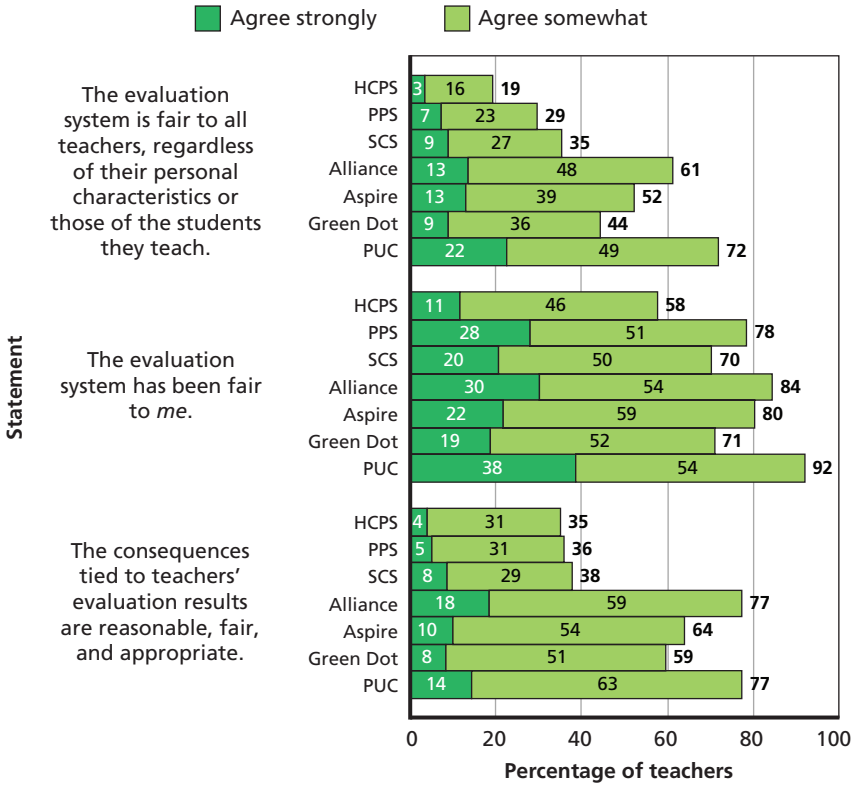
> I think there are too many key points, so that it's too easy to be unfair to teachers, I believe.

Especially in the three districts, teachers had concerns about the fairness of consequences tied to the evaluation system. We also asked teachers whether they thought that the consequences tied to teachers' evaluation results were "reasonable, fair, and appropriate." On this question, there was a very clear divide between teachers in the three districts, in which the percentages agreeing were always below 40 percent, and the CMOs, in which the percentages agreeing were typically above 60 percent (except in Green Dot, in which the percentages were 50 to 60 percent). In Chapters Four through Eight, we discuss some of the consequences that were likely in teachers' minds as they answered this question, such as linking placement, termination, or compensation decisions to effectiveness ratings.

Figure 3.10 shows the percentages of teachers agreeing in 2016 with the three survey statements about fairness. Although, for the sake of simplicity, we do not display the percentages for prior years, there was a general trend upward; that is, in most sites, the percentages agreeing with these statements about the evaluation system's fairness tended to increase from 2013 to 2016. (See the bottom three rows of Figure C.5 in Appendix C.) The increases from 2014 to 2016 are particularly notable for PUC; this is likely related to PUC's shift away from an overall effectiveness measure in that time frame. (See Figure 3.11.)

One fairness concern that many teachers had, particularly in the three districts, was a perception that observers were under pressure to limit the number of high ratings they gave. This concern was most prevalent in HCPS, in which, in every year from 2013 through 2016, 80 to 85 percent of teachers agreed with the statement, "Even if there are many highly effective teachers in a school, there is pressure to only rate a small number of them as very highly effective." In PPS, the percentages agreeing with this statement ranged from 70 to 80 percent, and, in SCS, the range was 60 to 75 percent. In the CMOs, the ranges were lower (typically 45 to 65 percent), especially in PUC in 2015 and 2016, when only about 30 percent of teachers agreed with the state-

Figure 3.10
Teachers' Agreement with Statements About the Fairness of the Evaluation System, Spring 2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly."

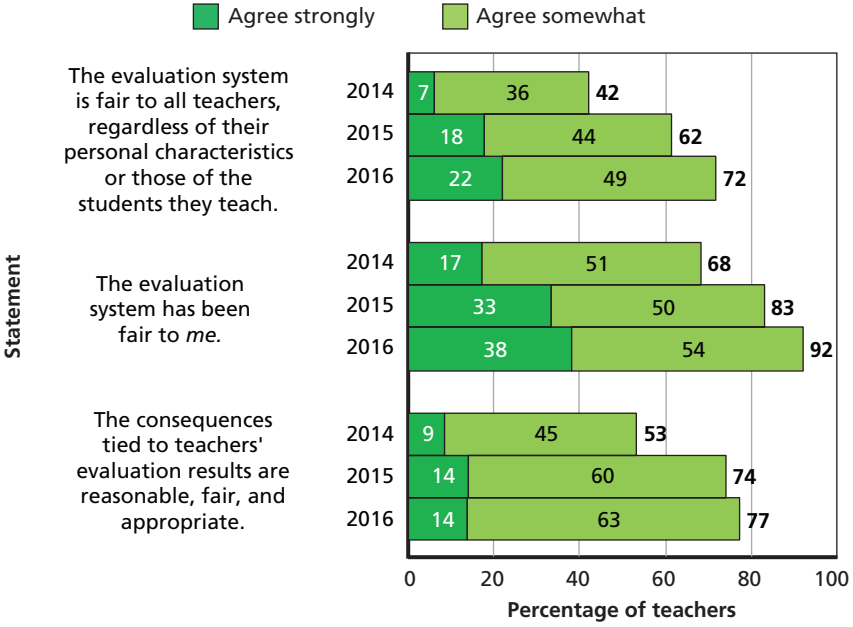
RAND *RR2242-3.10*

ment. To our knowledge, none of the sites had a policy restricting the number or percentage of teachers who could get high ratings.

Perceived Effects

In theory, teacher-evaluation processes are used not only to judge and rate teachers but also to help them improve, and each of the IP sites has teacher improvement as a major goal. Improvement might come about because teachers are motivated by the "carrot and stick" of their

Figure 3.11
PUC Teachers' Agreement with Statements About the Fairness of the Evaluation System, Springs 2014–2016

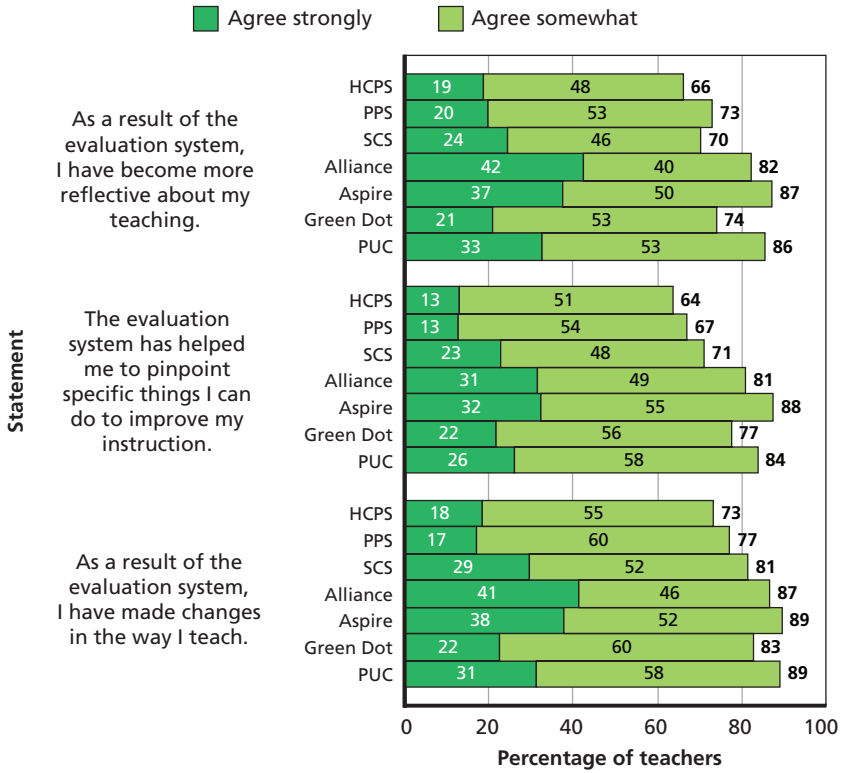


NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly."
RAND *RR2242-3.11*

systems to simply "teach better," but many would argue that improvement is more likely to come as a result of feedback received from others and through a process of self-reflection. In this section, we explore the extent to which teachers perceived their evaluations as useful.

Most teachers thought that the evaluation system had helped them improve their teaching, although some reported negative side effects. Majorities of teachers in all seven sites agreed in 2016 that their site's evaluation system had helped them become more reflective about their teaching, pinpoint specific things they could do to improve their instruction, and make changes in the way they teach (see Figure 3.12).

Figure 3.12
Teachers' Agreement with Statements About the Evaluation System's Effects on Their Teaching, Spring 2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly."

RAND *RR2242-3.12*

Teachers in the CMOs were more likely than teachers in the districts to agree, and teachers in HCPS were the least likely to agree.[15]

Although many teachers reported that the evaluation system had helped them become more reflective and improve their instruction,

[15] In earlier years, differences between the districts and CMOs were not as large. (See the top three rows of Figure C.5 in Appendix C.) Trends over time differed by site and are not easy to describe, but, in general, percentages tended to decrease in the three districts (particularly on the middle item) and increase in the CMOs (particularly on the top item).

there were also reports of some less positive side effects on instruction. In 2016, 50 to 70 percent of teachers in each of the seven sites agreed that, "As a result of the evaluation system, I feel inclined to 'play it safe' with my instruction rather than take risks with new approaches that might not work."[16] As one Green Dot teacher commented in 2016,

> Sometimes I kind of hesitate to experiment if it doesn't adhere exactly to the model of TCRP, especially when I'm being observed. When I'm being observed, I feel I have to definitely follow the rubric, whereas, when I'm not watched as much, I'm more willing to take risks in the things I want to do in the classroom.

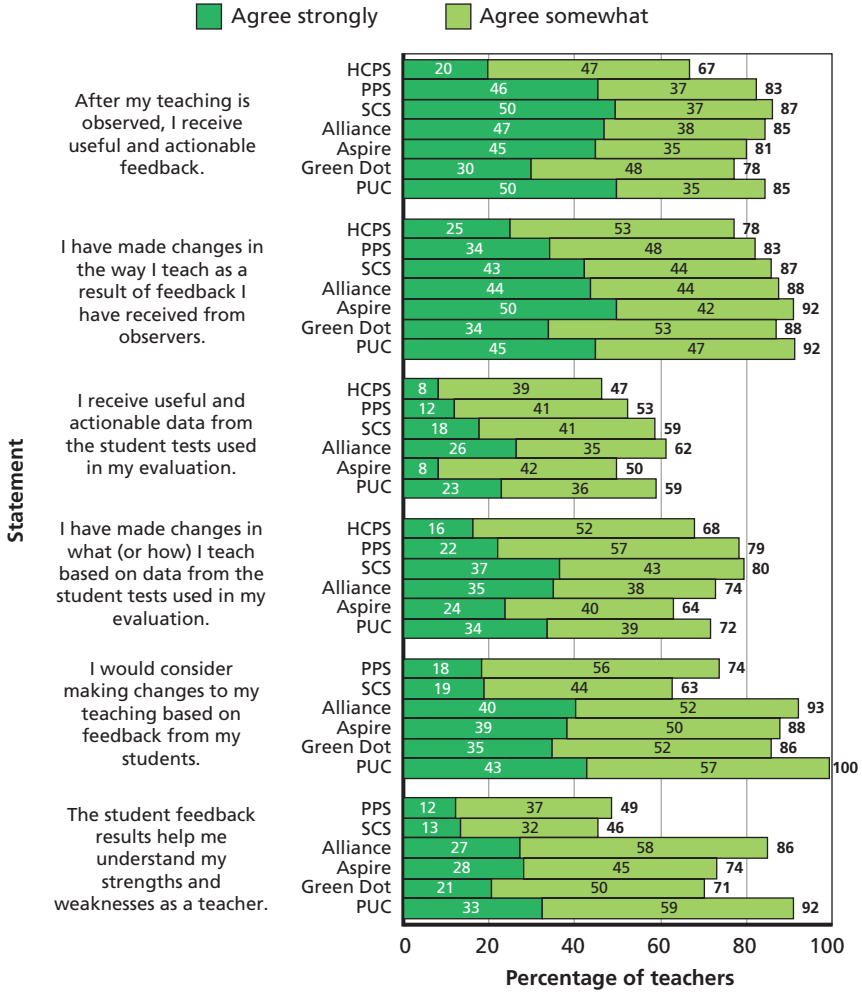
In addition, between one-quarter and one-half of each site's teachers tended to agree (from 2013 through 2016) that "The evaluation system is pushing me to teach in ways I don't think are good for my students."

A major mechanism by which the evaluation system might have spurred improvements in practice was the feedback teachers received on individual components. Figure 3.13 presents 2016 results on teachers' views of the usefulness of the three main evaluation system components (Figure C.6 in Appendix C presents results for all years):

- **observations:** Most teachers (around 80 percent) agreed that they had received useful feedback following observations and that they had made changes to their teaching based on observers' feedback. Agreement was lower in HCPS than in the other sites.
- **test scores:** About half of teachers indicated that the data from the student tests used in their evaluations were useful and actionable. Interestingly, higher percentages (i.e., 65 to 80 percent in 2016) of teachers reported that they had made changes in their teaching based on data from these tests.
- **student input:** Teachers in the CMOs were more likely than teachers in the districts to indicate finding value in student feedback results.

[16] We had not asked this question in previous years.

Figure 3.13
Teachers' Agreement with Statements About the Usefulness of Feedback from Evaluation Components, Spring 2016

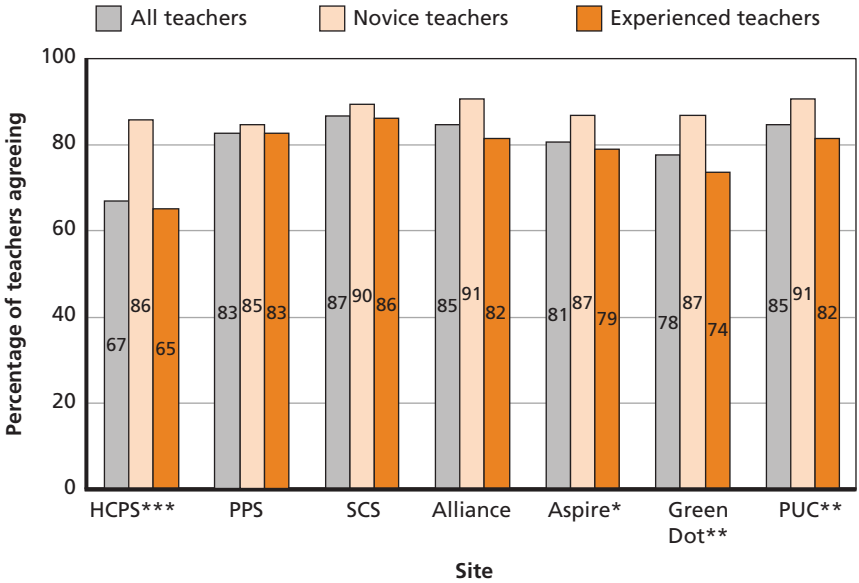


NOTE: We omit Green Dot from the items about test scores because Green Dot did not use test scores in teachers' evaluations in 2016. We omit HCPS from the items about student feedback because student feedback was not an evaluation component in HCPS. Omitted response categories are "disagree somewhat" and "disagree strongly."

RAND *RR2242-3.13*

Novice teachers were typically more likely than experienced teachers to agree with the statements about the usefulness of the components, particularly in HCPS.[17] Figure 3.14 shows an example.

Figure 3.14
Teachers' Agreement That, "After My Teaching Is Observed, I Receive Useful and Actionable Feedback," by Teacher Experience Level, Spring 2016



NOTE: The original response categories of "agree strongly" and "agree somewhat" have been combined. Omitted response categories are "disagree somewhat" and "disagree strongly." * denotes that the difference between novice and experienced percentages is significant at $p < 0.05$. ** denotes that the difference is significant at $p < 0.01$. *** denotes that the difference is significant at $p < 0.001$.

RAND *RR2242-3.14*

[17] Throughout this report, where we discuss differences in survey responses between novice and experienced teachers, we define *novice teacher* as a teacher with two years or less of teaching experience and *experienced teacher* as a teacher with more than two years of teaching experience.

Many teachers whom we interviewed commented on the usefulness of feedback based on observations. Comments included the following:

> RISE is helpful. You're getting observed; the principal gives you good information about what works and what doesn't—even the littlest things. When you do group or partner work in class, which I do a lot, [administrators] tell you when you need to spend a little more time helping the students figure it out before you give the answer. [Administrators] observe and see what you're doing. It helps you tweak what you're doing. (PPS teacher, 2016)

> If someone comes in and gives me a clinical look at how I'm teaching, I think that's good. I look for support from colleagues, but someone trained in the RISE process pushes my growth even more. Even though I'm intrinsically motivated [to improve my teaching practice], receiving feedback at a technical level and with a critical eye has really moved my teaching and my way of thinking about [my practice]. (PPS teacher, 2016)

> I find [feedback] very helpful. Anytime you are observed, it's an opportunity for outside eyes, different ways to think of grouping, opportunities you missed, different ways to think of questions you missed, so the feedback is great. (Green Dot teacher, 2014)

Summary

Research suggests that teachers' contributions to student achievement growth are substantial and that differences in teachers' performance are widespread. High-quality information on teachers' performance could help central-office leaders make better decisions about how to support teachers' professional growth, and such information could also contribute to improved hiring, placement, and compensation policies if it is used appropriately. At the same time, researchers and others have raised concerns about negative consequences of relying too heavily on test-based measures of TE. Research suggests that the best approach to measuring TE is one that draws on multiple sources of information.

Multiple-measure systems that take into account not only students' achievement growth but also teachers' classroom practices and stakeholders' responses could help mitigate some of the dangers of a system that uses only test-score data, while providing the kind of detailed evidence that could help inform efforts to improve teachers' performance.

A system of multiple measures of TE is the linchpin of the IP reforms, and each site designed such a system in collaboration with stakeholders. Sites assigned overall ratings to teachers, with the bulk of the weight coming from the achievement growth and classroom-observation measures. Most sites also included student feedback surveys, but the weight assigned to this component of the system was much smaller than the weight for the achievement and observation components. Analysis of the distributions of teachers' overall ratings indicates that, during the initiative, increasing percentages of teachers were rated in the highest categories and diminishing percentages were assigned to the lowest categories. However, the reasons for this apparent improvement are not clear.

The sites had to overcome several challenges that arose during implementation. In particular, interviewees indicated that the classroom-observation requirements placed a heavy burden on principals' time and might have detracted from other leadership responsibilities. In HCPS, the only site that used peer evaluators, the reliance on peers was generally unpopular among teachers, suggesting that use of peers might not be an ideal approach to reducing burdens on SLs. Another challenge stemmed from the lack of test scores for all subjects and grade levels in every site except HCPS, which prevented those sites from calculating achievement growth measures for all teachers and might have contributed to perceptions of unfairness among some teachers.

Majorities of teachers were aware of the components of the evaluation measures and expressed favorable opinions of these measures, particularly the observation component. For example, in the spring of 2016, large majorities of teachers who responded to the survey indicated that the observations provided valid measures of their effectiveness as teachers to either a large or moderate extent, whereas the percentages who said the same about the achievement measures, student

surveys, and overall composite measure were smaller. Except in HCPS, most teachers held favorable views of the observation rubric, observers, and the number and length of observations. In the three districts, roughly half of teachers agreed that the teacher-evaluation system did a good job distinguishing effective from ineffective teachers. CMO teachers, as well as SLs across sites, expressed more-favorable opinions about accuracy than district teachers did.

One potentially noteworthy finding is that, although many teachers disagreed that the evaluation system was fair to all teachers, majorities reported that the system had been fair to *them*. This finding might reflect the fact that large majorities of teachers received high ratings on the evaluation, and, in fact, teachers who received high ratings were more likely than low-rated teachers to agree that the system was fair to them. Finally, most teachers agreed that the evaluation system had helped them improve their instruction, but some mentioned side effects, such as a perceived need to "play it safe" and avoid taking risks or exploring new instructional approaches. This inclination to avoid experimentation might be desirable if the practices that evaluation systems measure clearly represent the highest quality of teaching but could be detrimental if teachers become unwilling to try out potentially promising strategies that might ultimately be more effective than those emphasized by the current systems.

CHAPTER FOUR

Recruitment, Hiring, Placement, and Transfer Policies

> Our teaching staff will only be as good as what we are allowed to select from. We have to strengthen our pool of eligible candidates . . . .
>
> —SL, written comment in survey

> The *best* thing I have in my school is [the teacher residency program]. The residents train at my building all year and they are the top candidates for my hiring.
>
> —SL, written comment in survey

In Brief

How did sites modify their recruitment, hiring, placement, and transfer policies to achieve the initiative's goals of more-effective newly hired teachers, an improvement in overall TE, and greater access to effective teaching for LIM students?

Our analysis found that all the sites modified their recruitment and hiring policies somewhat during the IP initiative, primarily by broadening recruitment efforts, automating the application and screening process, or, in some sites, facilitating hiring for hard-to-staff schools. In addition, some of the sites developed residency programs and partnerships with local colleges and universities to increase the pool of future teacher candidates. Even so, the sites had difficulty attracting effective teachers to high-need schools, and persistent teacher turnover was a particular problem for the CMOs. Also, changes to placement and transfer policies were relatively uncommon.

SLs were generally aware of their sites' policies and thought that the hiring process worked fairly well; moreover, perceptions of the hiring process improved in PPS and SCS over time. SLs perceived residency programs as having beneficial effects, and, in sites that offered incentives for effective teachers to transfer to high-need schools, many, but not all, SLs at high-need schools perceived a benefit. Many teachers indicated that they would reject transfers to schools not of their choosing, but, over time, fewer and fewer SLs reported experiencing the loss of good teachers from, and the assignment of unwanted teachers to, their schools.

Introduction

Recruitment and hiring procedures partly determine the effectiveness of a district's or CMO's teacher workforce. The quality of the teacher candidates whom a site attracts contributes to the quality of the teachers it employs. Recruitment and hiring are often inextricably linked to policies on the placement of new teachers in schools and the transfer of experienced teachers between schools. For example, LIM students' access to effective teaching is determined partly by policies on transfer preferences given to experienced teachers, by the timing of employment offers from low-performing schools, and by any other incentive policies that influence teacher placement. The IP initiative developers and site leaders recognized that effective recruitment, hiring, placement, and transfer policies were important in improving TE overall, as highlighted in the theory-of-action model in Figure 1.1 in Chapter One.

In this chapter, we first discuss what the literature indicates are best practices with respect to recruitment and hiring. Then, we describe the recruitment, hiring, placement, and transfer policies enacted by each site as part of its IP initiative. Following that, we summarize broad features of the sites' implementation processes and the challenges they faced. Finally, we examine teacher and SL attitudes toward the sites' recruitment, hiring, placement, and transfer policies and teacher outcomes associated with them. We close with a summary of our findings.

Best Practices with Respect to Recruitment and Hiring

Maintaining a high-quality teacher workforce begins with recruiting and hiring, and the need to attract and retain good teachers is a regular topic of discussion in educators' professional journals (e.g., Tooms and Crowe, 2004; Clement, 2013; Yaffe, 2015). Research shows that many districts, particularly those serving LIM students, have difficulty attracting and retaining enough qualified candidates (Guarino, Santibañez, and Daley, 2006; Loeb, Kalogrides, and Beteille, 2012). For example, Lankford, Loeb, and Wyckoff found that schools serving low-income students have more new teachers, more teachers with limited education backgrounds, and fewer certified teachers than schools serving high-income students (Lankford, Loeb, and Wyckoff, 2002). For example, researchers at Harvard University studied four large school districts and found that first-year teachers were far more likely than more-experienced teachers to be placed with students who started the year performing considerably behind their peers (Center for Education Policy Research, 2012). Placement into high-need schools makes a new teacher's job more challenging and increases the likelihood that the teacher will leave teaching. In fact, schools serving LIM students have higher teacher turnover than other schools and, as a result, are often staffed with higher percentages of new and inexperienced teachers (Kini and Podolsky, 2016). This leads to a "revolving door" of more inexperienced teachers serving in more high-need schools while more-effective teachers work with other students.

However, research suggests that changing recruitment, hiring, transfer, and placement policies can help stop the revolving door of lower-qualified teachers being overrepresented in high-need schools. Although there is limited empirical research on the effectiveness of specific recruitment strategies (Allen, 2005), there is considerable research on the working conditions that influence teachers' decisions to enter, stay in, or leave the profession (Ladd, 2011; Johnson, Kraft, and Papay, 2012). Podolsky and her colleagues found that teachers' decisions about where to work and whether to remain in the profession were greatly influenced by their perceptions of school leadership and administrative support, opportunities for professional collaboration and shared

decisionmaking, reduced accountability and testing pressures, and adequate resources and facilities to support teaching and learning (Podolsky et al., 2016). Ladd found that working conditions, particularly school leadership, were highly predictive of North Carolina teachers' planned and actual movement (Ladd, 2011). Similarly, Johnson, Kraft, and Papay found that a positive work context—culture, leadership, relationships among colleagues—predicted Massachusetts' teachers' likelihood, independent of the students' demographic characteristics, of remaining in teaching (Johnson, Kraft, and Papay, 2012).

In addition to these general features that make a school more or less attractive to teachers, researchers have found that districts and schools can increase their attractiveness to teacher candidates by changing a variety of recruitment policies at the school and district levels (Guarino, Santibañez, and Daley, 2006). For example, Podolsky and her colleagues found that late hiring—that is, late in the hiring cycle (e.g., just before the start of the school year)—had negative effects on teacher candidates' decisions to accept job offers (Podolsky et al., 2016). Schools that made such offers were more likely to be turned down by their preferred candidates. They also found that districts that could not communicate quickly and effectively using modern technology were less successful in attracting candidates. This led to a recommendation to expedite the hiring process to make sure decisions are made early and based on adequate information (e.g., candidates have a chance to experience the environment in which they will work, and school staff have an opportunity to see the candidate working with students). Similarly, a study of four hard-to-staff districts found that their recruitment efforts identified high-quality candidates, but, by the time offers of employment were made, many of the most qualified had taken teaching jobs elsewhere (Levin and Quinn, 2003).

Many of the recruitment and hiring problems identified in the literature occur at the district level and involve systems and procedures that are inefficient and not optimized to support effective action. For example, TNTP identified procedural problems that interfere with hiring efforts, including experienced teachers' late notification of their intent to retire or leave, negotiated requirements giving more-experienced teachers priority when filling vacant positions (vacancy

transfer rights), and poor internal forecasting (Levin and Quinn, 2003). These problems can be addressed by changes in district procedures (e.g., districts can revise their transfer timelines so vacancies are identified in the spring before the school year ends and positions can be advertised earlier). Further, districts can institute policies that give the schools with the greatest need an advantage in the hiring process (e.g., give these schools early opportunities to review candidates and give them priority in hiring).

Other educators suggest broader efforts districts can make that will benefit the hiring process, such as investing in the development of high-quality principals, who create environments in which teachers feel supported, respected, and appreciated, and including teachers in school decisionmaking to foster a more positive school culture (Podolsky et al., 2016; Ingersoll, 1999; Brill and McCartney, 2008). As teachers grow more satisfied with working conditions in their schools, they will be less likely to leave—thus reducing the number of vacancies—and new candidates will be more likely to accept offers of employment. Similarly, Johnson, Kardos, and their colleagues recognized that recruitment and hiring efforts were also influenced by school-level practices, including the school's ability to ensure a good fit between the candidate and the teaching position; connect new teachers with appropriate mentors; and employ a well-defined, standards-based curriculum to support the teacher while permitting sufficient flexibility (Johnson, Kardos, et al., 2004).

Various groups have tried to develop concise recruiting and hiring recommendations based on research. One of the clearest checklists comes from a county office of education in California, which offers a thoughtful set of specific suggestions based on both research and real-world experience (Tulare County Office of Education, 2008). At the broadest level, the authors argue for the need to focus the organization's mission at all levels around recruiting and hiring, to enhance the district's efforts in marketing and outreach, and to streamline internal policies and procedures to remove obstacles and maximize information flow. Making recruitment a priority entails an investment of time, people, and resources. They offer specific suggestions to districts toward that end, ranging from surveying teachers annually to ascertain

their future plans; to using an effective screening and selection tool; to designating a knowledgeable single point of contact for each candidate who can answer questions and address their concerns. As these examples suggest, even small things that make candidates feel valued and welcome can improve recruitment (Decker, 2015).

Some districts offer additional compensation to recruit teachers to work in schools in low-income neighborhoods, but the research on such placement-incentive programs suggests that they have limited effects. On the one hand, incentives can be effective at increasing recruitment and hiring numbers. On the other hand, the additional rewards do not make up for deficiencies in working conditions often found in these schools, and, where those deficiencies persist, teachers are less likely to remain (David, 2008). As David (2008) summarizes, "even when bonuses succeeded in drawing teachers to the poorest schools, such incentives could not compensate for the lack of support they encountered in these schools, which in turn contributed to the departure of many of these teachers" (p. 4).

In areas where candidate pools are insufficient, a longer-term recruitment and engagement strategy might be needed (Allen, 2005; Center for Public Education, 2008). For example, some districts have begun working with local HSs and colleges to create pathways into the profession (Podolsky et al., 2016; Tulare County Office of Education, 2008). Many such grow-your-own teacher-preparation models try to increase the hiring pool and build future candidates' connections with a particular district or school. An example of the latter type comes from Fulton County Schools, which redesigned its student-teacher program to use evidence from the Georgia Teacher Keys Effectiveness System to ensure that student teachers are placed with better-matched mentors and develop stronger ties to the school (L. Jones, 2017).

Teacher residency programs are another example of grow-your-own efforts that are spreading in places with teacher shortages. In these programs built along the lines of medical residency programs, teacher residents receive stipends while being trained and complete a yearlong, in-school residency during which they practice teaching under the guidance of a teacher-mentor. Teachers who are accepted into residency programs are often asked to make multiyear commitments to teaching

in their districts, so they are more likely to remain (National Center for Teacher Residencies, undated).

Finally, many of the factors that influence the teacher labor market are outside the control of individual schools or districts, but they might be responsive to state intervention. For example, a piece by Carver-Thomas and Darling-Hammond recommends that, to address the California teacher shortage, the state enact scholarships or loan-forgiveness programs for teachers agreeing to work in high-need areas, promote teacher residency models and other programs to boost supply in critical subjects and locations, and reduce barriers to reentry for retired teachers (Carver-Thomas and Darling-Hammond, 2017). Such programs are beyond the capacity of individual districts, but districts with hiring shortages can advocate for them and should be prepared to take advantage of them if they are created.

IP Recruitment, Hiring, Placement, and Transfer Policies

All the sites had recruitment, hiring, placement, and transfer policies prior to the IP initiative, which they reviewed and, in many cases, revised as part of their initiative design and implementation efforts. We did not fully document the preinitiative practices, but we did monitor changes that occurred during the initiative. Among the modifications that occurred in one or more sites were changes in recruitment and hiring policies to attract larger pools of stronger candidates, make better choices among them, and confirm hiring decisions earlier. There were also changes in placement and transfer policies, such as giving preferences so that high-need schools were able to employ more-effective teachers.

In this section, we begin by summarizing recruitment, hiring, placement, and transfer policies in the districts, then we describe the policies in the CMOs, and, at the conclusion of the section, we offer a brief synthesis pertaining to all seven sites. In the following paragraphs, in Tables 4.1 and 4.2, and in Appendix D, we present details.

Each district made changes to recruitment, hiring, transfer, and placement policies to meet its needs and be responsive to its

Table 4.1
Recruitment, Hiring, Transfer, and Placement Policies in the Districts Before and During the IP Initiative

Policy Area	HCPS	PPS	SCS
Recruitment	**Prior to 2009–2010 and ongoing:** HR recruits at job fairs, conducts targeted advertising campaigns through social media, and holds hiring fairs for difficult-to-staff schools. These efforts continued and broadened during the initiative (e.g., advertising in Puerto Rico, as well as Florida). District tracks recruitment data through a partnership with TNTP. Starting in 2010–2011, the mentor program was promoted as a benefit for new teachers hired by the district.	**Prior to 2009–2010 and ongoing:** HR recruits at job fairs, at local schools of education, and on social media. During the initiative, recruiting focused on increasing diversity.	**Prior to 2009–2010 and ongoing:** HR recruits at job fairs, at local schools of education, and on social media. **2009–2010 and ongoing:** During the initiative, the district contracted with TNTP to direct recruitment efforts. TNTP worked to increase the number of applicants, begin recruitment efforts earlier in the year, expand recruiting to more teacher-preparation programs, and partner with alternative certification programs, including TFA and teacher residency programs (e.g., Memphis Teacher Residency).

Table 4.1—Continued

Policy Area	HCPS	PPS	SCS
Screening	**Prior to 2009–2010 and ongoing:** Teacher applies to district for specified open positions. Central HR screens for qualifications and then passes resumes to schools, where principals review qualifying applicants for fit.	**2009–2010 and ongoing:** Teacher submits application, resume, three essays, and Gallup Insight survey to district. HR-trained teachers screen for qualifications and assess candidate's performance on application. By state law, only the top 10% of applicants move onto the "eligible list." HR passes resumes to principals, who select candidates to interview.	**Prior to 2011–2012:** TNTP manages the district's screening process. Applicants complete a paper application and telephone interview.
			2011–2012 and ongoing: The application process is completed online.
			2013–2014 and ongoing: Telephone screening revised to align more with the TEM rubric; applicants with the highest scores are referred to principals first.
	2013–2014 and ongoing: Teachers apply online through central AppliTrack system, which screens for qualified applicants using the Teacher Fit tool, a survey based on the TE rubric. It was piloted in 2013–2014 and launched for new hires in 2015–2016.		**2015–2016 and ongoing:** Applicants from other districts with TEM scores of 3, 4, or 5 are referred directly to principals without further screening.

Table 4.1—Continued

Policy Area	HCPS	PPS	SCS
Hiring	**2013–2014 and ongoing:** Each school has its own selection process. Typically, the principal conducts interviews and makes the final hiring decision. The district moved the hiring process six weeks earlier to start in April rather than June. Training in hiring strategies is incorporated into the principal pipeline and is given to all aspiring principals and APs. This training is also offered to all principals as an option for their own PD.	**Prior to 2009–2010 and ongoing:** Principals must interview PPS internal candidates first. Principal sends recommendations to HR, which transfers and assigns remaining displaced teachers to schools. External candidates can be interviewed and hired into the district before internal candidates are assigned to specific schools. However, internal candidates must be assigned to specific schools before external candidates. PPS internal candidates are assigned to specific schools from March to June; external candidates are hired and assigned from June to August. Principals receive guidance and resources (e.g., interview questions) but do not receive training explicitly in hiring strategies.	**2009–2010 through 2012–2013:** Principals were required to interview SCS internal candidates first; they could interview external candidates if no internal candidates applied. They sent preferences to HR, which made the final match of candidates to schools. Seniority determined placement if there were multiple internal candidates. Hiring occurred throughout the year when vacancies were known. Teachers were assigned to positions if not picked by a principal. **2013–2014 and ongoing:** All candidates go into the same hiring pool (i.e., principals are not obligated to interview internal candidates before external). Preferences are sent to HR, which hires based on "mutual consent" (i.e., teacher and principal both agree). The district no longer has an obligation to provide positions to internal candidates. Teachers are no longer assigned to positions. Hiring occurs throughout the year when vacancies are known. TNTP coaches principals to use TE data in hiring when available.

Table 4.1—Continued

Policy Area	HCPS	PPS	SCS
Hiring for hard-to-staff schools and positions	**2013–2014 and ongoing:**[a] One-time recruitment bonus ($1,000) and annual retention bonus ($2,000) to work in one of the 30 high-need POWER3 schools. (TE rating of HE at HCPS or in another school district is required to be eligible.) **2012–2013 through 2014–2015:** District paid a salary differential of 5% of base salary for experienced teachers or 2% of base salary for first-year teachers to work in any of the 50 high-need Renaissance schools. **2014–2015 through 2015–2016:** The salary differential changed to $1,000 for first-year teachers, $2,300 for teachers with two to ten years of experience, and $3,600 for teachers with 11 years or more. NBPTS-certified teachers receive an additional $4,500.	**2013–2014 and ongoing:** In staffing-support schools, principals must interview internal candidates first.[b] Principals can interview external candidates after internal candidates have been assigned to schools. If no internal candidates apply, external candidates can be hired before the internal transfer season is complete. HR solicits preferences and matches accordingly. HR ensures that teachers with NI or F ratings are not placed in staffing-support schools. HR offers external candidates higher step placement on the salary schedule to work in these schools. Additional payments range, on average, from $4,040 to $6,060.[c]	**2012–2013 and ongoing:** In Innovation Zone (iZone) schools, HR approval is required for hiring of teachers with TEM scores below 4.[d] Teachers can choose a recruitment bonus of $1,000 or a retention bonus of $1,000; half of the retention bonus is payable in December and half in May.

Table 4.1—Continued

Policy Area	HCPS	PPS	SCS
Transfer and placement	**Prior to 2009–2010 and ongoing:** Each school hires its own teachers. Central office transfers a teacher only if the teacher's position is eliminated. In that case, teachers are assigned in order of seniority (through 2013–2014) and then TE score (2014–2015 and ongoing). Teachers who want to change schools voluntarily apply directly to the school.[e] Principals must give transferring HCPS teachers priority over new hires except at Renaissance schools. Principals are required to check prior TE scores and can use them as a factor. Seniority is not a consideration.	**Prior to 2009–2010 and ongoing:** PPS posts vacancies each spring; school-based committees interview candidates and make recommendations to HR. PPS tries to use a mutual-match process in which principals and teachers submit preferences to HR and HR assigns placements. PPS is obligated to provide jobs for all internal candidates in order of seniority and will assign teachers to positions, if necessary.	**2009–2010 through 2012–2013:** Internal SCS candidates were hired into positions in order of seniority and assigned to positions if needed. **2013–2014 and ongoing:** The district no longer has an obligation to provide positions to internal candidates. Internal SCS teachers are no longer assigned to positions if not selected. Seniority no longer plays a role in hiring internal candidates.

SOURCES: Annual interviews with central-office staff in each of the IP districts.

NOTE: NBPTS = National Board for Professional Teaching Standards.

[a] The final HCPS POWER grant ended in September 2017.

[b] Staffing-support schools are those that PPS designates its highest-need schools.

[c] In 2015–2016, a teacher with certification in a high-need area *or* a teacher being placed in a high-need school earned, on average, an additional $4,040 annually through placement on an advanced step on the salary schedule. A teacher certified in a high-need area *and* placed in a high-need school earned, on average, an additional $6,060 annually.

[d] The iZone, or Innovation Zone, is a subset of SCS "designed to turn around underperforming schools." Its mission is "to move schools from the bottom 5% to the top 25% in the state of Tennessee" (SCS, undated).

[e] As of April 2017, HCPS transfers also have the option of applying through AppliTrack.

Table 4.2
Recruitment, Hiring, Transfer, and Placement Policies in the CMOs Before and During the IP Initiative

Policy Area	Alliance	Aspire	Green Dot	PUC
Recruitment	**Prior to 2009–2010 and ongoing:** HR recruits at job fairs, universities, TFA, and LinkedIn (for math and science positions), and through referrals from TFA alumni. Principals also do their own recruiting.	**Prior to 2009–2010 and ongoing:** HR recruits at job fairs, colleges, websites, open houses, and interview days hosted by Aspire.	**Prior to 2009–2010 and ongoing:** HR recruits using referrals from current staff, job fairs, and TFA.	**Prior to 2009–2010 and ongoing:** HR recruits at job fairs and universities, through partnerships with UC Los Angeles (UCLA) extension and the Claremont Colleges, and through TFA.
			2011–2012 only: The CMO has a residency program with Loyola Marymount University.	
			2010–2011 and ongoing: The CMO created an HC department to focus on recruitment, retention, PD, transitions, and performance management.	**2010–2011 only:** Math and science residency program with Loyola Marymount University was used as recruitment tool. (It was discontinued because of implementation problems.)
		2014–2015 and ongoing: HR also recruits through TFA. Principals can do their own recruiting.	**2015–2016 and ongoing:** It increased the number of student teachers.	

Table 4.2—Continued

Policy Area	Alliance	Aspire	Green Dot	PUC
Recruitment, continued	**2011–2012 through 2012–2013:** A residency program with Loyola Marymount University was used as a recruitment tool for high-need subjects.	**2013 and ongoing:** It expanded recruitment to East Coast sources.	**2014–2015 and ongoing:** It established an ongoing partnership with CSUDH, to review and provide PD support for preservice students. It hired student ambassadors at target colleges to help recruit.	**2014–2015 and ongoing:** It has a residency program with Loyola Marymount University for alumni of PUC.
	2014–2015 only: It had a residency program with the University of the Pacific in all subject areas.	**2010–2011 and ongoing:** It has a teacher residency program.	**2015–2016 and ongoing:** It has increased use of social media.	

Table 4.2—Continued

Policy Area	Alliance	Aspire	Green Dot	PUC
Screening	**Prior to 2009–2010 and ongoing:** Teacher applies online through EDJOIN or directly to the school. HR screens all candidates for proper credentials and security.	**Prior to 2009–2010 and ongoing:** Teacher applies online through EDJOIN or directly to the principal. Initial phone screening can be by HR or the principal if the principal so chooses.	**Prior to 2009–2010 and ongoing:** Teacher applies online through EDJOIN or directly to the school. HR conducts more-extensive screening, including phone interview, submission of a lesson plan, and group interview at the central office. When a principal identifies a direct-application candidate to hire, HR screens for eligibility and informs the principal.	**Prior to 2009–2010 and ongoing:** Teacher applies through PUC online job platform. HR screens for appropriate credential and teaching experience.
	2012–2013 through 2015–2016: HR adopted additional screening of online candidates using phone and in-person interviews and created a pool of screened candidates, which about 35% of principals choose to utilize.		**2015–2016 and ongoing:** HR adopted a shortened process for highly qualified candidates (e.g., TFA alumni).	

Table 4.2—Continued

Policy Area	Alliance	Aspire	Green Dot	PUC
Hiring	**2009–2010 through 2014–2015:** Each school had its own hiring process. The principal made the final hiring decision. No training was provided for principals. Typically, hiring was from March to August.	**Prior to 2009–2010 and ongoing:** A candidate is interviewed by a panel of school parents and teachers and conducts a sample lesson for teachers in his or her subject area and administrators. The principal makes the final hiring decision. New principals receive one-day interview training; each year before any interviewing, HR holds a training session for all staff involved in hiring. Typically, hiring is from March to August.	**Prior to 2009–2010 and ongoing:** A candidate is interviewed by a school hiring team, conducts a sample lesson rated using the observation rubric, and responds orally to standard teaching scenarios. The principal makes the final hiring decision. HR provides an interviewing workshop for new principals and hiring teams. Typically, hiring is from March to August.	**Prior to 2009–2010 and ongoing:** Hiring is coordinated centrally. Principals review applications then attend central-office interview days and observe an interview of a panel of candidates and, the next day, a candidate teaching a demo lesson. The principal makes the final hiring decision. HR conducts hiring training sessions twice a year.
	2015–2016 and ongoing: The CMO has a standard procedure for its hiring process.			**2015–2016 and ongoing:** The process has shortened from two days to one day. Typically, hiring is from March to August.

Table 4.2—Continued

Policy Area	Alliance	Aspire	Green Dot	PUC
Hiring for hard-to-staff schools and positions	Some principals provide signing bonuses for math or science teachers, but that is rare.	**2013–2014 and ongoing:** It offers $10,000 incentives for HE and master teachers to move to low-achieving schools participating in the Aspire Focus School Incentive Program. Occasional stipends go up to $2,500 for hard-to-fill positions, especially for interim positions. Teacher residents who are hired for focus schools receive hiring bonuses.	**2014–2015 and ongoing:** The CMO makes early offers for hard-to-staff positions, contingent on the expected position materializing.	Same as for other schools and positions

Table 4.2—Continued

Policy Area	Alliance	Aspire	Green Dot	PUC
Transfer and placement	**Prior to 2009–2010 and ongoing:** Each school hires its own teachers. The central office does not place or transfer teachers. A teacher wanting to change schools must apply directly to the desired destination school.	**Prior to 2009–2010 and ongoing:** Each school hires its own teachers. The central office does not place or transfer teachers. A teacher wanting to change schools must apply directly to the desired destination school.	**Prior to 2009–2010 and ongoing:** Each school hires its own teachers. The central office does not place or transfer teachers. A teacher wanting to change schools must apply directly to the desired destination school. If layoffs are necessary, teachers are placed on a reemployment list for 12 months and are offered any vacant positions that meet their qualifications. If a teacher refuses the position, that teacher is removed from the reemployment list.	**Prior to 2009–2010 and ongoing:** Each school hires its own teachers. The central office does not place or transfer teachers. A teacher wanting to change schools must apply directly to the desired destination school.

SOURCES: Annual interviews with central-office staff in each of the IP CMOs.

NOTE: CSUDH = California State University, Dominguez Hills.

state and local context. After the IP initiative began in 2009–2010, each of the three districts adjusted its procedures for recruitment, screening, hiring (and hiring for hard-to-staff schools and positions), and transfer and placement; Table 4.1 provides a summary. In the area of recruitment, SCS greatly broadened its outreach efforts, contracting with TNTP to expand and accelerate recruitment. PPS also modified its recruitment efforts, albeit later in the initiative, with the goal of attracting more minority candidates. All three districts had screened applicants centrally before referring qualified candidates to principals for review, and each adopted some type of online application process, mostly during the IP initiative. In HCPS and SCS, the screening process was modified to gather information about the extent to which candidates possessed the characteristics of effective teaching embodied in the district's effectiveness rubric. In all three districts, principals played a key role in hiring decisions; their hiring recommendations were informed by interviews and their own review of applications, but it was the central HR office that made the final offer of employment based on principals' recommendations.

Other staffing changes made by the districts include moving the hiring process earlier in the year (HCPS and SCS), expediting the referral of internal candidates with high effectiveness ratings (SCS), and training principals to make better use of information about effectiveness in hiring (SCS). In HCPS, moving up the hiring timeline by six weeks was successful in helping HR and SLs fill more open positions before the start of the school year than they had in the past. In addition, SCS complied with a state policy change made in 2014–2015 whereby all candidates for teaching positions go into the same hiring pool; internal candidates (i.e., teachers transferring from other schools within the district) no longer have preference.

The CMOs continued their active recruitment efforts, expanded hiring activities, and initiated partnerships with local colleges and universities to develop future candidates. Table 4.2 summarizes the CMOs' efforts to optimize their recruitment, hiring, placement, and transfer efforts during the IP initiative. The CMOs have always been attentive to recruitment and hiring policies because they operate in highly competitive California labor markets and because they

are nontraditional schools competing with generally higher-paying districts. In the CMOs, both the central office and individual principals are active in recruiting, and all the CMOs established relationships with local colleges to recruit candidates. Green Dot formed a partnership with a local university whose students had similar backgrounds to the students served by the CMO. The partnership gives Green Dot more access to potential candidates at an earlier stage by giving Green Dot a spot on the university's review panel for teacher-training applicants, as well as a role in providing PD for teacher trainees and hosting trainees as they do their student teaching. Two of the CMOs also launched teacher residency programs to grow their own candidates.

Like in the districts, principals in the CMOs had primary responsibility for hiring decisions, but school staff members (and even parents) were involved extensively in reviewing and hiring. Several of the CMOs did extensive vetting of candidates through interviews and by having them teach sample lessons.

With the exception of Green Dot, which has a teachers' union, teachers in the CMOs have no special transfer rights to move among schools, and those wishing to change schools have to apply just like external candidates do. All the CMOs serve primarily LIM students, so, with the exception of Aspire, there are no special financial incentives to work in particular schools.

Many of the sites adopted technology to improve their recruitment, application, and screening processes. New technology played a role in many sites' efforts to improve recruitment and hiring. For example, HCPS piloted AppliTrack in 2013–2014, and the program was launched district-wide in 2015–2016. AppliTrack makes it easier for HR staff to screen candidates for eligibility and for principals to access information about candidates. Similarly, SCS adopted an online application system with the help of TNTP, and the CMOs continued to use the online EDJOIN system to facilitate the application process.

Sites increased their efforts to welcome and support new teachers through additional onboarding and mentoring activities. Some teacher support activities make schools more attractive to teacher candidates and are sometimes considered to be part of the recruitment and hiring process, as well as part of the PD process. One example

is the HCPS mentor program (discussed in more depth in Chapters Six and Eight). According to the district, the program was used as a recruiting tool to make HCPS an attractive place to teach. One HCPS administrator said, "We have teachers coming in [who] will lie and say they have no experience because they want to get a mentor." Similarly, many of the CMOs expanded their coaching staffs and targeted their efforts primarily at new teachers. SCS also provided each teacher who was new to the district with a peer mentor; this was a popular program, according to many teachers we interviewed.

Implementation Challenges

The sites' recruitment and hiring policies influence how they interact with the local labor market and the higher education community, and the local context had a significant influence on these policies. As noted above, the sites were generally successful in broadening their outreach efforts and improving their application and screening processes. However, both the districts and the CMOs also experienced challenges related to staffing, as described in this section.

Sites' efforts to attract effective teachers to transfer to high-need schools were not very successful. Many sites attempted but found it difficult to entice effective teachers to transfer to high-need schools. For example, HCPS added small salary incentives to work in hard-to-staff schools (see Table 4.1), but these efforts to entice effective teachers to transfer to high-need schools had limited success. Central-office staff in HCPS reported that teachers were reluctant to transfer to high-need schools despite the cash incentive and extra support because they believed that obtaining a good VAM score would be difficult at a high-need school. Similarly, PPS's CL roles (discussed in more detail in Chapter Eight) were designed as incentives for effective teachers to transfer to high-need schools. According to interviews with teachers and central-office staff, lack of interest in the program among some teachers was due, in part, to a reluctance to transfer schools.

In many cases, local and state contexts presented challenges to sites' efforts to recruit, hire, and place more-effective teachers.

SCS recruitment efforts had to address challenges raised by both the merger between legacy MCS and legacy SCS and the creation of the state-managed ASD in Tennessee, to which the lowest-performing 5 percent of schools were assigned. Controversy preceding the merger, which took about 18 months to resolve, created uncertainty for teachers, and central-office staff were worried that teachers would seek new positions outside the district. In fact, we were told that some central-office staff sought new positions during the lead-up to the merger. The ASD, which took over some of the district's lowest-performing schools, also presented a challenge by drawing preferred candidates away from the district. On the other hand, policy changes at the state level in Tennessee, particularly those enacted as a result of Tennessee's RTT grant, facilitated some of SCS's staffing reforms. In particular, the state eliminated collective bargaining laws, which gave the teachers' association much less power to influence staffing decisions, and the state eliminated seniority as a factor in hiring decisions.

Local conditions also influenced the implementation of recruitment and hiring changes in PPS but in very different ways. In the early years of the grant, budget shortfalls and declining enrollment limited the hiring of new teachers. Even without shrinking enrollments, PPS generally has few open positions each year because turnover is low. The district contract with the teachers' union also links teacher seniority to transfer and placement policies, which, according to some SLs, constrains their ability to make decisions about which teachers work in their schools and limits the hiring of external candidates. Teachers' contract provisions also contributed to derailing PPS's plans to implement its own teacher-training and pipeline program, in the form of two teacher academies. Coincidentally, the academies, which were to train new teachers, were to be implemented in the same year that PPS budget shortfalls and enrollment declines forced PPS to furlough about 180 teachers. Because PPS's contract with the union, as well as state law, required teachers to be furloughed by seniority, the new teachers coming to the academies—because they would be the newest—would be the first to be furloughed. PPS and the union were unable to negotiate an exception for the academy teachers, and, ultimately, the academies were not implemented.

The local labor market presented a challenge to the CMOs. Despite their active recruitment efforts, they had difficulty finding enough candidates who were committed to working with the low-income, low-performing, minority populations they served and with the college-ready-for-all culture they encouraged. Aspire's website specifies that its culture "[f]uels teammates to come to work every day and persist in the face of adversity" (Aspire, undated). A Green Dot central-office staff person captured the challenge of working in a tight labor market in the fall of 2014: "We're seeing higher-quality candidates because of our vetting and hiring process. Where we're struggling is [with] the number of candidates in the pool."

Teacher turnover created challenges for some of the IP sites. The CMOs are relatively young organizations, with PUC having established the earliest school in the group in 1999, and, during the initiative period, they were all expanding and adding schools. Like the schools themselves, most of the teachers working in them were young, and, for many, the CMOs provided their first teaching jobs. In addition, many teachers were recruited from TFA and intended to remain in teaching for only a few years. As a result, the CMOs experienced a high degree of teacher turnover and a continuing need for new teachers. (In Chapter Eleven, we examine the success of teacher-retention efforts.) In addition, starting in 2014, the improved economy meant more competition for veteran and HE teachers, who could find jobs in other school districts that might have offered higher salaries, shorter calendars, and more-advantaged students. As a consequence, the CMOs struggled to compete for the best teachers and teacher candidates. As one HR administrator said in 2014, "With lack of candidates comes people we probably would not have selected, and they have not lasted." Turnover was one factor that led the CMOs to initiate residency programs and partnerships with universities to enhance the pipeline of new teachers. PPS, meanwhile, found it challenging to keep experienced teachers from transferring out of the district's highest-need schools. To address this, PPS designated these schools as staffing-support schools and implemented policies to reduce turnover in them (as described in Table 4.1 and Appendix D).

SCS experienced some challenges in their efforts to supplement their internal HR capacity by engaging with an external organization. Prior to the initiative, legacy MCS had an inefficient teacher application process and a limited candidate pool, according to central-office staff.[1] To address these shortcomings, the district established a relationship with TNTP prior to receiving the IP grant; once it received the grant, the district used some of the funds to expand this relationship. TNTP functioned to some degree as the district's HR office and was responsible for implementing most of the staffing levers. (See Appendix D for additional details.) Unfortunately, there were tensions between TNTP and the district's HR office throughout the initiative. Central-office and TNTP staff told us that members of the HR staff were opposed to the partnership with TNTP and were disinclined to collaborate with it, although this relationship reportedly improved over time. Although the relationship between HR and TNTP created some problems, the engagement with TNTP achieved many of its goals. In particular, central-office and TNTP staff told us that TNTP was successful in expanding the applicant pool, both in terms of numbers and geographic distribution of applicants and in terms of automating and streamlining the application and screening process, the vacancy notification process (by which principals notified HR that they had vacancies), and the internal transfer process (prior to 2014–2015, when the rules were changed).

Teacher and SL Reactions to Recruitment, Hiring, Placement, and Transfer Policies

To examine educators' perceptions of the sites' recruitment, hiring, placement, and transfer policies, we explore their input, from surveys and interviews, on three broad topics:

- awareness: Did SLs know about the policies?
- endorsement: Did SLs and teachers approve of the policies?

[1] This discussion refers primarily to legacy MCS, prior to the merger with legacy SCS.

- perceived effects: Did SLs find the policies useful?

Our information on perceptions of the sites' policies related to recruitment, hiring, placement, and transfer comes almost entirely from SLs because surveys and interviewers did not generally ask teachers about these topics.

Awareness

To gauge SLs' awareness of the sites' policies related to recruitment, hiring, placement, and transfer, we asked them in our annual spring surveys whether their site did any of the following things:

- "Work with external organizations to hire new, high-quality teachers (for example, TNTP or Teach For America)"
- "Train/hire teachers through a residency program"
- "Offer incentives for *effective* teachers (for example, based on teacher evaluation results or ratings) to transfer to a high-needs school"
- "Offer incentives for *any* teacher, regardless of effectiveness, to transfer to a high-needs school"
- "Consider teacher effectiveness in addition to (or instead of) seniority when making decisions about teacher transfers."

In sites that worked with external organizations to hire teachers, most SLs were aware of this practice. Indicating an awareness of the policies described in Tables 4.1 and 4.2, high percentages of SLs in SCS, Alliance, Green Dot, and PUC reported that their site worked with external organizations to hire teachers; these percentages exceeded 80 percent in most years, especially 2012 and later. Percentages were somewhat lower in Aspire (50 to 80 percent, depending on the year) and much lower in HCPS (30 to 50 percent) and PPS (15 to 40 percent), again in accordance with the fact that these sites did not make heavy use of external organizations for recruitment.[2]

---

[2] In the past, high-poverty schools have faced greater challenges recruiting teachers, and some might wonder whether SLs in those schools were equally aware of the recruitment efforts. However, we found no consistent difference in responses to this question from SLs

Except in SCS and Alliance, most SLs knew whether their site had a residency program in any given year. SLs' reports of whether their site trained or hired teachers through a residency program were also generally consistent with sites' policies as reported in Tables 4.1 and 4.2. Nearly all Aspire SLs indicated that the CMO had a residency program every year from 2013 (the first year we asked the residency question) through 2016, which was consistent with Aspire's robust and ongoing residency program. In PUC, nearly every SL reported a residency program in 2015 and 2016, when the residency program with Loyola Marymount University was reestablished. Few SLs in PPS (20 percent or less) reported a residency program in any year (there was not one), and relatively few SLs (20 to 40 percent) in HCPS or Green Dot reported a residency program in any year except for Green Dot in 2016 (50 percent), possibly reflecting the partnership that Green Dot established with CSUDH the previous year. In SCS, 40 to 60 percent of SLs correctly reported a residency program. The results for Alliance are somewhat less consistent with the site's use of a residency program.[3]

In HCPS, SCS, and Aspire, which offered—or began to offer—incentives for effective teachers to transfer to high-need schools, a majority of SLs knew about these incentives. We also asked SLs whether their site offered incentives for *effective* teachers (with effectiveness based on teacher-evaluation results or ratings) to transfer to a high-need school, as well as whether their site offered incentives for *any* teacher, regardless of effectiveness, to transfer to a high-need school. Demonstrating an awareness of HCPS's incentives for high-performing teachers to teach in the POWER3 schools, most SLs in HCPS (65 to 85 percent) indicated that the district offered incentives

---

when we split them into two groups based on schools' percentages of LIM students. This was true for both of the two types of LIM comparisons we made: comparing high-LIM schools (i.e., schools with at least 80 percent LIM students) with all other schools in the site and comparing schools above and below the site-wide median percentage of LIM students. (See "Survey Methods" in Appendix A.)

[3] There was no consistent relationship between responses from SLs in low- and high-LIM schools, using either specification of LIM as described earlier.

for effective teachers to transfer to a high-need school.[4] At the same time, however, only 40 to 50 percent of HCPS SLs reported that there were incentives for *any* teacher to transfer to a high-need school, when such incentives existed for teaching in the Renaissance schools, so this could suggest some lack of awareness.[5] SCS and Aspire both instituted incentives, partway through the initiative, for effective teachers to transfer to high-need schools, and, correspondingly, the percentage of SLs reporting such incentives increased after they were implemented.[6]

It is harder to make sense of responses about transfer incentives from school leaders in PPS, Alliance, Green Dot, and PUC in light of those sites' policies. For example, after the spring of 2012, the percentages of SLs reporting incentives for effective teachers to transfer to high-need schools decreased in PPS and increased in the three CMOs; neither change corresponds directly to identifiable policy changes. The response patterns in these sites might reflect the complexity of the policies in those sites, some degree of misunderstanding, or both.

---

[4] As one might expect, in all three years tested (2014, 2015, and 2016), SLs in high-LIM schools (i.e., schools with at least 80 percent LIM students) were significantly more likely ($p < 0.01$) than other SLs to report that such incentives existed.

[5] Contrary to what one might expect, SLs in high-LIM schools (i.e., schools with at least 80 percent LIM students) were no more likely than other SLs to report that such incentives existed in any of the three years tested (2014, 2015, and 2016), and, in fact, in 2014, SLs in high-LIM schools were significantly *less* likely ($p < 0.001$) than other SLs to report that such incentives existed.

[6] In SCS, in which the incentives to transfer to high-need schools began in the 2012–2013 school year, the percentage of SLs reporting such incentives increased from about 30 percent in the springs of 2011 and 2012 to 60 percent in the spring of 2013 (and then 45 to 55 percent from 2014 through 2016). In Aspire, only 13 percent of SLs reported the existence of the incentives in the springs of 2011 and 2012, but the percentage skyrocketed to 84 percent in 2013 and stayed high through 2015 before dropping somewhat to 58 percent in 2016. In the spring of 2013, half of Aspire SLs also reported incentives for *any* teacher to transfer, but that year was a fluke (i.e., the percentages were much lower in the other years) and likely represented temporary confusion as the new program kicked in.

In SCS, as one might expect, SLs in high-LIM schools (i.e., schools with at least 80 percent LIM students) were significantly more likely ($p < 0.001$) than other SLs to report, in all three years tested (2014, 2015, and 2016), that such incentives existed. In Aspire, however, SLs in high-LIM schools were no more likely than other SLs to report that such incentives existed in any of the three years.

SLs' awareness of whether effectiveness played a role in transfer decisions was consistent with site policy in HCPS and PPS but less so elsewhere. Finally, from 2013 through 2016, we asked SLs whether their site considered teacher effectiveness in addition to (or instead of) seniority when making decisions about teacher transfers. In HCPS, the percentage saying yes rose from about 50 percent in the first two years to about 60 percent in 2015 and 2016, corresponding with a policy change made in the district. In PPS, less than one-quarter of SLs said yes in any year, a pattern that is also consistent with PPS's lack of change to policies regarding seniority-based transfer during the initiative. Results are more difficult to interpret in SCS, in which about 60 percent of SLs said that effectiveness was considered in addition to seniority in 2013, 2014, and 2015, but the percentage dropped to 49 percent in 2016. This pattern appears to be at odds with the fact that, prior to the 2014–2015 school year, seniority played a role in transfers in SCS, but, in subsequent years, it did not; one would thus expect the percentages of SLs indicating "effectiveness rather than seniority" to increase rather than decrease. The change might have affected primarily central-office decisions and thus could have been invisible to SLs, or perhaps transfers were rare enough that the factors affecting them did not seem very salient. In the CMOs, meanwhile, the survey question had limited applicability because it was framed as a question about site-wide policy rather than about individual schools; thus, we do not report the CMO results here.

Endorsement

Most SLs thought that the processes by which teachers were hired to their school worked well; in PPS and SCS, perceptions improved during the grant period. In every site except PPS, more than 80 percent of SLs agreed each year from 2013 through 2016 (i.e., the years the question was asked) that "the processes by which teachers are hired to my school work well," and typically 40 to 60 percent agreed strongly. In PPS, the percentages agreeing (and agreeing strongly) were considerably lower but increased significantly and consistently during the four years; in 2013, 52 percent agreed (and 6 percent agreed strongly), but,

by 2016, 78 percent agreed (and 21 percent agreed strongly).[7] We are not sure why there was such an increase in agreement in PPS, but it was not driven by SLs in high-LIM schools, who were significantly less likely than leaders at lower-LIM schools to agree in 2014 and 2015.[8]

Not surprisingly, very few SLs in the CMOs agreed that central-office procedures "sometimes require my school to take on a teacher who is not a good fit for the school," and up to 40 percent of CMO SLs indicated that the statement did not even apply to them. Green Dot was something of an exception, with 20 percent or less reporting that the statement was "not applicable"; of the rest, 40 to 60 percent agreed that procedures sometimes led them to take on teachers who were not a good fit. SLs in Green Dot schools with LIM percentages above the median were significantly more likely than leaders in schools with below-median LIM percentages to agree with this statement in 2014 and 2015.

Even so, the Green Dot percentages were much lower than those in the three districts. About 80 percent of HCPS leaders agreed each year from 2012 through 2016 that they were sometimes compelled to take on poorly fitting teachers; in PPS, the percentage agreeing exceeded 85 percent in all but one of the years. In SCS, which did make major changes to its hiring procedures during the initiative (see Table 4.1), the percentage agreeing started out high—85 percent in 2012—but then dropped steadily through 2015 (60 percent), before rising slightly

---

[7] For percentage agreeing, the increases from 2013 to 2014 and from 2014 to 2015 were both significant ($p < 0.05$), but the increase from 2015 to 2016 was not, despite being of a similar magnitude (8 to 10 percentage points). For percentage agreeing strongly, the increases from 2013 to 2014 and from 2015 to 2016 were both significant, but the increase from 2014 to 2015 was not.

[8] This was the case for both of the two types of LIM comparisons we made: comparing high-LIM schools (i.e., schools with at least 80 percent LIM students) with all other schools in the site and comparing schools above and below the site-wide median percentage of LIM students.

to 66 percent in 2016.[9] In HCPS and SCS, SLs at higher-LIM schools were generally less likely than SLs at lower-LIM schools to agree.[10]

During the initiative, SLs and teachers became less likely to believe strongly that teacher-evaluation results should play a role in school placement decisions. We asked both SLs and teachers how much they agreed that "teachers' evaluation results should be factored in to decisions about how teachers are placed at schools." In the three districts, about 90 percent of SLs agreed somewhat or strongly every year from 2011 through 2016, so we do not show the results in a figure. These percentages were fairly steady over time, but the percentages agreeing *strongly* showed marked decreases after 2013 in all three districts, sometimes after showing increases in earlier years. Most notably, the percentage agreeing strongly in SCS started out in 2011 at 49 percent, rose to 62 percent in 2013, and then dropped steadily down to 44 percent in 2016.[11] However, these decreases were likely more a result of changed perceptions about the overall value of teachers' evaluation results for distinguishing effective from ineffective teachers than they were about placement policies per se (see Figure 3.7 in Chapter Three). Among SLs in the CMOs, agreement with the statement tended to be 5 to 10 percent lower than in the three districts (10 to 15 percent if the percentages reporting that the statement was not applicable are included in the totals). The CMO SL percentages also showed greater variability over time, with no clear trends discernible.

For their part, teachers were never that keen on the use of evaluation results in placement decisions in any of the seven sites, but they, like the SLs in the districts, became even less likely to agree over time.

---

[9] Each year's change to the following year was statistically significant ($p < 0.05$).

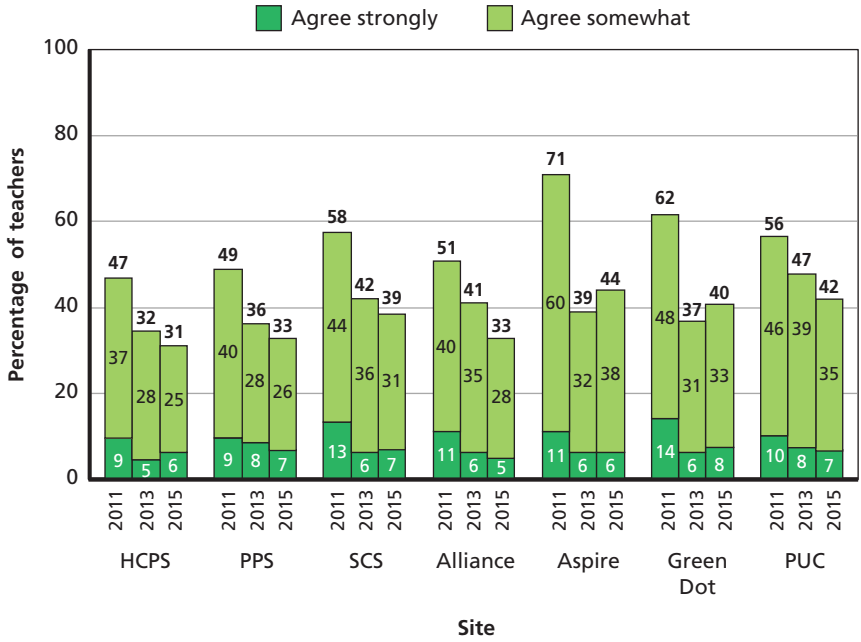
[10] This was true for both of the two types of LIM comparisons we made: comparing high-LIM schools (i.e., schools with at least 80 percent LIM students) with all other schools in the site and comparing schools above and below the site-wide median of percentage of LIM students.

[11] Each year's change to the following year was statistically significant ($p < 0.05$).

(See Figure 4.1.) Novice teachers were more likely than experienced teachers to agree with the statement.[12]

We also asked teachers and SLs whether they thought that effective teachers in low-need schools should be offered incentives to transfer to high-need schools. Majorities of both groups thought that they should. In HCPS, about 90 percent of SLs agreed every year from 2013 through 2016, with about 50 percent agreeing strongly each year. In

Figure 4.1
Teachers' Agreement That "Teachers' Evaluation Results Should Be Factored in to Decisions About How Teachers Are Placed at Schools," Springs 2011, 2013, and 2015



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly." We asked this question of teachers only in the years shown.

RAND *RR2242-4.1*

[12] Comparisons between novice teachers and other teachers were made in 2013 and 2015. In 2015, novice teachers were more likely than other teachers to agree in all sites except Green Dot, and significantly so in HCPS, SCS, and PUC. In 2013, novice teachers were more likely to agree in the three districts and Alliance, and significantly so in SCS and Alliance.

PPS and SCS, the percentages agreeing were 5 to 10 percentage points lower than in HCPS, but, like in HCPS, about 50 percent agreed strongly. No obvious time trends were visible in any of the three districts, but SLs in higher-LIM schools were more likely to agree strongly than SLs in lower-LIM schools.[13] In the CMOs, the percentages of SLs agreeing with the statement were too variable to summarize concisely.

Among teachers who were presented with this statement only in 2013 and 2015, the percentage agreeing strongly or somewhat ranged from about 60 percent to 80 percent, and the percentage agreeing strongly was typically 20 to 40 percent. There were not large differences between sites, nor were there systematic differences between 2013 and 2015; there also were not consistent differences between teachers at high-LIM schools (defined as schools with at least 80 percent LIM students) and teachers at lower-LIM schools.

Many teachers reported that they would rather leave teaching than be forced to transfer to a school that was not of their choosing. As reported in Tables 4.1 and 4.2, none of the sites had a policy of forcing teachers to transfer to other schools, but we nevertheless asked teachers how they might react if faced with a forced transfer. In most sites and years, 50 to 60 percent of teachers agreed strongly or somewhat that "I would rather leave teaching than be forced to transfer to a school that is not of my choosing"; 20 to 30 percent agreed strongly. In HCPS, the percentages were higher, with closer to 70 percent agreeing, including almost 40 percent agreeing strongly. There were also differences based on teacher and school characteristics: Experienced teachers, teachers with higher effectiveness ratings, and teachers in

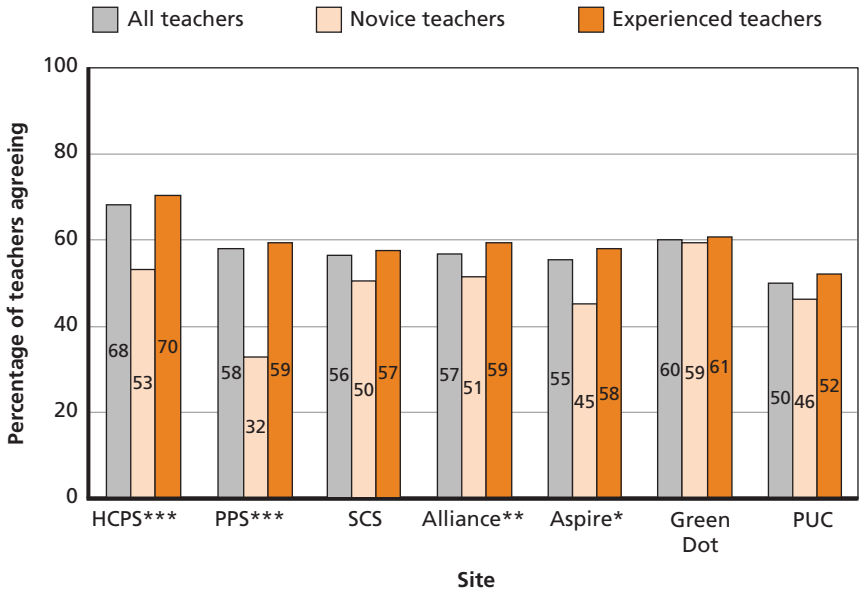
[13] For 2015 and 2016, we compared "agree strongly" responses from SLs in higher-LIM schools (schools in the top half of the within-site LIM distribution) and SLs in lower-LIM schools (schools in the bottom half of the distribution). In 2016, SLs in top-half LIM schools were more likely to agree strongly with the statement than SLs in bottom-half LIM schools in every site except Alliance, and significantly so in HCPS, PPS, SCS, and PUC. In 2015, SLs in top-half LIM schools were more likely to agree strongly with the statement than SLs in bottom-half LIM schools in the three districts and Alliance, and significantly so in HCPS and PPS.

lower-LIM schools were all more likely to agree with the statement (see Figures 4.2, 4.3, and 4.4).[14]

Perceived Effects

As described earlier, we asked SLs in our annual surveys whether their site worked with external organizations to hire teachers; had a residency program; offered incentives for teachers to transfer to high-need schools; and considered teacher effectiveness in transfer decisions. For

Figure 4.2
Teachers' Agreement That "I Would Rather Leave Teaching Than Be Forced to Transfer to a School That Is Not of My Choosing," by Teacher Experience Level, Spring 2015

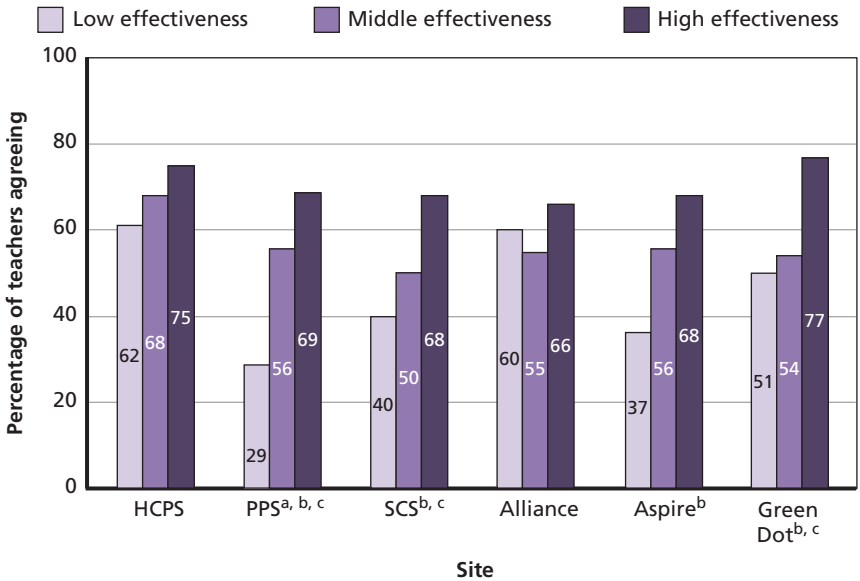


NOTE: We have combined the original response categories of "agree strongly" and "agree somewhat." Omitted response categories are "disagree somewhat" and "disagree strongly." * denotes that the difference between novice and experienced percentages is significant at $p < 0.05$. ** denotes that the difference is significant at $p < 0.01$. *** denotes that the difference is significant at $p < 0.001$.

RAND *RR2242-4.2*

[14] Figures show results from 2015. Results from 2013 were generally similar.

Figure 4.3
Teachers' Agreement That "I Would Rather Leave Teaching Than Be Forced to Transfer to a School That Is Not of My Choosing," by Prior-Year TE Rating, Spring 2015

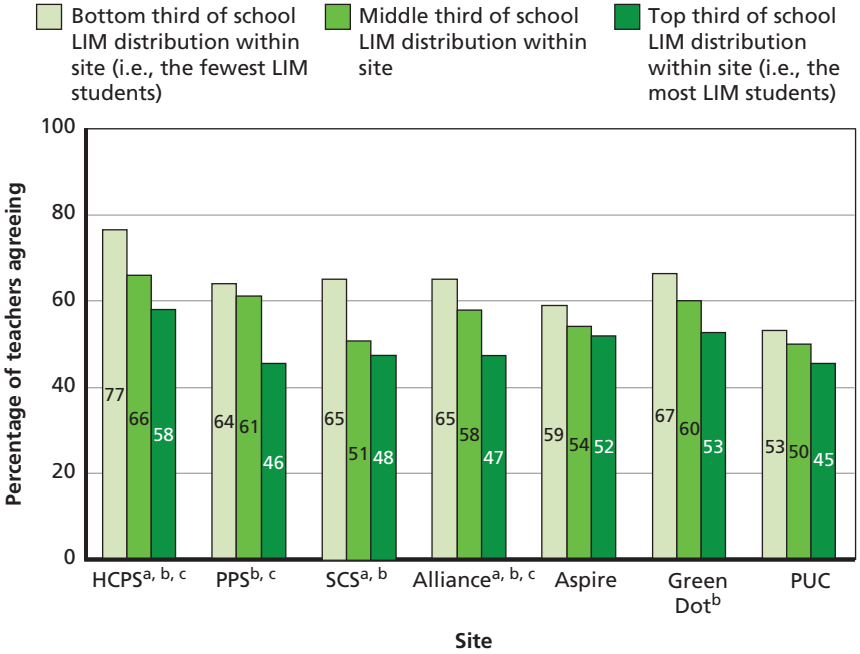


NOTE: We based definitions of *low*, *middle*, and *high* on sites' own categorizations; for details, see Appendix A. Effectiveness ratings were not available for PUC. We have combined the original response categories of "agree strongly" and "agree somewhat." Omitted response categories are "disagree somewhat" and "disagree strongly."
[a] Low–middle difference is significant at $p < 0.05$.
[b] Low–high difference is significant at $p < 0.05$.
[c] Middle–high difference is significant at $p < 0.05$.
RAND *RR2242-4.3*

those who said that their site had any of these policies, we then asked whether their school had benefited, been hurt, or neither, as a result of the policy. We summarize those findings here.

According to SLs, external organizations had neutral or positive effects on hiring. In SCS, Alliance, Green Dot, and PUC, which worked with external organizations to hire new teachers, most SLs reported that this policy had had either a beneficial or a neutral (i.e., neither positive nor negative) effect on their school; very few (less than

Figure 4.4
Teachers' Agreement That "I Would Rather Leave Teaching Than Be Forced to Transfer to a School That Is Not of My Choosing," by School's Percentage of Students with LIM Status, Spring 2015



NOTE: We have combined the original response categories of "agree strongly" and "agree somewhat." Omitted response categories are "disagree somewhat" and "disagree strongly." Tertiles are of the within-site distribution of schools by percentages of their student bodies who are LIM students.
[a] Bottom–middle difference is significant at $p < 0.05$.
[b] Bottom–top difference is significant at $p < 0.05$.
[c] Middle–top difference is significant at $p < 0.05$.
RAND *RR2242-4.4*

15 percent) reported that the policy had hurt their school. SLs in the three CMOs were the most likely to say that their school had benefited, typically with 60 to 70 percent of SLs reporting a benefit. In SCS, about 50 percent reported a benefit.

SLs perceived residency programs as having beneficial effects. Some sites initiated residency programs during the initiative; in those that did, a majority of SLs thought that these programs were beneficial.

Aspire SLs, especially, tended to respond favorably about their site's residency program; more than 70 percent in all years—and 97 percent in 2015—said that their school had benefited from the residency program. In PUC, which had a residency program in 2015 and 2016, 60 percent of SLs in 2015 and 74 percent in 2016 said that their school had benefited, and most of the rest indicated a neutral effect. Green Dot was similar; 66 percent reported a benefit in 2016 when the residency program was in place. In SCS, in which only about half of SLs indicated the use of a residency program, about half of those SLs reported having benefited, and nearly all the rest reported a neutral effect.

In sites that offered incentives for effective teachers to transfer to high-need schools, many, but not all, SLs at high-need schools perceived a benefit. In the sites that had clear policies of offering effective teachers incentives to transfer to high-need schools (HCPS, SCS, and Aspire), majorities of SLs reported a neutral effect on their school; very few (typically less than 10 percent) reported a negative effect. However, it is probably more instructive to focus on the perceptions of leaders in the high-need schools that these incentives targeted.[15] In schools with at least 80 percent LIM students, a majority of SLs in HCPS and SCS perceived a benefit from the policy in 2014 and 2015, as well as in HCPS in 2016. Moreover, in HCPS and SCS, SLs at higher-LIM schools were more likely than those at lower-LIM schools to say that their school had benefited.[16] For instance, in HCPS in 2016, 56 percent of SLs at high-LIM schools (defined as schools having at least 80 percent LIM students) thought that their school had benefited, compared with only 21 percent at other schools. The direction of this

[15] We did not have lists of the schools classified by each site each year as high need, so we used our LIM classifications as a proxy.

[16] We compared responses from SLs at higher-LIM and lower-LIM schools for 2014, 2015, and 2016. The differences were significant for HCPS and SCS in all three years for both of the two types of LIM comparisons we made: comparing high-LIM schools (i.e., schools with at least 80 percent LIM students) with all other schools in the site and comparing schools above and below the site-wide median percentage of LIM students. In Aspire, none of the six comparisons (three years, two LIM specifications) yielded a significant difference, and, in fact, in 2014 and 2015, the two groups of schools within Aspire had very similar percentages reporting a benefit to the school.

difference is in the expected direction, but even at high-LIM schools, not all leaders perceived a benefit; the incentive might not have been sufficient to attract teachers.[17] As one HCPS respondent from a high-need school stated in an interview, "It's hard to find good teachers to come and work here when they can go to another school with less stressful situations."

According to SLs, the addition of effectiveness as a criterion in transfer decisions had positive or neutral effects. In HCPS, which instituted a policy to consider TE in addition to seniority in transfer decisions, about half the SLs who were aware of this policy said that their school had benefited. In SCS, which followed the state policy change to encourage consideration of effectiveness and eliminate consideration of seniority as a factor in transfer, about 70 percent of the SLs who were aware of the policy said that their school had benefited. Nearly all the remaining SLs in both sites reported a neutral effect.

Over time, fewer SLs in the districts reported that they lost good teachers or were forced to accept particular teachers because of personnel policies. We also asked SLs how much they agreed that, "More often than is good for my school, good teachers are forced to leave my staff because they are 'bumped' due to seniority or other policies." Less than half of the SLs in HCPS and SCS agreed that good teachers were bumped because of policy, and there was a notable decrease over time in the percentages agreeing with the statement in all three districts (see Figure 4.5).[18] The percentages agreeing were higher in PPS than in the other districts, which is consistent with the furlough and transfer policies in the district during the period. In the CMOs in which mandatory transfers did not occur, nearly all SLs disagreed with the statement or indicated that it did not apply.

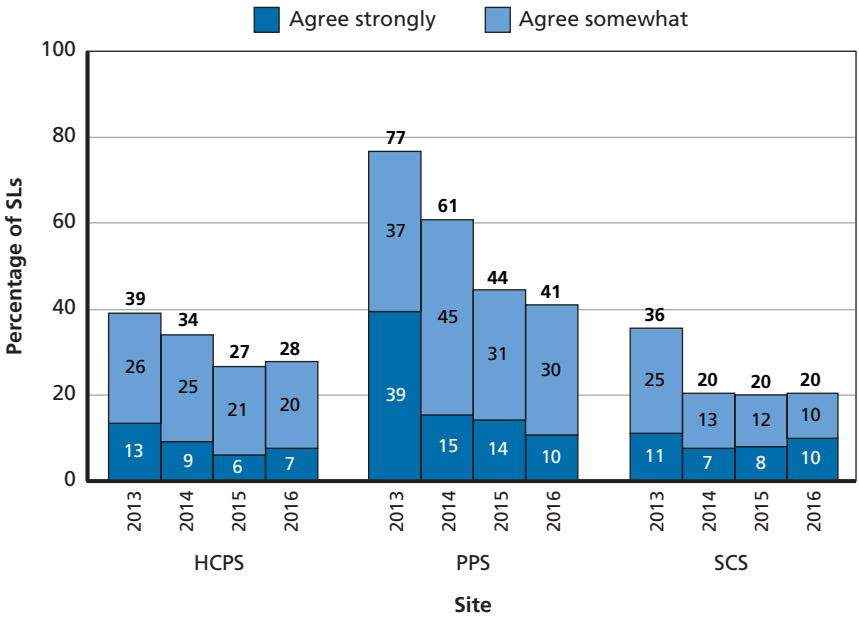
In addition, although most SLs in the three districts reported having experienced forced transfers *into* their school, in both SCS and

---

[17] The definition of *high LIM* used in the comparisons might also not exactly match sites' designations of which schools were considered high-need schools.

[18] In all three sites, the decline from 2013 to 2014 in the percentage agreeing was significant. In HCPS and PPS, the decline from 2014 to 2015 was also significant. In no site was the change from 2015 to 2016 significant.

Figure 4.5
District SLs' Agreement That, "More Often Than Is Good for My School, Good Teachers Are Forced to Leave My Staff Because They Are 'Bumped' Due to Seniority or Other Policies," Springs 2013–2016



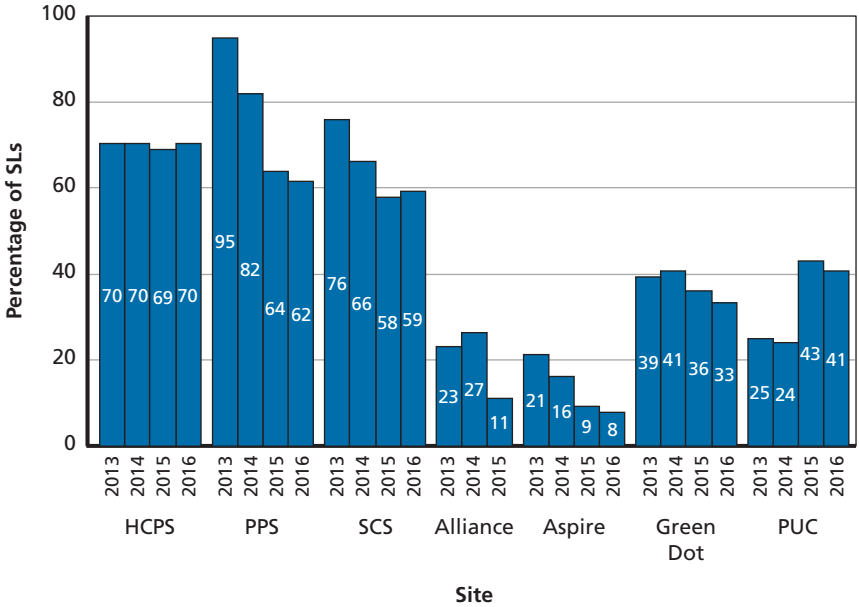
NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly." The question as it appeared in the survey also had a "not applicable" option, but we coded responses of "not applicable" as missing in the calculation of the percentages shown. We do not show CMOs because of large percentages of "not applicable" responses.

RAND *RR2242-4.5*

PPS (but not HCPS), the percentage reporting such transfers declined from 2013 to 2016—that is, there was a decrease in the percentage of SLs indicating that their school had teachers who had transferred "from elsewhere in the district who were assigned to my school based on district policy (for example, 'pool' or 'surplused' teachers)" (see Figure 4.6).

In the three districts, where transfer is broadly applicable, most SLs were satisfied with the performance of transferring teachers, and their satisfaction appears to have improved during the initiative. The changes were gradual and not enormous but are noticeable nonetheless

Figure 4.6
SLs Reporting That Their School Had Teachers Assigned to the School Based on Site Policy, Springs 2013–2016



NOTE: We asked this question only of SLs who reported that their schools had at least one new teacher in the current school year. We inadvertently did not ask the question of Alliance SLs in 2016.

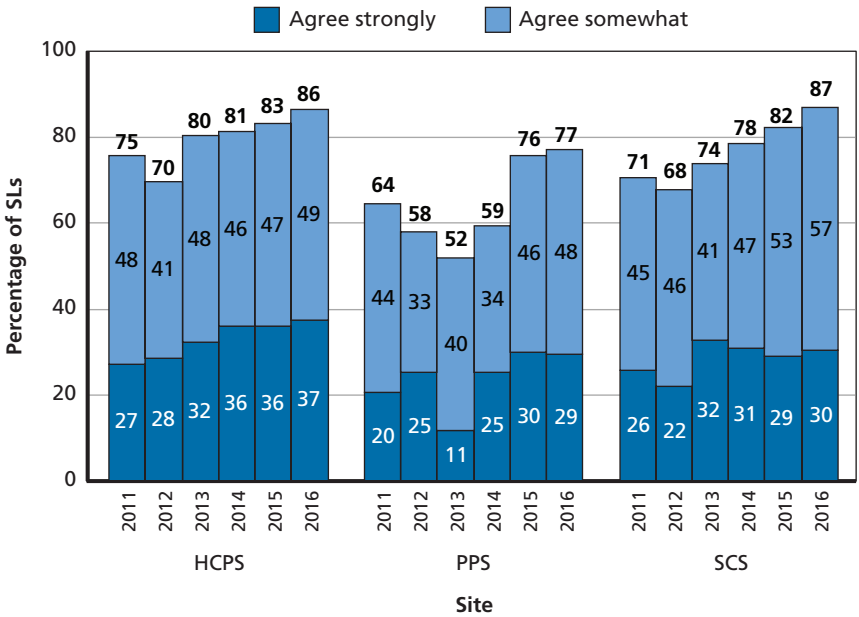
RAND *RR2242-4.6*

(see Figure 4.7). In all three districts, the increase in the percentage agreeing from 2011 to 2016 was statistically significant.

Teacher Outcomes Associated with Recruitment, Hiring, Placement, and Transfer Policies

We cannot isolate the effects that the sites' recruitment, hiring, placement, and transfer policies had on the overall effectiveness of teachers or on changes in student performance. In Chapter Ten, we examine trends over time in the effectiveness of novice teachers compared with

Figure 4.7
District SLs' Agreement That, Overall, They Were Satisfied with the Performance of Teachers Who Transferred into Their School from Elsewhere in the District That Year, Springs 2011–2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly." The question as it appeared in the survey also had a "not applicable" option, but we coded responses of "not applicable" as missing in the calculation of the percentages shown. We do not show the CMOs because of large percentages of "not applicable" responses.

RAND *RR2242-4.7*

that of experienced teachers, and, in Chapter Thirteen, we present evidence about the initiative's impact on student outcomes.

Summary

Education research suggests that sites can improve teacher recruitment, hiring, transfer, and placement in a variety of ways, including better marketing and outreach, more-timely identification of vacancies, earlier and more-efficient application and screening procedures, better infor-

mation for candidates (e.g., clearer, more-detailed job postings) and SLs (e.g., candidate screening tools), reducing or eliminating transfer preferences given to experienced teachers, and more-supportive school environments. In addition, sites can improve their long-term prospects by engaging in grow-your-own programs, such as targeted residencies. Although the IP sites did not receive explicit guidance about how to improve their recruitment and hiring efforts, they made changes that were consistent with many of the recommendations from the literature.

All sites modified their recruitment, hiring, placement, or transfer policies to some extent during the IP initiative, primarily by broadening recruitment efforts, automating the application and screening process, and adopting better screening or interviewing tools. In some sites, modifications were made in response to changes in the local or state context. In addition, some sites enacted mentoring programs to support newly hired teachers, and some developed residency programs and partnerships with local colleges and universities to increase and diversify the pool of future teacher candidates. Despite the changes, the sites continued to have difficulty attracting effective teachers to high-need schools, and persistent teacher turnover remained a problem in the CMOs.

SLs were generally aware of their sites' policies and thought that the hiring process worked fairly well; perceptions of the hiring process improved in PPS and SCS over time. Most SLs perceived residency programs as having beneficial effects, and many SLs in high-need schools perceived policies that offered teachers incentives to transfer to such schools as being beneficial. In addition, over time, fewer and fewer SLs reported experiencing the loss of good teachers from, and the assignment of unwanted teachers to, their schools.

CHAPTER FIVE

Tenure and Dismissal Policies

> There have always been systems in place to remove bad or ineffective teachers. It takes administrative leadership and diligence to do what is best for their school and district.
>
> —teacher, written comment in survey

In Brief

Did the sites reform their tenure and dismissal policies in an effort to reduce the number of low-performing teachers? How did teachers respond to their sites' policies?

Our analysis indicates that the three districts set specific criteria based on their new evaluation systems to identify low-performing teachers who might be denied tenure, placed on improvement plans, or considered for dismissal or nonrenewal of their contracts. In contrast, the CMOs, which did not offer tenure, took the new evaluation results into consideration but, except for Green Dot, did not establish specific criteria that triggered action for teachers with performance problems. Across the districts and CMOs for which data are available, about 1 percent of teachers were dismissed for poor performance in 2015–2016, the most recent year for which data are available. Our survey and interview data indicate that teachers in the districts and CMOs were aware of their sites' policies, and the percentage of teachers reporting that they understood their sites' dismissal policies increased over time. On the other hand, the percentage of district teachers reporting that evaluation results should be used to make tenure deci-

sions declined over time. Few teachers in either the districts or the CMOs reported that they were worried about being dismissed.

Introduction

This chapter reviews the IP sites' approaches to reforming tenure, as well as policies regarding teacher dismissal. One of the central premises underlying the IP initiative was that sites could improve the quality of the teaching workforce if they based tenure and dismissal decisions on systematic teacher-evaluation data. In this chapter, we examine the arguments that educators and researchers made for reforming tenure and dismissal policies, the evidence base supporting these reforms, the tenure and dismissal policies and practices the sites adopted and implemented, and teacher and SL attitudes toward the idea of basing key decisions about tenure and dismissal on evaluation results. (In Chapter Three, we describe the evaluation system itself, and, in Chapter Eleven, we discuss the impact that tenure policies and other IP levers had on teacher retention.)

Best Practices for Tenure and Dismissal

Teacher tenure is a common but not universal feature of public education in the United States. Most states have laws specifying that, after teaching for a probationary period of typically two or three years, teachers who meet their districts' standards will be awarded tenured status.[1] Tenured teachers can generally be dismissed only for specified reasons, such as criminal activity, immoral behavior, unprofessional conduct, or incompetence (poor performance). Dismissal involves various procedural protections, typically involving a right to a hearing, with a specified appeal process. (See Bireda, 2010.)

---

[1] In 2015, according to the National Council on Teacher Quality, one state required one probationary year before receiving tenure, three states required two, 31 states required three, five states required four, six states required five, and three did not offer tenure. Two states had no tenure policy.

Tenure became part of the public education landscape more than 100 years ago, initially as a mechanism to protect teachers from potential interference by school administrators (McGuinn, 2010). Advocates of tenure feared that teachers might be fired for speaking out against district policy or that administrators might attempt to fill positions with political supporters (Baratz-Snowden, 2009). In addition, they thought that the job security offered by tenure might help compensate for working conditions or salaries that might not otherwise be competitive (Baratz-Snowden, 2009).

State law typically requires that teachers be given the opportunity to remediate performance problems or other behavior prior to being given notice of dismissal (Bireda, 2010). States generally require a specific remediation period, along with documentation of failure to improve, prior to initiating the dismissal process (Chait, 2010). It is common for districts to place teachers with performance that might lead to dismissal on improvement plans, specifying goals and providing a timeline and specific supports (for example, coaching). There is some evidence that being identified as low performing can lead some teachers to leave teaching voluntarily, reducing the need for formal dismissals (Dee and Wyckoff, 2015).

In practice, in most districts in the United States, almost all teachers are awarded tenure at the end of the probationary period, and dismissal of tenured teachers for performance-related reasons has traditionally been quite rare. Data from the 2011–2012 Schools and Staffing Survey (the most recent year for which data are available) indicate that, on average, only 1.1 teachers per district were dismissed for poor performance in 2010–2011—less than one percent of the teacher workforce (NCES, undated [a], Table 8).[2] Of these, 1.0 were nontenured teachers, and 0.2 were tenured.[3] Because the great majority of teachers

[2] NCES, undated (a), Table 8 (average number of teachers per public school district in 2011–2012, average number of teachers per district in the previous year [2010–2011] who were dismissed or did not have their contracts renewed for any reason, including poor performance, by teachers' tenure status and state), reports an average of 187 teachers total per district, which implies that the overall dismissal rate averaged about 0.6 percent.

[3] Because of rounding, the number of probationary and tenured teachers dismissed does not sum to the total (1.1).

are tenured, this implies that the rate of dismissal for tenured teachers is far below 1 percent.[4] Some have argued that the very low rate of dismissal of tenured teachers for performance reflects the cost in time and resources of the various hearings and appeals involved. It can cost a district as much as $250,000 to complete the process of dismissing a tenured teacher (Goldhaber and Walch, 2016).

Some scholars, policymakers, and activists have argued for reform of the tenure process by citing the fact that few teachers are denied tenure or dismissed for performance once they have received tenure. Four types of reforms have been suggested (McGuinn, 2010):

- *Extend the probationary period.* This might provide the opportunity for administrators to gather more information on a teacher's effectiveness—including both data on value added and classroom observations—prior to making a decision. If tenure is awarded in the third year, which is common, there are only two years of data on which to rely in making a decision.
- *Increase the standard for moving from probationary status to tenure.* Currently, nearly all teachers who reach the end of the probationary period are awarded tenure (Weisberg et al., 2009). Some reformers argue that, to be awarded tenure, new teachers should meet a defined standard of effectiveness—for example, at least the average performance of teachers in the district.
- *Require teachers to maintain a satisfactory level of measured effectiveness to retain tenure.* Some argue that tenure can reduce teachers' motivation to excel. Requiring teachers to maintain a reasonable level of performance might help to sustain their motivation.

---

[4] NCES, undated (a), Table 8, does not report the average number of nontenured and tenured teachers per district, so we cannot compute precise dismissal rates for these two groups. But data from the 2011–2012 Schools and Staffing Survey teacher survey indicates that 55.6 percent of teachers were tenured, 14.7 were nontenured, and 29.7 percent were in schools that did not offer tenure. As indicated earlier, NCES reports an average of 187 teachers per district. Thus, the average number of teachers without tenure was 14.7 + 29.7 = 44.4 percent of 187, or 83. As reported by NCES, the number of nontenured teachers dismissed for performance each year was 1.0; thus, the percentage of nontenured teachers dismissed was 1.0 ÷ 83 = 1.2 percent. The rate for tenured teachers was about 0.2 percent.

- *Streamline the dismissal process for tenured teachers.* Some argue that the current protections in state law make dismissing tenured teachers nearly impossible unless they have demonstrated egregious failures of performance; reducing the number of hearings or appeals might reduce the cost of terminating low-performing teachers.

Other reformers have argued that tenure should be eliminated altogether, replaced by annual teaching contracts. Under this arrangement, teachers would be evaluated annually, and teachers meeting a defined performance criterion rehired (Finn, 2017).

Despite these arguments in support of tenure and dismissal reform (to which we refer jointly as *tenure reform*), it is not a foregone conclusion that reform would have positive effects, even if it led to an increase in the exit rate for low-performing teachers. For example, although reformers argue that raising standards for earning or maintaining tenure would increase teacher motivation to excel, it might instead be demoralizing. In addition, increasing the number of dismissals would exacerbate teacher turnover. There is some evidence that teacher turnover has a negative impact on student achievement (Ronfeldt, Loeb, and Wyckoff, 2013), in part because teachers new to teaching are, on average, less effective than more-experienced teachers (see Chapter Ten) and in part because, when one teacher replaces another, the new teacher must spend time to learn the school's curriculum, culture, and students. To have a net positive effect, tenure reform would need to ensure the replacement of low-performing teachers with teachers who are higher in effectiveness, on average.

Given the continuing policy interest in tenure reform, it is perhaps surprising that the empirical literature bearing on tenure is quite small. Some studies have investigated the effects that changes in tenure policy have on teacher exits—particularly the exits of low-performing teachers. For example, Strunk, Barrett, and Lincove found that, after Louisiana eliminated tenure, Louisiana teachers' exits from their schools increased significantly (Strunk, Barrett, and Lincove, 2017). Other studies have examined the effects of introducing systematic teacher evaluation. Loeb, Miller, and Wyckoff examined a tenure reform in

New York City, in which the central office provided clearer guidance to principals on tenure decisions and permitted principals to extend the probationary period for teachers about whom they were uncertain (Loeb, Miller, and Wyckoff, 2015). The introduction of the reform led principals to extend the probationary period for about 40 percent of low-performing teachers, and these "extended" teachers left their schools at significantly higher rates than teachers awarded tenure. Furthermore, the teachers who replaced the exiting teachers were higher in average achievement.

Other studies have examined the impact on teacher attrition of providing principals or teachers with more systematic data on teacher performance. For example, Cullen, Koedel, and Parsons, in a study in the Houston Independent School District, found that providing information to principals on TE led to an increase in the exit rate of low-performing teachers (Cullen, Koedel, and Parsons, 2016). Sartain and Steinberg, in a randomized experiment in which some Chicago schools were assigned to provide teachers with feedback based on systematic classroom observations, found that low-performing teachers in the treatment condition were more likely than control teachers to exit in the year following the intervention (Sartain and Steinberg, 2016). In another experimental study, conducted in the New York City schools, Rockoff, Staiger, Kane, and Taylor provided treatment principals with information on their teachers' value added (Rockoff et al., 2012). At the end of the year, the exit rate of low-performing teachers was higher in treatment schools than in control schools.

In a related study, focusing on the District of Columbia's IMPACT set of reforms, including providing systematic teacher-evaluation data to principals and teachers, Adnot and her colleagues examined the average achievement of teachers who entered D.C. public schools to replace those who left (Adnot et al., 2017). They found that, on average, teachers who left their schools were replaced by teachers with significantly higher overall effectiveness ratings and higher average value added (significant in math but not reading).

Apart from studies focused on the effects that tenure and dismissal reform have on teacher exits, a few studies have examined how the introduction of reform affects teacher motivation. For example,

Goldhaber, Hansen, and Walch found that extending the probationary period from three to four years in North Carolina and from two to three years in Washington reduced teacher absenteeism among probationary teachers, which might be viewed as a proxy for motivation (Goldhaber, Hansen, and Walch, 2016). Similarly, Jacob found that a Chicago policy that gave principals the flexibility to dismiss probationary teachers for any reason without the typical documentation and hearing process reduced the absenteeism of probationary teachers (Jacob, 2010).

Although these studies suggest that tenure reform might increase the exit rate of lower-performing teachers, and these teachers might be replaced by those higher in effectiveness, most of the studies do not examine the overall net impact on student achievement. Several simulation studies have attempted to assess the fraction of teachers who would need to be replaced (through tenure reforms or other policies) to have a policy-relevant impact on overall student achievement. In one of the first studies of this kind, widely cited at the time the IP initiative got under way, Hanushek found that dismissing the lowest 5 to 10 percent of the teaching force would have a measurable positive impact on student achievement (Hanushek, 2009). For purposes of the simulation, Hanushek assumed that the deselection would be based on a perfectly reliable measure of TE. Subsequent studies have explored the implications of basing dismissal decisions on measures with typical reliability, varying the number of years of TE data available in making the dismissal decision. In general, these studies have focused on the fraction of novice teachers dismissed at the end of the probationary period because this is a decision point of particular policy interest. Staiger and Rockoff, for example, found that, if dismissal decisions are based on one year of VAM data, the maximum gain in overall achievement would be obtained by dismissing a very high proportion of novice teachers—as many as 80 percent, depending on assumptions (Staiger and Rockoff, 2010). But meaningful gains would be obtained by dismissing fewer teachers, and the proportion could be lowered if more-reliable information on performance were available. Rothstein, in a simulation that took teacher supply into account, concluded that, if decisions are based on two years of VAM data, an exit

rate of the lowest 20 percent of novice teachers would lead to a noticeable improvement in student achievement, but an increase in teacher salaries would be required to attract the additional hires required under this policy (Rothstein, 2015). (For additional analyses, see Goldhaber and Hansen, 2010, and Winters and Cowen, 2013.)

These simulation results are based on assumptions that reflect teacher career patterns in regular public school districts; in particular, they typically assume a potential teaching career of 30 years. Thus, decisions made about whether to retain novice teachers after their first few years of teaching can have long-term consequences for the teacher distribution. It is not clear whether the results would be similar in charter schools, in which teaching careers are typically much shorter.[5]

Taken together, the empirical and simulation work suggests that tenure reform and changes in dismissal policies have the potential to lead to improvements in student outcomes, at least in regular district schools. But achievement of the desired outcomes would require setting the bar for receiving tenure considerably higher than it is in current practice to have an appreciable impact on student achievement. In the following section, we examine the tenure and dismissal policies in the IP sites when the initiative got under way and the reforms on which the sites embarked.

IP Tenure and Dismissal Policies

All seven IP sites were expected to implement policy changes that would increase exits of low-performing teachers. In the three districts, this included policies for tenure, as well as dismissal. In the CMOs, however, only the latter applied because the CMOs did not offer tenure to their teachers. Table 5.1 summarizes tenure and dismissal policies

[5] The literature on teacher attrition in charter schools is limited, but the available evidence indicates that the exit rate for low-performing teachers in the first few years of teaching is similar in charters and regular district schools, suggesting that charters do not, on average, apply higher standards than regular district schools do for retaining teachers. See Cowen and Winters, 2013.

Table 5.1
District Tenure and Dismissal Policies Through 2015–2016

Policy Area	HCPS	PPS	SCS
Requirements for tenure	**Prior to July 2011:** Three years of effective teaching in the district	**Prior to July 2011 and ongoing:** Six semesters of satisfactory performance. The policy predated the IP reform, but the definition of *satisfactory performance* changed under the initiative.	**Prior to July 2011:** Completion of six semesters (three years) of satisfactory performance
	July 2011: Florida abolished tenure for newly hired teachers.		**For teachers hired in and after July 2011:** Completion of a 45-month (five-year) probationary period with TE scores of 4 or 5 in the past two years. Once tenure is granted, a TE rating of 3, 4, or 5 is required to maintain tenure.
	2011–2012 through 2015–2016: HCPS teachers were hired as probationary and promoted to nonprobationary status in lieu of tenure after three consecutive years of teaching in the district, provided they had a TE rating of E or higher each year and without any reservation by an administrator.[a]		
Number of tenured teachers in 2015–2016 out of total number of teachers	8,407 teachers (about 50% of HCPS teachers) had tenure grandfathered under the old system, and 7,116 (about 42%) had nonprobationary status.	Approximately 1,596 (about 85% of all PPS teachers)	No information available

Table 5.1—Continued

Policy Area	HCPS	PPS	SCS
Number of teachers eligible for tenure who received it in 2015–2016	No information available	52 teachers eligible; all received tenure, out of a cohort of 109[b]	No information available
Circumstances leading to placement on improvement plan	**Prior to 2010–2011:** Low rating under the previous performance rating system	**Prior to 2011–2012:** Principal discretion based on observation	**Prior to 2012–2013:** No information available
	2010–2011 and ongoing: Rating of U or NI on the TE measure	**2011–2012 through 2012–2013:** Poor performance on RISE	**2012–2013 and ongoing:** Rating of 2 or lower on two or more observation rubric indicators (out of seven scored indicators) or overall TEM score of 1 or 2
		2013–2014 and ongoing: Rating of NI or F on the TE measure in the previous year. Principals have discretion to place a pretenure teacher on a plan based on observation results.	
Number of teachers on improvement plans in 2015–2016	135 (less than 1% of HCPS teachers, data from 2016–2017)	55 teachers (about 3.2% of all PPS teachers)	No information available

Table 5.1—Continued

Policy Area	HCPS	PPS	SCS
Circumstances leading to dismissal for poor performance	**Prior to 2010–2011:** Persistent low ratings under previous system without evidence of growth, at principal's discretion	**Prior to July 2011 and ongoing:** For tenured teachers, two consecutive U ratings; for pretenured teachers, one U rating. The policy predated the IP reform, but the definition of U or S changed under the initiative.	**Prior to 2012–2013:** No information available
	2010–2011 and ongoing: For tenured and nonprobationary teachers, two consecutive U or three consecutive NI ratings. Tenured teachers put up for dismissal can appeal to the board. For probationary teachers, one U or NI rating, but a principal can opt to put any teacher on an improvement plan instead of dismissing him or her.		**2012–2013 and ongoing:** For teachers hired after the 2010–2011 year, principals can dismiss nontenured teachers with a TE score of 1 or 2 and lack of improvement over the course of the year with approval from the HR office. Teachers hired after 2010–2011 can also, once tenured, lose tenure for poor performance as described earlier. Once tenure is lost, a teacher can be dismissed for poor performance at the end of the year.

Table 5.1—Continued

Policy Area	HCPS	PPS	SCS
Number of teachers dismissed for poor performance in 2015–2016	84 nonrenominated; 57 resigned in lieu of being nonrenominated (less than 1% of all HCPS teachers)	One of the 55 on improvement plans (less than 1% of all PPS teachers)	No information available

SOURCES: Annual interviews with central-office staff in each of the IP districts.

[a] Between 2011–2012 and 2015–2016, HCPS granted nonprobationary teachers protections similar to those provided under tenure in the past.

[b] PPS hires teachers in cohorts (i.e., all the teachers hired in a given school year). The cohort that was eligible for tenure in 2015–2016 initially had 109 teachers, 57 of whom left the district before earning tenure.

in the districts, and Table 5.2 summarizes the policies in the CMOs; additional details are available in Appendix E.

District Tenure and Dismissal Policies

In 2009–2010, when the three districts were awarded IP grants, each of the states in which the districts were located (Florida, Pennsylvania, and Tennessee) had a three-year probationary period for new teachers, prior to tenure. In Pennsylvania, the state policy did not change appreciably over the period of the IP initiative. Both Florida and Tennessee, however, made major changes in the state law, both of which took effect in July, 2011, just as the IP initiative was picking up steam:

- In Florida, the state abolished tenure for newly hired teachers, instead putting teachers on an annual contract and requiring teachers to be evaluated each year. In response to the state change, HCPS retained a three-year probationary period and considered a teacher nonprobationary (rather than tenured) if he or she met the district's performance standards at the end of a probationary period (see Table 5.1). A nonprobationary teacher was subject to annual evaluation and nonrenewal of the contract if he or she failed to meet specified effectiveness rating standards.
- In Tennessee, the probationary period was extended to five years, and each teacher was required to obtain a rating of at least E to maintain tenured status.[6] SCS followed this new approach to awarding tenure (see Table 5.1).

In addition to the changes made in HCPS and SCS in response to changes in state law, all three districts established new criteria based on their new evaluation systems that teachers were required to meet to achieve tenure (or, in HCPS, nonprobationary status). In general, the sites set the equivalent of one or more years of effective performance as the standard to move from probationary to tenured (or nonprobationary) status, and, because most teachers met this standard each year (as

[6] According to the Tennessee policy, a tenured teacher who receives a rating of less than effective for two consecutive years is returned to probationary status.

Table 5.2
CMO Tenure and Dismissal Policies Through 2015–2016

Policy Area	Alliance	Aspire	Green Dot	PUC
Requirement for tenure	Not applicable. Employment agreements are for one year and renewed or not renewed each year.	Not applicable; does not offer tenure. All teachers are hired on an at-will basis.	Not applicable; does not offer tenure. All teachers are hired on an at-will basis.	Not applicable; does not offer tenure. All teachers are hired on an at-will basis.
Number of tenured teachers in 2015–2016 out of total number of teachers	Not applicable	Not applicable	Not applicable	Not applicable
Number of teachers eligible for tenure who received it in 2015–2016	Not applicable	Not applicable	Not applicable	Not applicable

Table 5.2—Continued

Policy Area	Alliance	Aspire	Green Dot	PUC
Circumstances leading to placement on improvement plan	**Prior to 2012–2013:** No information available	**Prior to 2012–2013:** No information available	**Prior to 2012–2013:** No information available **2012–2013:** Pilot test of new criteria	**Prior to 2012–2013:** No information available
	2012–2013 through 2015–2016: Poor performance based on classroom observations, professional responsibilities, or overall effectiveness rating	**2012–2013 and ongoing:** Poor performance as judged by SL based on formal and informal observations, the results of interim student assessments, and stakeholder survey results	**2013–2014 through 2014–2015:** Established observation rubric score levels for placement on improvement plan **2015–2016 and ongoing:** Pilot test lowered the minimum required scores to place on improvement plan.	**2012–2013 and ongoing:** Poor performance as judged by SL based on formal and informal observations, the results of interim student assessments, and stakeholder survey results
Number of teachers on improvement plans in 2015–2016	No information available	No information available	Typically three to eight teachers per year	Estimated to be less than 3%

Table 5.2—Continued

Policy Area	Alliance	Aspire	Green Dot	PUC
Circumstances leading to dismissal for poor performance	**Prior to 2012–2013:** No information available	**Prior to 2012–2013:** No information available	**Prior to 2012–2013:** No information available	**Prior to 2012–2013:** No information available
	2012–2013 through 2015–2016: Continued poor performance after efforts to improve, at discretion of SL	**2012–2013 and ongoing:** Continued poor performance after efforts to improve; at discretion of SL	**2012–2013 and ongoing:** Continued poor performance after efforts to improve; at discretion of SL	**2012–2013 and ongoing:** Continued poor performance after efforts to improve; at discretion of SL
Number of teachers dismissed for poor performance in 2015–2016	No information available	(Estimate) 1–2% of teachers	No teachers were dismissed.	Three or four out of 272 teachers (about 1 percent)

SOURCES: Annual interviews with central-office staff in each of the IP CMOs.

shown in Chapter Three), most new teachers received tenure at the end of their probationary period. For example, in PPS in 2015–2016, all 27 teachers who reached the end of the probationary period that year were awarded tenure. (Comparable data are not available for HCPS and SCS.)

All three districts also set criteria based on the teacher-evaluation system to identify teachers who should be placed on improvement plans, the first step in moving toward dismissal for both tenured and nontenured teachers. These criteria ordinarily involved one or two years of performance below the effective level. Again, because most teachers received ratings of E, few were identified for improvement plans. In 2015, 55 teachers in PPS (about 3.2 percent of all teachers in the district) were on improvement plans. Fewer teachers were eventually dismissed—11 of the 55 teachers on improvement plans—less than 1 percent of the teacher workforce. In HCPS, similarly, less than 1 percent of teachers were dismissed.[7]

CMO Tenure and Dismissal Policies

Unlike the three districts, none of the CMOs offered tenure prior to or during the initiative. Instead, teachers were hired on an at-will basis. Once hired, teachers in the CMOs were retained or terminated at the

---

[7] The information on the number of teachers dismissed for low performance was provided by central-office staff during interviews or subsequent communications. To check the reliability of the information, we examined data from teacher administrative records, which we discuss in Chapter Twelve. Although the administrative records allow us to estimate the number of teacher separations (exits), they do not indicate whether teachers exited because of dismissal for low performance or for voluntary reasons, so we cannot conduct an exact comparison. But by linking the empirical data on teachers' exits to teachers' effectiveness ratings, we can infer the approximate number who might have left because of low performance. The overall results are consistent with those reported earlier from our interviews. For example, in HCPS, for 2014–2015 (the most recent year for which we have data), 89 teachers who were rated U (the lowest level) left the district, and 88 who were rated NI left the district, a total of 177. This is 36 higher than the 84 + 57 = 141 who were not renominated or who resigned because of low performance according to our district contacts (see bottom row of Table 5.1). The difference might be accounted for by routine teacher exits (e.g., departures due to illness or relocation). The administrative record data indicate a strong relationship between teacher ratings and exits. For example, 72 percent of the teachers rated U left, compared with 34 percent of those rated NI and 8 percent of those rated HE.

discretion of the principal. (Table 5.2 summarizes tenure and dismissal policies in the CMOs, and we provide additional details in Appendix E.) Although the principals drew on the formal evaluation system developed as part of the IP initiative in assessing teacher performance, they combined this information with informal observations rather than establishing specific criteria analogous to those established by the districts.

Like in the districts, all four CMOs provided improvement plans for teachers identified as low performing. In Alliance, Aspire, and PUC, the process leading to being placed on an improvement plan was not formalized. But in Green Dot, the one unionized CMO, the steps in the improvement plan were specified as part of the union contract from 2013–2014 on, although those steps were not mandatory. In all four CMOs, given the at-will contracts, principals could also respond to poor performance by deciding not to renew a teacher's contract. Nonrenewal for low performance in the CMOs can be viewed as the analog of dismissal in districts with tenure.

Implementation Challenges

Because of the job security implications for teachers, changes to tenure and dismissal policies are particularly likely to raise potential implementation challenges. Our interviews with central-office administrators, principals, and teachers highlighted three issues that sites faced.

The sites struggled to balance reforms to tenure and dismissal policies designed to remove low-performing teachers with improvement policies designed to enhance the performance of all teachers. According to our interviews with central-office staff, all the sites tried to focus their IP work on the improvement of the effectiveness of the teaching workforce through the identification of low-performing teachers and provision of supports to help them to improve, rather than dismissal. This strategy was particularly salient in PPS, in which district budget constraints limited hiring of new teachers, even to replace teachers who left (whether voluntarily or not). But even in the other sites, the focus was largely on improving teacher quality

rather than dismissal. In HCPS, for example, although a teacher could be dismissed for one to three years of poor performance (depending on seniority), principals often opted to put these teachers on performance plans instead, in hopes of improving their performance. And in the CMOs, which, as charters, had the freedom to refrain from renewing teacher contracts at will, the emphasis nevertheless was on improving performance and retaining teachers who demonstrated a commitment to improving instructional practices.

The fact that teacher-evaluation results were used as the basis for tenure and dismissal decisions might have led some principals to avoid giving low observation ratings that would reduce a teacher's composite score.[8] This issue arose in PPS early in the rollout of the new composite measure of TE. In the pilot year (2012–2013), 14 percent of teachers received composite ratings in the lowest two performance levels. No stakes were attached to the ratings that year, but, if they had been, these lowest-rated teachers would have been put on improvement plans and would have had to improve the next year or be dismissed. In the next year, when stakes were attached to the ratings, the proportion of teachers rated at the lowest two levels on the composite measure fell to 3 percent. According to central-office staff, principal ratings on one observation component—3f, which measured student growth in nontested grades and subjects—largely accounted for the shift. When principals assigned ratings in the pilot year, they did not realize that 3f carried a lot of weight in the composite, enough to exert substantial influence on the overall rating and thus the consequences. In the next year, they gave far fewer low ratings on 3f, with the result that far fewer teachers were rated NI or F on the composite. In other words, central-office staff suggested that principals might have changed their approaches to scoring some components over time once they saw how the measures combined to produce the composite. In particular, some central-office staff told us that the change in the way

[8] This concern has also been raised in recent studies of teacher-evaluation systems. For example, Grissom and Loeb, in a study of an urban district, compared principals' high- and low-stakes ratings of teachers (Grissom and Loeb, 2017). They found that both high- and low-stakes ratings were largely positive, but the high-stakes ratings were more positive.

principals rated 3f in 2013–2014—as well as the way they scored the SLOs that replaced 3f in subsequent years—might have been partly because principals did not want to have so many teachers on improvement plans, which create a lot of work for the principal. These central-office staff members commented that, prior to the IP initiative, about 3 percent of teachers were on improvement plans and that principals might have felt most comfortable with a percentage in that range.

There were also concerns about inflation in the other districts. As one HCPS central-office administrator commented in 2016, "I would hate to see us slip back into a place where everybody got these great evaluations and they're not that great." However, this was less of a concern in the CMOs, in which principals typically have considerable discretion over whether a teacher is identified for improvement or not renewed and thus would be less likely to feel pressure to give higher ratings.

In most sites, the composite TE measure for a given school year could not be computed and released until the following fall, which complicated its use in tenure and dismissal decisions. In all sites except PPS and SCS, teacher scores on the composite measure of TE for each academic year were generally not available until the fall of the subsequent year, given the timing of the exams that generated the data (they occurred in the spring) and the time required to generate the needed measures of student achievement growth and combine them with observation and other scores.[9] For this reason, in making decisions about whether teachers should be renewed, principals tended to rely mainly on observation data. In addition, principals often found that the observation dimension scores and qualitative information were at least as useful as the formal composite scores in making decisions about improvement plans, tenure, and dismissal.

[9] PPS used VAM scores from a year prior, which allowed it to produce composite scores earlier. SCS issued a provisional TE measure at the end of the school year, which had current data for everything but the VAM scores and achievement measures; SCS central-office staff justified this based on indications that the VAM scores and achievement data tended not to change teachers' composite scores very much.

Teacher and SL Reactions to Tenure and Dismissal Policies

To examine teachers' and SLs' perceptions of sites' tenure and dismissal policies, we explore their input, from surveys and interviews, on four broad topics:

- awareness: Did teachers and SLs know about the tenure and dismissal policies?
- endorsement: Did teachers and SLs approve of the tenure and dismissal policies?
- fairness: Did SLs think that dismissal policies were fair?
- perceived effects: What types of effects did teachers and SLs think that the tenure and dismissal policies had had?

For each of the four topics, we begin by discussing tenure and then discuss dismissal. In the "Awareness" section, at the beginning of the discussion of dismissal, we also discuss improvement plans.

Awareness

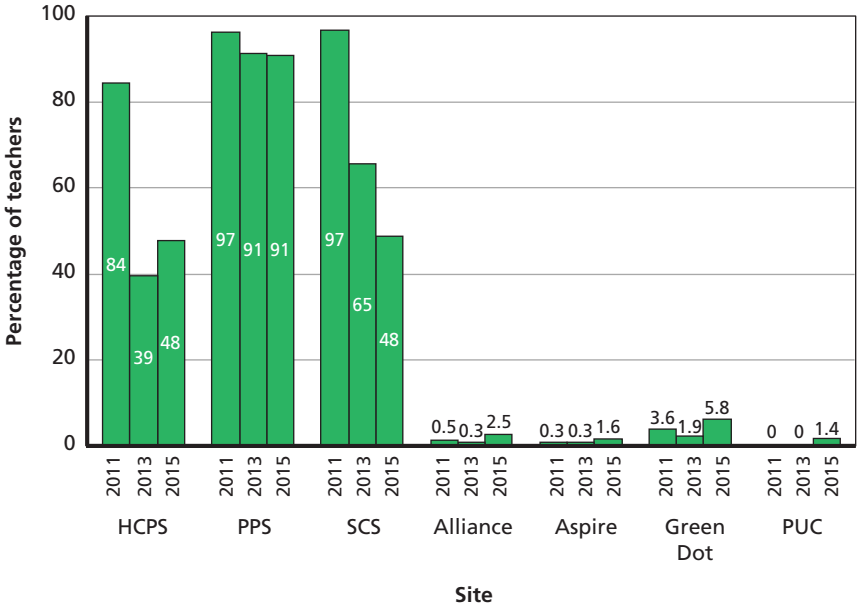
In general, teachers were aware of their sites' policies related to tenure and dismissal and of changes made to these policies during the period studied. This section examines tenure and dismissal in turn.

Tenure

In the three districts, teacher awareness of tenure policies reflected the changes made during the study period, as well as the complexity of the specific tenure policies in some of the sites. In the spring of 2011, all three districts awarded tenure. Accordingly, in the spring 2011 survey, most teachers in HCPS and nearly all teachers in PPS and SCS indicated that their site granted tenure (see Figure 5.1).

In HCPS, the percentage of teachers saying that the district offered tenure dropped dramatically after 2011, to 39 percent in 2013 and 48 percent in 2015. The fact that about half the teachers said that tenure was granted and about half did not suggests that there was

Figure 5.1
Teachers Reporting That Their Site Granted Tenure, Springs 2011, 2013, and 2015



NOTE: For 2011, omitted response categories are "no" and "don't know." For 2013 and 2015, omitted response categories are "no," "don't know," and "it used to, but it no longer does." On the teacher survey, we asked about tenure in only 2011, 2013, and 2015.

RAND *RR2242-5.1*

considerable ambiguity on the matter, and it is understandable why.[10] Although, as noted earlier, the state of Florida abolished tenure for newly hired teachers in the summer of 2011, HCPS maintained a distinction between probationary status (for teachers in their first three years of teaching in the district) and nonprobationary status (for teachers with more than three years of experience). Some teachers might

[10] "Don't know" was one of the options and was selected by sizable percentages of HCPS teachers in 2013 (10 percent) and 2015 (19 percent).

have considered this basically a continuation of tenure,[11] while others might have viewed the state's abolition of tenure as a rationale for not saying "yes" to the tenure question. Of the 48 percent who, in 2015, said that tenure *was* offered, most acknowledged that "the nature of tenure has changed in the past few years," while relatively few said that there had "not been any changes to tenure policies or the meaning of tenure in the past few years." (Earlier years' surveys did not offer these response options.) Thus, all told, the evidence suggests that most HCPS teachers were, in fact, aware of the tenure-related changes made in HCPS following 2011.

SCS also saw a drop in the percentage of teachers reporting that the district offered tenure: from 97 percent in 2011, to 65 percent in 2013, to 48 percent in 2015. Like in HCPS, this suggests considerable ambiguity about the matter, especially in 2015.[12] Because tenure officially existed in SCS in all three years in which the teacher survey included questions about tenure (2011, 2013, and 2015), it is not clear why only about half of the teachers indicated that there was tenure in 2015. It could be related to the change in policy allowing for *losing* tenure. As a result of this change, some teachers might have thought that tenure was no longer truly tenure, even if it was still *called* tenure.[13] Like in HCPS, most SCS teachers who said that there was tenure thought that the nature of tenure had changed in the past few years.

In PPS, in which tenure policy essentially did not change during the initiative, most teachers—though not quite as many as in 2011—continued to correctly say in 2013 and 2015 that the district granted tenure.

---

[11] A further source of ambiguity on the matter might have come from the survey itself, which offered a definition of *tenure* that specifically noted that "the term 'tenure' may or may not be used."

[12] Exactly like in HCPS, 10 percent of SCS teachers in 2013 and 19 percent of SCS teachers in 2015 answered "don't know."

[13] The definition of *tenure* given in the survey might have inadvertently encouraged this perspective: "By 'tenure,' we mean a status awarded to teachers, typically after a multi-year probationary period, in which they enjoy enhanced job security and perhaps other projections and benefits."

Virtually no teacher in any of the CMOs said that his or her site granted tenure, which is consistent with the fact that none of the CMOs does, in fact, offer tenure. Accordingly, all remaining findings related to attitudes about tenure are limited to the three districts.

In all three districts, the percentage of teachers reporting that they understood the criteria their districts used for awarding tenure peaked in 2013,[14] perhaps because the change to policy (in HCPS and SCS) was more recent at that time (see Figure 5.2). Teachers in all three districts also became more likely to say that tenure was now more difficult to earn than it used to be.[15] This was likely a reflection of a more rigorous evaluation process (see Chapter Three) than of a change to tenure per se.

Only in SCS did a majority of teachers—and SLs—think that *keeping* tenure had become more difficult for teachers (see Figure 5.3). This, too, is generally consistent with the three districts' policy changes. Although all three districts allowed for dismissal of experienced teachers based on performance, only SCS had a policy explicitly providing for "loss of tenure" (i.e., a return to probationary status) for low-performing tenured teachers. The fact that only teachers hired in or after July 2011 could lose tenure is a complicating factor in interpreting the SCS survey results. That could explain why the percentage agreeing with the statement is not even higher, for either teachers or leaders.

Dismissal

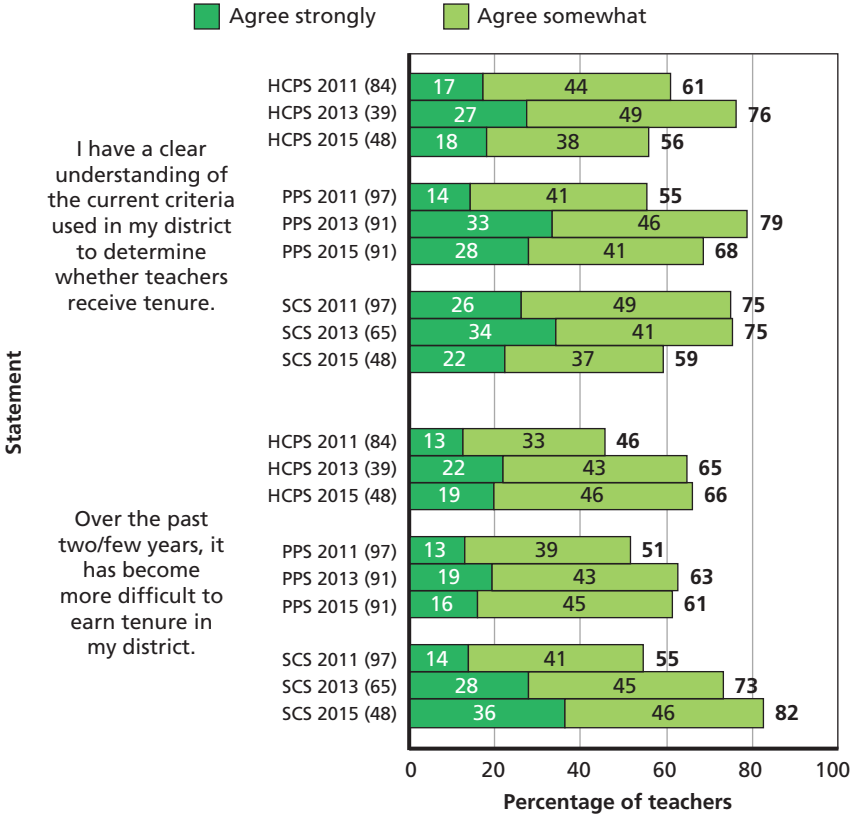
In all sites except one, a majority of teachers were aware that their evaluation results could be used to place them on improvement plans or to initiate dismissal. As described earlier, all seven sites (the districts and the CMOs) had provisions for placing teachers on improvement plans based either on the overall effectiveness measure or on components of it. In six of the sites, a majority of teachers—albeit a small majority—indicated awareness that their evaluation results could

---

[14] The increase in the percentage agreeing (strongly and somewhat) from 2011 to 2013 was statistically significant ($p < 0.05$) in HCPS and PPS. The decrease in the percentage agreeing from 2013 to 2015 was significant in all three sites.

[15] The increase in the percentage agreeing (strongly and somewhat) from 2011 to 2015 was statistically significant ($p < 0.05$) in all three sites.

Figure 5.2
Teachers' Agreement with Statements About Receiving Tenure, Springs 2011, 2013, and 2015

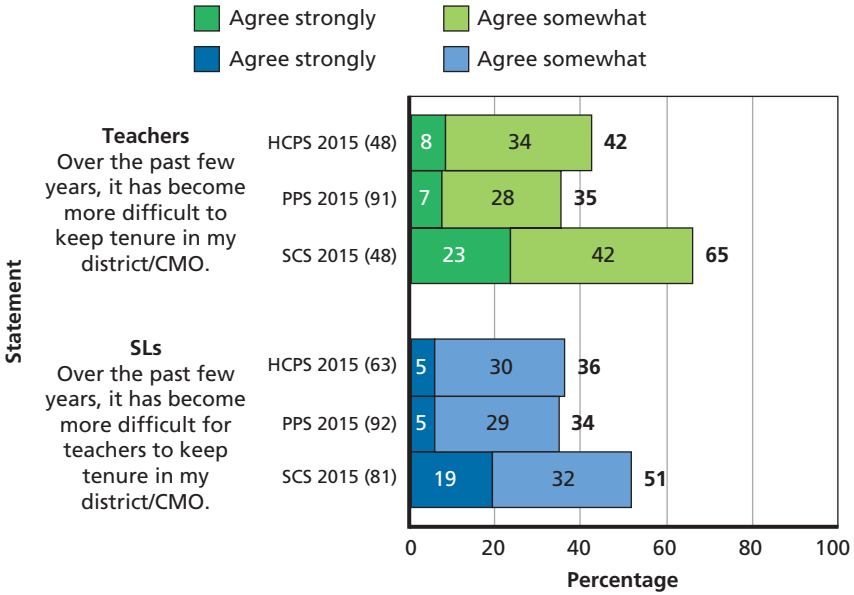


NOTE: Numbers in parentheses next to the site names are the percentages of teachers who indicated that their sites grant tenure. The percentages in the bars are among the teachers who indicated that their sites grant tenure. On the second item, the phrasing of the question changed; in 2011 and 2013, the question asked about "the past two years," whereas, in 2015, the question asked about "the past few years." Omitted response categories are "disagree somewhat" and "disagree strongly."

RAND *RR2242-5.2*

be used to determine whether they entered some type of improvement status (see Figure 5.4). Because, in theory, any teacher could do poorly enough on an evaluation to warrant such placement, the percentages indicating awareness of the *possibility* arguably should have been higher,

Figure 5.3
Teachers' and SLs' Agreement That Keeping Tenure Had Become More Difficult, Spring 2015



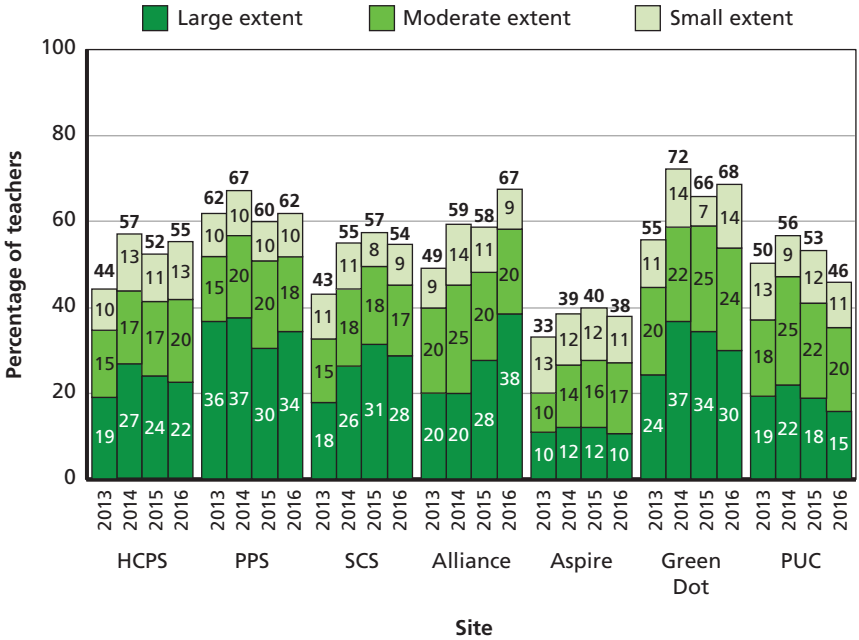
NOTE: Numbers in parentheses next to the site names are the percentages of respondents who indicated that their site grants tenure. The percentages in the bars are among the respondents who indicated that their site grants tenure. We asked this question only in 2015. Omitted response categories are "disagree somewhat" and "disagree strongly."

RAND *RR2242-5.3*

but because, in practice, very few teachers actually received ratings low enough to be put on improvement plans, it might not be surprising that relatively few teachers saw this as a possibility worth mentioning. Moreover, by the time teachers took the survey each spring, they might have already had some observations under their belt and could thus have had some sense of their evaluation results for the year.

The one site that differed from the others was Aspire, which had noticeably lower percentages of teachers saying that their evaluation results would be used to determine whether they entered into some type of improvement status. For the equivalent question on the SL survey, about the use of teacher-evaluation results for placing teachers

Figure 5.4
Teachers Reporting That Their Evaluation Results Would Be Used to a Large, Moderate, or Small Extent to Determine Whether They Would Enter into Some Type of Probationary Status (e.g., Employee Improvement Plan), Springs 2013–2016



NOTE: Omitted response categories are "not at all" and "don't know." We did not ask this question in 2011.

RAND *RR2242-5.4*

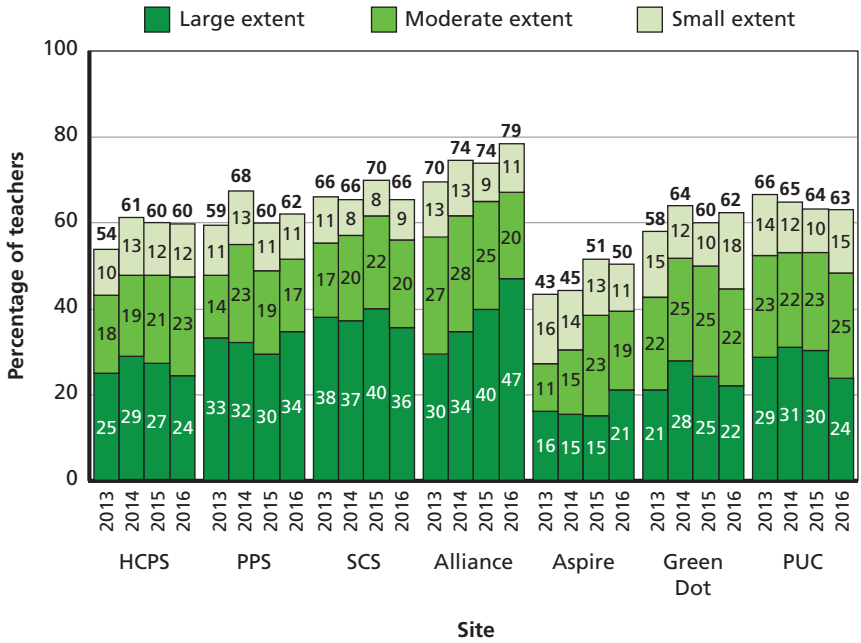
on improvement plans, Aspire had similarly low percentages compared with those of the other six sites. The low percentages might be because, although Aspire does use improvement plans, they are not mentioned in the site's TE guidebook, so teachers might not be aware that they are a possibility. In addition, according to central-office staff, the improvement plan process varies by school and region and is used more for teachers struggling with professionalism than for those struggling with instructional delivery.

As described earlier, HCPS and PPS both had provisions for dismissing teachers based on effectiveness ratings. In PPS, two consec-

utive poor ratings warranted dismissal for tenured teachers (one for pretenure teachers), while, in HCPS, two consecutive U or three NI ratings could be grounds for dismissal. SCS had a similar policy for teachers hired in or after July 2011. The CMOs documented eligibility for dismissal based on observations, the other evaluation measures, and teacher improvement efforts. However, dismissal was at the discretion of the principal and not mandated by evaluation results.

The percentages of teachers saying that their evaluation results would be used to determine whether they were "qualified to continue teaching" (see Figure 5.5) were higher than the percentages reporting the use of evaluation results for placement on an improvement plan in

Figure 5.5
Teachers Reporting That Their Evaluation Results Would Be Used to a Large, Moderate, or Small Extent to Determine Whether They Were Qualified to Continue Teaching, Springs 2013–2016



NOTE: Omitted response categories are "not at all" and "don't know." We did not ask this question in 2011.

RAND *RR2242-5.5*

five of the seven sites (see Figure 5.4). This suggests that the potential for dismissal might have been more salient in teachers' minds than the possibility of going into improvement status, even though dismissal typically required multiple years of poor performance, whereas placement into improvement status typically required only one poor performance rating.

The percentages of teachers reporting the potential for evaluation-based dismissal were highest in Alliance (in which the percentages not only were relatively high but also increased over time). We do not have a clear explanation for the high percentages in Alliance. SCS had the next-highest percentages of teachers reporting the potential for evaluation-based dismissal, followed closely by PPS and HCPS. Given that SCS teachers hired before July 2011 were not subject to dismissal for poor performance, the relatively high percentages in SCS—exceeding 60 percent each year—are somewhat surprising, especially in consideration of the similar percentages in PPS and HCPS, in which there was no such restriction. Like with the improvement plan finding, the percentages were lowest in Aspire.

Interviews, particularly in PPS and SCS, supported the survey finding that teachers might have been more concerned about dismissal than was warranted. Especially in the year or so after the new evaluation systems were rolled out, there was confusion among teachers about what would happen if they received low evaluation scores. Some teachers reported that others were confused, while some displayed confusion themselves. Relevant quotes from teachers include the following:

> Discomfort and rumor [say], if you get a 1 on anything, they'll fire you. (SCS teacher, 2012)

> [Observations are] stressful . . . . [They affect] whether or not you'll have a job the next year because of the scores. It's a lot of pressure. (SCS teacher, 2013)

> [I]f a teacher doesn't perform well, then that teacher is fired. That level of stress and anxiety is bleeding into Pittsburgh. Some teachers are feeling that the district is out to get them. Teachers say, "If I get two unsatisfactory [ratings] in a row, then I am

> fired!" I say to those teachers, "Do you know how difficult it is to get two unsatisfactory ratings?!" I think the district needs to do more to explain the reality to teachers. Instead of a bad game of telephone causing misinformation. . . . Misconceptions need to be corrected through more awareness and clear communication. (PPS teacher, 2013)

> And it may not be accurate, but, if you have a low VAM score, then they're going to kick you out. That . . . is the perception. (PPS teacher, 2015)

Over time, increasing percentages of teachers reported understanding the criteria for dismissal, but even by the spring of 2015, only about half indicated that they understood the criteria. From 2011 to 2015, every site had an increase in the percentage of teachers saying that they had a clear understanding of the criteria used to make decisions about teacher dismissal (see Figure 5.6).[16] This could reflect clearer understanding of the measurement of TE, as well as clearer understanding about termination policies per se. Nevertheless, even in the most recent year in which we asked this question (2015), only about half of each site's teachers reported having a clear understanding of the dismissal criteria. To the extent that principal discretion was involved in dismissal decisions, such as in the CMOs, a lack of clear understanding might not be surprising. And some teachers who thought they understood the criteria might not have. One HCPS teacher we interviewed in 2014 commented,

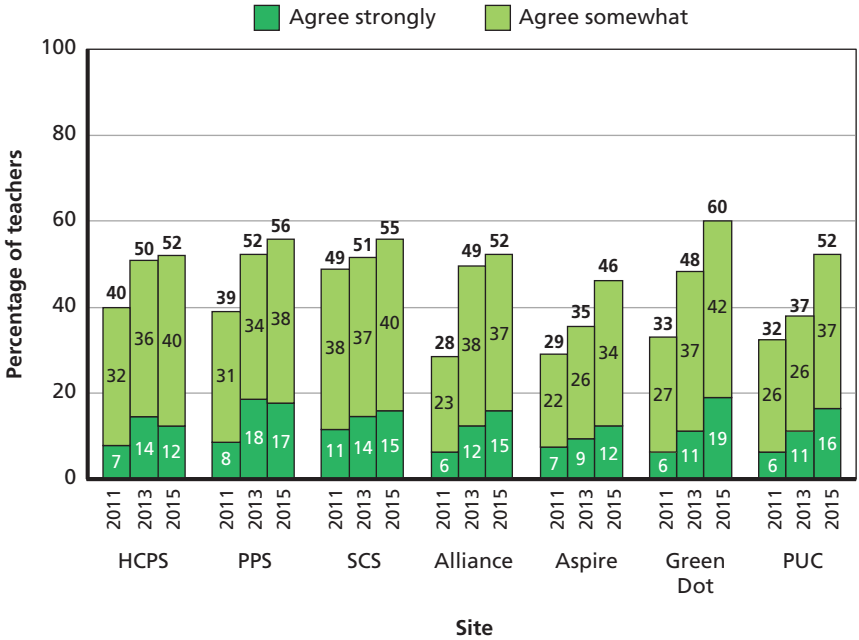
> If you would put me on the spot, I have a pretty good understanding of it. That we have a window of like a couple years to improve, we get more evaluations, and I think the reality is [that, if] it's three years in a row you're unsatisfactory, I think they can get rid of you.

In fact, HCPS's criterion for dismissal of nonprobationary teachers was two consecutive U ratings or three consecutive NI ratings.

---

[16] The increase in the percentage agreeing (strongly and somewhat) from 2011 to 2015 was statistically significant ($p < 0.05$) in all seven sites.

Figure 5.6
Teachers' Agreement That They Had a Clear Understanding of the Criteria Used to Make Decisions About Teacher Termination, Springs 2011, 2013, and 2015



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly." On the teacher survey, we asked about dismissal in only 2011, 2013, and 2015.

RAND *RR2242-5.6*

Endorsement

Tenure

Over time, fewer teachers and SLs thought that tenure should be linked to teacher-evaluation results. We asked teachers and SLs who reported that their site granted tenure whether they agreed that "Tenure should be linked to teachers' evaluation results."[17] (As noted earlier, because the CMOs did not offer tenure, we show results only

[17] Although this survey question was not about survey respondents' support for the specific tenure policies enacted in their site, the three districts did, as discussed earlier in this chapter, base decisions about tenure (or equivalent) on teachers' evaluation results during the

for districts.) As Figure 5.7 shows, there was a decline over time in all three districts in the percentages of both teachers and SLs agreeing with the statement.[18] This suggests decreasing support for policies linking tenure to evaluation results, although, in HCPS and SCS, there was also a large drop in the percentage of respondents answering the question because only respondents who said that their site granted tenure were routed to it. We are not sure why support for linking tenure to evaluation declined, but it might reflect a general fall in teachers' faith in the evaluation system, as reported in Chapter Three. SLs were more likely than teachers to agree that tenure should be linked to evaluation results.[19]

Among teachers, the opinions of novice teachers (who have not yet reached the tenure determination) might arguably be of greater relevance than the opinions of more-experienced teachers, and results indicate that the views about tenure held by novice and experienced teachers did, in fact, differ. In 2015, novice teachers in PPS and SCS were more likely than experienced teachers to agree with the statement about linking tenure to evaluation. In both sites, more than half of novice teachers agreed, while less than half of experienced teachers agreed.[20] In HCPS, the percentages of novice teachers and experienced teachers agreeing were roughly equivalent (43 and 45 percent, respectively).

---

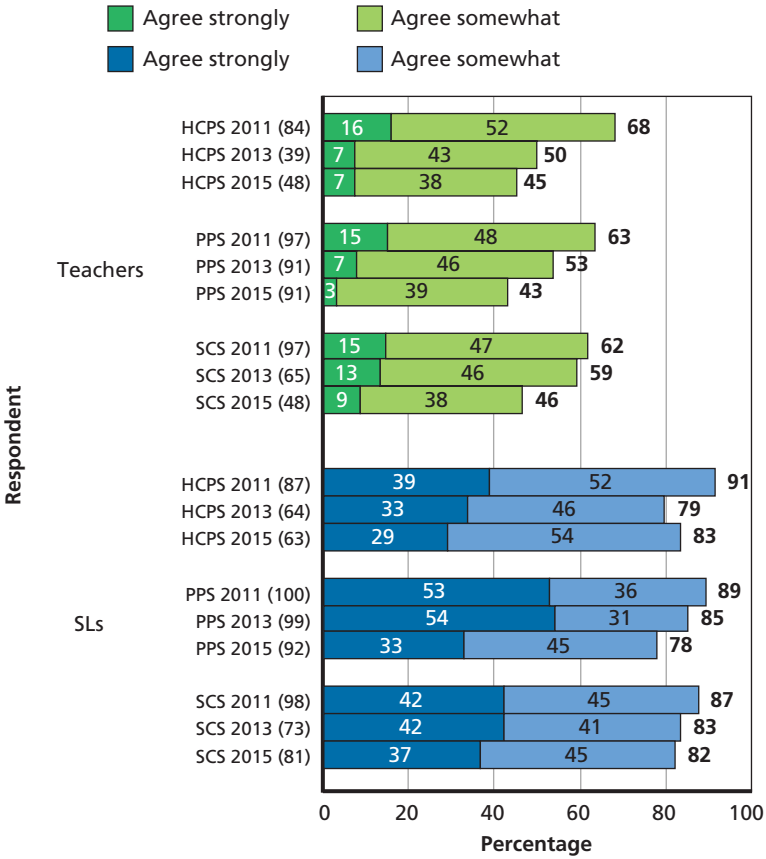
years studied. Therefore, we use this survey item as a proxy to gauge support for the districts' policies.

[18] The decrease in the percentage agreeing (strongly and somewhat) from 2011 to 2015 was statistically significant ($p < 0.05$) in all three sites for both teachers and SLs.

[19] The results indicating that SLs had more-positive attitudes about linking tenure to evaluation reflect a general pattern: SLs tended to report more-positive attitudes than teachers about many of the policies included on the survey.

[20] In SCS, there was a statistically significant difference ($p < 0.05$) between teachers with two years of experience or less (72 percent of whom agreed) and teachers with more than two years of experience (45 percent of whom agreed). In PPS, 54 percent of teachers with two years of experience or less agreed, higher than the 42 percent of teachers with more than two years of experience, but this difference was not statistically significant.

Figure 5.7
Teachers' and SLs' Agreement That Tenure Should Be Linked to Teachers' Evaluation Results, Springs 2011, 2013, and 2015



NOTE: Numbers in parentheses next to the site names are the percentages of respondents who indicated that their site grants tenure. The percentages in the bars are among the respondents who indicated that their site grants tenure. Omitted response categories are "disagree somewhat" and "disagree strongly."

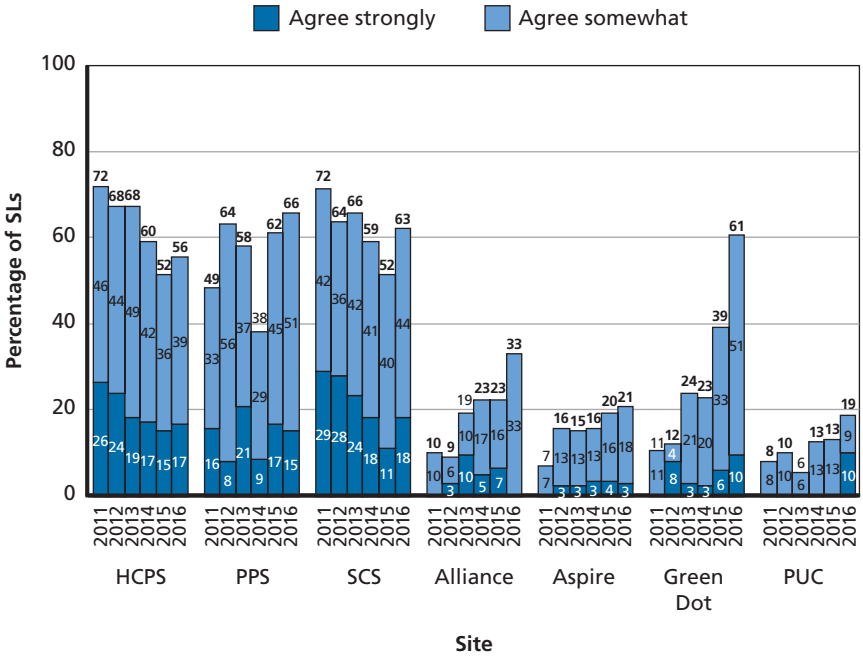
RAND *RR2242-5.7*

Dismissal

Over the course of the initiative, SLs in HCPS and SCS became less likely to report obstacles to teacher dismissal, but the opposite was true in the CMOs. We asked SLs how much they agreed that "The termination/dismissal procedures in my [district or CMO] are so

burdensome that most school administrators try to avoid using them." In HCPS and SCS, more than 70 percent of SLs agreed with the statement in 2011, but the percentage agreeing dropped fairly steadily through 2015.[21] (See Figure 5.8.) Each of the two sites had an uptick in 2016. PPS started out lower and ended up higher than the other two districts, but percentages in the intervening years—particularly a large dip in 2014—make a trend difficult to discern. In the CMOs, the per-

Figure 5.8
SLs' Agreement That the Termination Procedures in Their Site Were So Burdensome That Most School Administrators Tried to Avoid Using Them, Springs 2011–2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly."

RAND *RR2242-5.8*

[21] The decrease in the percentage agreeing (strongly and somewhat) from 2011 to 2015 was statistically significant ($p < 0.05$) in HCPS and SCS. The increase from 2015 to 2016 was also significant in both sites.

centages agreeing were consistently very low, but they increased over time in all four CMOs.[22] Thus, it seems that, although school leaders in at least two of the districts perceived an easing of obstacles to teacher dismissal, SLs in the CMOs perceived the opposite.

Principals we interviewed in SCS echoed the survey finding that dismissing teachers had become easier. As one SCS principal said in 2015,

> [Dismissing teachers has] been going very well for principals. We're finally getting the opportunity to really get rid of some people. Rather than the olden days, when you had that magical tenure and it was damn near impossible [to dismiss a teacher].

Fairness

Nearly all SLs thought that low-performing teachers had sufficient opportunities to improve before being dismissed, and few outside of SCS thought that teachers were dismissed unfairly. We asked SLs two questions pertaining to the fairness of dismissal policies.[23] The first was whether "teachers who are in danger of being dismissed based on performance currently have sufficient opportunities and time to improve prior to actually being dismissed." As Figure 5.9 shows, almost all SLs agreed at least somewhat, and, except in SCS, majorities agreed strongly.

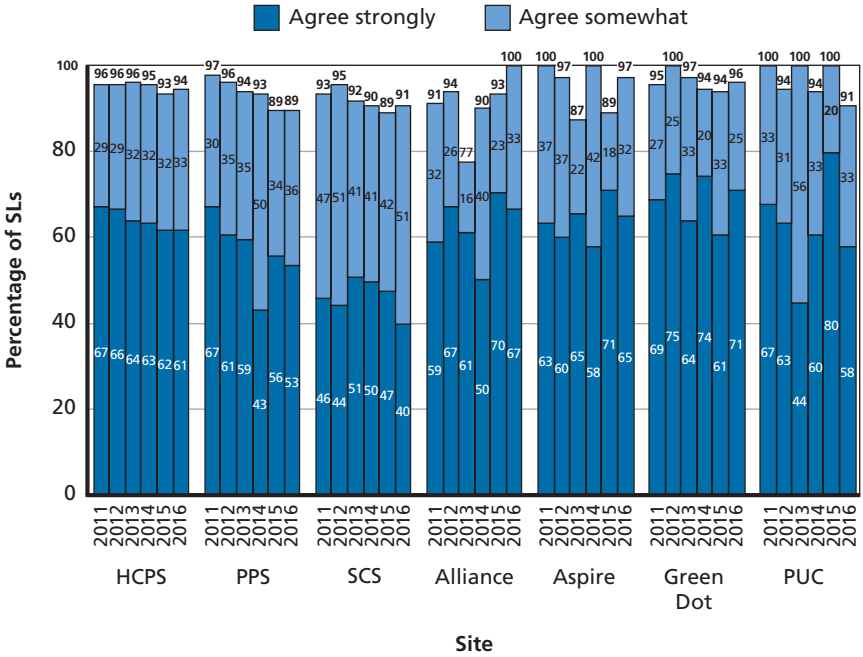
The second question was whether "it sometimes happens that a teacher who is actually very good gets dismissed." The percentages of SLs agreeing with this statement were very low, particularly in HCPS and the four CMOs (typically 10 to 20 percent of SLs agreed). Percentages were, however, notably higher in SCS, ranging from 32 percent to

---

[22] The increase in the percentage agreeing (strongly and somewhat) from 2011 to 2016 was statistically significant ($p < 0.05$) in Aspire and Green Dot but not in Alliance or PUC. In Alliance, the increase from 2011 to 2015 was statistically significant, so the lack of significance of the increase from 2011 to 2016 might have been due to the small sample in 2016.

[23] Because we expected that relatively few teacher respondents would themselves have experienced a tenure decision under the revised tenure policies and that hardly any would have been through dismissal proceedings, the teacher survey did not include any questions about the perceived fairness of these policies.

Figure 5.9
SLs' Agreement That "Teachers Who Are in Danger of Being Dismissed Based on Performance Currently Have Sufficient Opportunities and Time to Improve Prior to Actually Being Dismissed," Springs 2011–2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly."

RAND *RR2242-5.9*

41 percent. Thus, SLs in SCS do seem to have had some concerns about the fairness of teacher dismissal in the district, at least compared with those in the other sites. Because of technicalities in SCS's weighting scheme and evaluation requirements, the dismissal of teachers whom SLs perceived as "good" (i.e., teachers who had decent observation scores) might have been more likely in SCS than in the other sites.

Perceived Effects

Tenure

During the initiative, teachers and SLs became less likely to think that tenure was protecting bad or ineffective teachers. We asked

teachers and SLs whether they agreed that, "As currently implemented in my district, tenure protects bad or ineffective teachers." To the extent that the districts successfully implemented policies making tenure decisions more rigorous, one would expect to see declining agreement with the statement, and, in fact, that is what we see (Figure 5.10). In all three districts, between 2011 and 2015, both teachers and SLs became less likely to think that tenure, as implemented in their district, protected bad or ineffective teachers.[24] Although this decline could indicate that the policy changes related to tenure were perceived as being at least somewhat successful, the declining percentages could also reflect other factors. For example, perhaps, during the study period, people perceived a decline in the prevalence of bad or ineffective teachers, and this, rather than a change in tenure policy, caused the declining agreement with the statement.

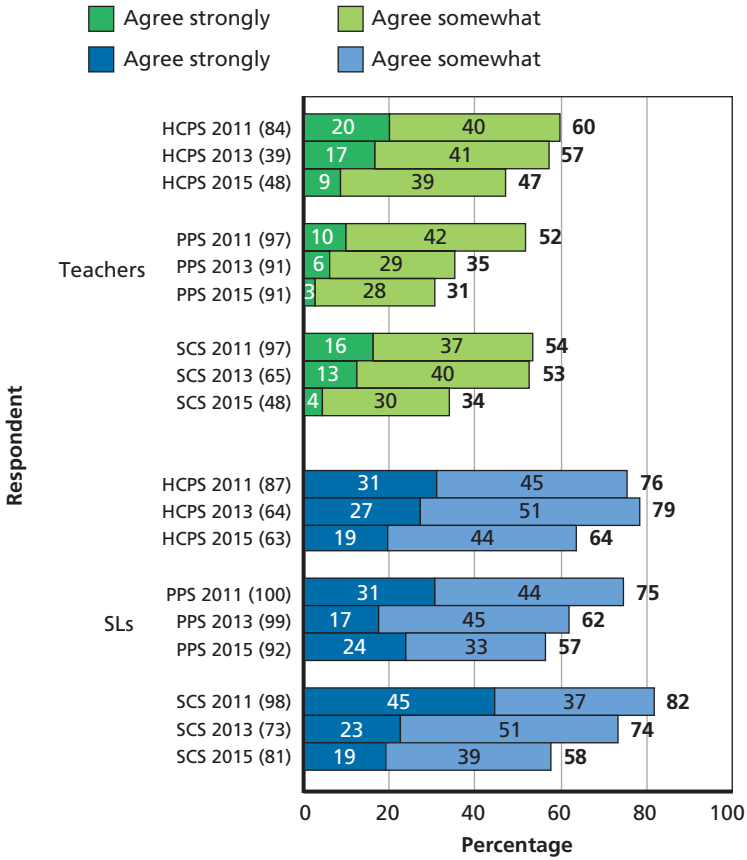
Dismissal

Few teachers reported worrying about dismissal, but teachers with low effectiveness ratings were more likely than higher-rated teachers to report worrying. In theory, fear of dismissal—which might reasonably accompany stricter or more-formalized dismissal policies tied to more-rigorous evaluations—could motivate teachers to improve. Figure 5.11 shows that very few teachers reported being worried about dismissal. Four sites, however, did see a significant uptick in percentages in 2013,[25] when the effectiveness criteria established as part of the initiative—and the potential for dismissals based on them—were newer. Because very few teachers were actually dismissed, it makes sense that most teachers' worries would ease after the low incidence of dismissal became known. In 2015, in every site except Aspire, teachers who had received low evaluation ratings were, as one would

[24] The decrease in the percentage agreeing (strongly and somewhat) from 2011 to 2015 was statistically significant ($p < 0.05$) in all three districts for both teachers and SLs.

[25] The increase in the percentage agreeing (strongly and somewhat) from 2011 to 2013 was statistically significant ($p < 0.05$) in all four sites (PPS, SCS, Aspire, and PUC), as was the decrease from 2013 to 2015.

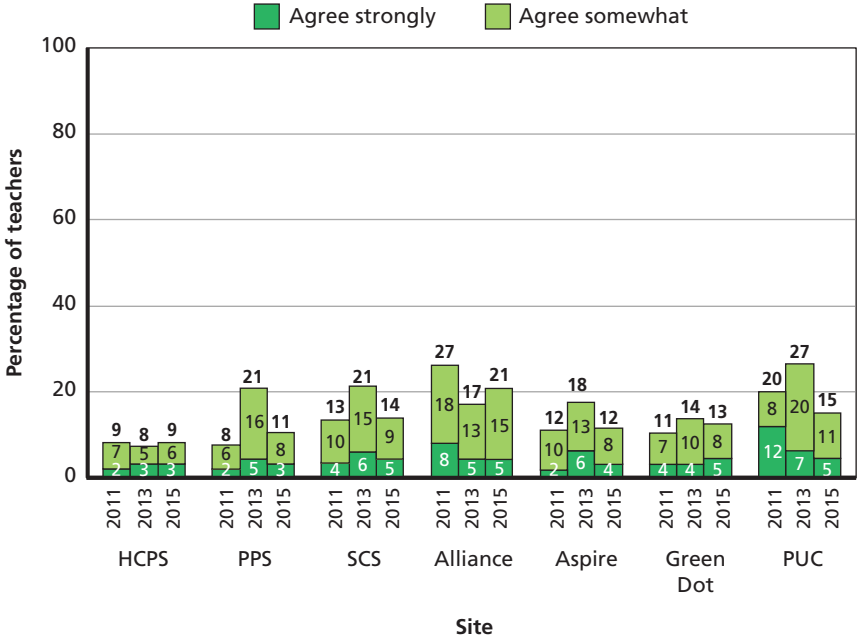
Figure 5.10
Teachers' and SLs' Agreement That, as Implemented in Their District, Tenure Protected Bad or Ineffective Teachers, Springs 2011, 2013, and 2015



NOTE: Numbers in parentheses next to the site names are the percentages of respondents who indicated that their site grants tenure. The percentages in the bars are among the respondents who indicated that their site grants tenure. Omitted response categories are "disagree somewhat" and "disagree strongly."

RAND *RR2242-5.10*

Figure 5.11
Teachers' Agreement That "I Am Really Worried About Being Dismissed," Springs 2011, 2013, and 2015



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly."
RAND *RR2242-5.11*

expect, more likely than teachers with higher ratings to agree with the statement.[26]

In PPS in particular, interviews suggested a somewhat higher degree of worry than the survey results did. Examples of teacher comments include the following:

> It's a little disconcerting to know that there are so many factors that they say they take into account that we have no control over that could dictate whether we have our job or not. Let's be clear. It

[26] Only in HCPS was this difference statistically significant ($p < 0.5$). That the difference was not significant in more of the sites is likely due at least in part to the small numbers of low-rated teachers.

> hasn't happened yet. I think it's more paranoia. . . . It's definitely a paranoia issue. (PPS teacher, 2012)

> You sometimes hear things that play on your fears and anxieties. You worry because you want to keep your job, but it's hard because you're not getting a lot of positive reinforcement. (PPS teacher, 2013)

> The whole evaluation system puts a lot of pressure on teachers; we worry about getting fired if we get a bad score. (PPS teacher, 2013)

In every site, at least one-third of teachers said that they would seriously consider leaving teaching if they received a very low evaluation rating. In some cases, low evaluation results might have led teachers to think about leaving voluntarily, perhaps after being "counseled out" by SLs or "counseling themselves out." Indeed, about half the teachers in HCPS and 30 to 45 percent in the other sites indicated that they would seriously consider leaving teaching if they received a very low evaluation rating (see Figure 5.12). From 2013 to 2016, the percentages agreeing rose significantly in PPS and SCS and declined significantly in PUC.[27]

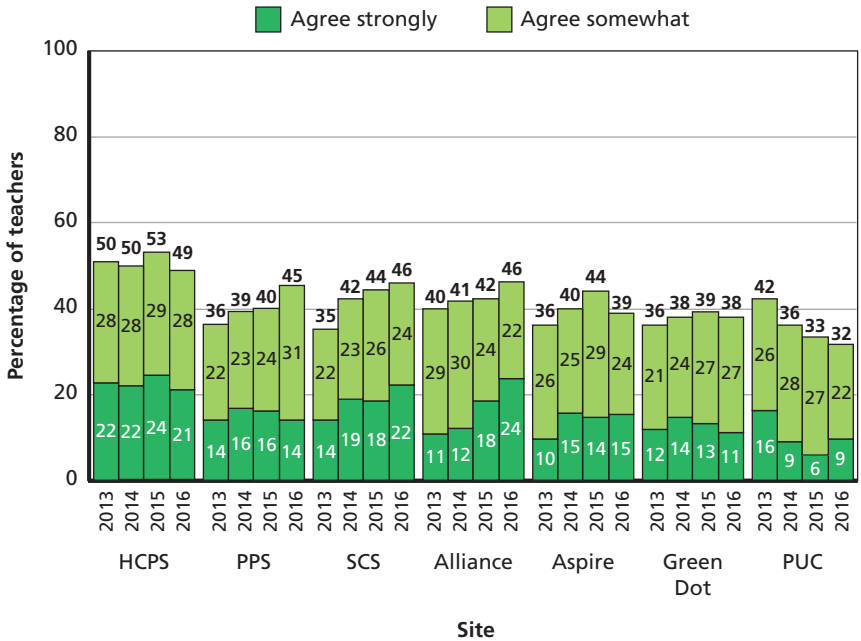
Some teachers and SLs we interviewed did mention voluntary departures. Examples include the following:

> Since the inception of this program [the IP initiative in HCPS,] . . . a lot of my friends either retired early or quit teaching altogether. I'm right behind them. I'm leaving next year . . . . All of [a] sudden, I'm mediocre, and it has nothing to do with my age—I'm still on fire in the classroom, I love the kids, I love my administration, [but] this program has destroyed my love for this vocation. (HCPS teacher, 2013)

---

[27] The results reported here concern teachers' responses to a hypothetical question about what the respondent would do if he or she a received low rating. Chapter Eleven reports results on teachers' stated intent to remain teaching in their site and the relationship between their intent to stay in teaching and their effectiveness rating.

Figure 5.12
Teachers' Agreement That, "If I Received a Very Low Evaluation Rating, I Would Seriously Consider Leaving Teaching," Springs 2013–2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly." We did not ask this question in 2011.

RAND *RR2242-5.12*

> At our school, I don't know of anybody [who] has been transitioned out solely based on their evaluation. It has become evident to them as a person that they're not where they should be, they shouldn't be doing this teaching, and they've chosen to leave. I don't think we've had to force to have anyone leave based on numbers. (HCPS teacher, 2014)

> [The TLE initiative] made a lot of older teachers retire quicker than they wanted to. Not just [the TLE] by itself, but the whole system, the way it's going. (SCS teacher, 2016)

Summary

Reformers have advocated four types of actions they believe would improve the tenure and dismissal processes in public schools: extending the probationary period for teachers, increasing the performance standard to move from probationary status to tenure, requiring teachers to maintain a satisfactory level of measured effectiveness to retain tenure, and streamlining the dismissal process for tenured teachers. Over the course of the IP initiative, all three districts revised policies related to two of these actions: increasing the standard to move from probationary to tenured or nonprobationary status and raising the bar on effectiveness to continue as a tenured or nonprobationary teacher. At the same time, Florida and Tennessee, the states in which HCPS and SCS are located, made changes in state tenure policies, with Florida abolishing tenure and Tennessee extending the probationary (pretenure) clock from three to five years. Like the districts, the four CMOs, which did not have tenure, worked on raising the bar to continue as a teacher.

The sites confronted several challenges in implementing tenure and dismissal reforms. In particular, they struggled to balance reforms designed to dismiss low-performing teachers with improvement policies designed to enhance the performance of all teachers. In addition, the fact that the sites used teacher-evaluation results as the basis for tenure and dismissal decisions might have led some principals to avoid giving low observation ratings that would reduce a teacher's composite score. Also, in most sites, the composite TE measure for a given school year could not be computed and released until the following fall, which complicated its use in tenure and dismissal decisions.

Our survey and interview data for teachers and SLs suggest that the reforms that the sites initiated had a reasonable level of support among these two groups. In particular, the survey data indicate that most teachers were aware of the policies in their site. In addition, many agreed with the policies, although the percentage endorsing the policies fell over time. SLs in the districts saw fewer obstacles to dismissing low-performing teachers as the initiative progressed, but leaders in the CMOs saw increasing obstacles, perhaps due to the tightening labor market for teachers in California, where most of the CMOs were

located (Carver-Thomas and Darling-Hammond, 2017). In general, SLs saw the systems as fair. And relatively few teachers reported that they were worried about dismissal.

Despite these positive changes, the percentage of teachers who were denied tenure or dismissed because of low performance was low each year in the districts for which data are available. Although these rates are somewhat above the national averages discussed in the introduction (1.2 percent for nontenured teachers and 0.2 percent for tenured teachers), they are much lower than the rates envisioned by some reformers (for example, 20 percent denied tenure). Similarly, the CMOs dismissed few teachers (i.e., few teachers failed to have their contracts renewed because of poor performance). One reason for this is the fact that the evaluation systems in the districts and CMOs identified very few teachers as less than E. Thus, few were identified for improvement plans or dismissal.

Thus, our findings related to the tenure and dismissal lever provide a mixed picture. The sites were able to implement changes that made it somewhat more difficult for teachers to achieve tenure, as well as changes that made it easier to systematically identify ineffective teachers based on teacher-evaluation results, place them on improvement plans, and dismiss the ones who did not improve. But because few teachers were rated at the levels triggering such actions, the reforms did not directly affect a large number of teachers, although perhaps the reforms had indirect effects on a larger number of teachers, encouraging them to perform at levels that would not trigger specific actions.

CHAPTER SIX

PD Policies

> The most powerful impacts are from my debrief conversations with the person [who is] observing me, and, after that, it would probably be workshops with peers at a similar level, the differentiated PDs. Online resources are great, but [they] can be kind of overwhelming and hard to sift through at times.
>
> —Aspire teacher, interviewed in 2015 about PD policies

In Brief

What kinds of PD activities did the sites plan and deliver, especially to meet needs identified through the teacher-evaluation system? Did the sites monitor teachers' participation in PD? Did teachers view the PD as useful in improving their instruction and student learning?

All seven of the IP sites offered multiple types of PD: coaching, district- or CMO-wide workshops, school-based workshops, school-based teacher collaboration, access to online materials (lesson plans and readings), and access to videos of lessons illustrating teaching practices in relation to the site's observation rubric. Principals and other staff recommended PD based on teachers' evaluation results, but sites did not require teachers to participate in recommended PD, nor did they put systems in place to monitor teachers' participation or determine whether teachers' effectiveness improved after participating. In all sites, teachers reported awareness of the available PD activities. Although many teachers reported that their evaluation results influenced their choice of PD, more reported other influences, such as their own interests. Teachers generally believed that their PD was

useful in improving student learning. In addition, most teachers viewed school-based collaboration as useful in improving their effectiveness; fewer teachers viewed online materials and videos as useful.

Introduction

This chapter reviews the PD activities that the IP sites provided to support teachers in meeting needs identified through the teacher-evaluation system. A key assumption underlying the IP theory of action (see Chapter One) was that SLs and others would draw on information from the teacher-evaluation system to identify appropriate PD opportunities for teachers and that participating in PD would support teachers in improving their practice. In this chapter, we review the research literature on the use of teacher-evaluation results to guide PD, we examine the PD programs the sites put in place, and we describe teachers' perceptions of the PD available to them.

Best Practices for Evaluation-Linked PD

There is a vast literature on PD, and a full review is beyond the scope of this report. Here, we discuss the much smaller literature focused explicitly on the role of teacher evaluation in PD. Advocates of teacher-evaluation reform have argued that teacher evaluation can serve as a valuable source of information to guide teachers' PD. (See, for example, Weisberg et al., 2009.) The basic premise is that, by identifying teachers' strengths and weaknesses, evaluation results can help determine areas in which sites should focus their PD and teachers should invest their development efforts (Coggshall et al., 2012).

In much of the literature on the link between evaluation systems and PD, evaluation systems are viewed as providing a periodic source of actionable data on teachers' instruction, to be used in guiding decisions about the PD in which teachers should participate (see Goe, Biggers, and Croft, 2012). This might occur in several ways. For example, a teacher's annual evaluation results (including composite scores, com-

ponent rating, and any narrative information accompanying observations) could be used to craft the teacher's PD plan for the coming year.[1] Or, teachers and SLs could discuss appropriate PD activities during the feedback sessions that follow each classroom observation. Or, it could be left to the teacher to seek out PD that might address the areas of need identified during feedback sessions.

Perhaps in part because efforts to link teacher evaluation and PD have been fairly recent, there have been few rigorous efforts to assess whether drawing on evaluation results to make decisions about PD improves teacher and student outcomes. There is, however, some descriptive evidence. Koedel and his colleagues, drawing on Tennessee data linking teachers' evaluation ratings to survey data on their PD participation, found that teachers with lower evaluation scores tended to participate in more PD activities, which one might anticipate if evaluation ratings were driving PD choices (Koedel, Li, et al., 2015).[2] Furthermore, the relationship between evaluation scores and PD participation was stronger for ratings based on observations alone than for those based on the Tennessee composite evaluation measure, suggesting that the teachers might have sought PD based on feedback from observers rather than solely on the composite end-of-year rating.

In another study, Shakman and her colleagues examined a district in which principals were expected to prescribe one or two PD activities for each teacher for each of the four standards assessed and record the prescribed activities in a district data system (Shakman et al., 2016). Across all four standards, most teachers who received prescriptions participated in at least one, although less than 40 percent of teachers participated in all PD activities prescribed for them. Furthermore, as predicted by the premise underlying evaluation-linked PD, teachers who participated in at least one prescribed activity were more likely than teachers who did not to be rated at least P on their subsequent

---

[1] The plan, jointly developed by the teacher and the SL, would discuss the teacher's evaluation results, determine the practices to be targeted for improvement, and identify the specific PD activities to be undertaken over the year.

[2] Although Koedel et al. found that teachers with lower evaluation ratings participated in more PD, they lacked the data to assess whether the PD in which these teachers participated matched the PD recommended as part of their evaluations.

ratings. The literature describes a variety of tasks or activities that need to be carried out to set the groundwork for linking evaluation and PD. Although these ideas are based more on professional judgment than on research, there is some consensus in the literature about their value.

Cataloging Available PD

One key task is cataloging the PD opportunities available in a district or CMO and determining their value in terms of supporting the kinds of practices emphasized in the teacher-evaluation system. For example, Goe, Biggers, and Croft argued,

> An important step in planning for professional growth is to take inventory of current professional development opportunities and make decisions about what should be continued, what should be eliminated, and where gaps exist. There are two main points to keep in mind when making these decisions: the characteristics of high-quality professional development and the meaning of the scores that will be produced by the selected evaluation tools. (Goe, Biggers, and Croft, 2012, p. 16)

This can be a challenging step to undertake, in part because districts and CMOs generally offer a range of diverse PD opportunities, including short workshops (an hour or two during or after school), institutes (multiday activities, often offered on weekends or in the summer), online resources (e.g., readings and curriculum materials, webinars, and online courses), videos of instruction aligned with district standards, coaching, collaboration (often in the form of study groups, learning communities, or grade-level or subject-area teams), and mentoring (career development advice). The latter three can be particularly difficult to inventory in relation to the evaluation system.

Assessing the Quality of Available PD

A second task is ensuring the quality of the PD included in the district or CMO portfolio. One strategy for measuring quality is to assess the PD's consistency with features of high-quality PD recommended in the professional and research literature (see, for example, Garet, Porter,

et al., 2001; Desimone, 2009; and Coggshall et al., 2012).[3] But despite the consensus on these features, empirical studies of PD's impact on student achievement do not provide consistent support for their effectiveness, making it difficult for districts to draw conclusions about the quality of their existing PD and to make decisions about the types of PD in which to invest.[4]

Establishing Embedded Forms of PD

Another activity that proponents of evaluation-linked PD emphasize is fostering embedded forms of PD—that is, PD that is "grounded in day-to-day teaching practice, occurs regularly, [and] consists of teachers analyzing students' learning and finding solutions to immediate problems of practice" (Coggshall et al., 2012, p. 4). Coaching and collaboration are the most-common examples.

Proponents of evaluation-linked PD often recommend coaching (assigning a staff person to work with a teacher on improving classroom instruction), on the theory that the coach can tailor the PD to the teacher (Goe, Biggers, and Croft, 2012). There is some evidence that some coaching programs might be effective. In a recent review, Kraft, Blazar, and Hogan found a positive overall impact on student achievement across the studies of coaching in their sample (Kraft, Blazar, and Hogan, 2018). The overall effect, however, was driven largely by studies of literacy coaching, which constituted 16 of the 23 studies. Coaching had no effect in the four studies that examined its impact on mathematics achievement.

---

[3] For example, it has been hypothesized that PD is likely to be more effective if it is sustained over time, involves other teachers from the same school or grade level (so that they can reinforce what is learned), is embedded in regular classroom teaching (that is, based on curriculum materials, assessments, and other materials in use in the classroom), involves opportunities for active learning (such as practice with feedback), is coherent (building on earlier PD), and is focused on curricular content (e.g., mathematics or ELA).

[4] For example, in a recent research synthesis of 28 studies of the impact of PD, Kennedy found that PD with the features of high-quality PD described in the preceding footnote was no more effective, on average, than PD without those features (Kennedy, 2016). For example, among the PD interventions included in her review, those that were more intensive (in terms of hours) were no more effective than those that were less intensive.

Collaborative forms of PD (study groups, professional learning communities, and grade-level teams) might also be useful evaluation-linked PD strategies, on the theory that teachers who are struggling might learn from others who excel and teachers might engage in joint problem-solving with respect to issues identified in evaluation results (Coggshall et al., 2012). Unfortunately, there are few studies of the impact of collaborative forms of PD. But Papay and his colleagues provided some evidence that a low-performing teacher can benefit by working with a higher-performing teacher (Papay et al., 2016).

Apart from coaching and collaboration, the feedback that occurs after evaluation-related observations can itself be viewed *as a form of PD*, involving a teacher reflecting on his or her practice and perhaps coaching by the observer. (See Papay, 2012; and Rowan and Raudenbush, 2016.) A few studies have examined the impact of the feedback teachers receive after evaluation-related observations and have found that it can have positive effects. For example, Taylor and Tyler, in a study in Cincinnati, compared fourth- through eighth-grade teachers who received feedback four times during a school year based on the FFT rubric (once from the principal and three times from a peer observer) with teachers who did not receive feedback that year (Taylor and Tyler, 2012). Students of teachers who received feedback gained 0.1 standard deviations more in mathematics than the teachers who did not, although there was no difference in reading. Steinberg and Sartain found similar results in a randomized trial examining the impact of feedback using the FFT in Chicago, although the effects were statistically significant in reading but not mathematics (Steinberg and Sartain, 2015). In another recent study of the impact of performance feedback, Garet, Wayne, and their colleagues found that students in schools in which feedback was provided scored 0.05 standard deviations higher in mathematics than students in control schools did. There was no statistically significant impact in reading, however (Garet, Wayne, et al., 2017).[5]

---

[5] The intervention in the Garet, Wayne, et al. study involved three components: feedback for teachers based on four classroom observations per year, provision of scores for value

Establishing the Infrastructure to Support Teachers and Leaders in Identifying PD and Tracking Participation

Another task in establishing a system of evaluation-linked PD is building an infrastructure to support SLs in making wise recommendations about PD and helping teachers follow through on them. Coggshall and her colleagues, for example, recommended that teachers routinely complete PD activity logs, which "typically include detailed descriptions of their significant learning opportunities" (Coggshall et al., 2012, p. 17). In a similar vein, Curtis and Wiener recommended implementing online systems for registering for PD and tracking participation (Curtis and Wiener, 2012). Curtis and Wiener also suggested that districts examine their policies to see whether any of them might impede basing PD decisions on evaluation results:

> Are there district policies, including contract terms, that impede the use of evaluation information in supporting teachers' growth and development (e.g., coaches cannot have access to evaluation information; professional development is determined solely at teacher's discretion rather than developed as part of professional growth planning)? If so what needs to be done to align these policies to the goals of the evaluation system? (p. 40)

Integrating the Parts of the System

Little is yet known about the feasibility and effectiveness of implementing a full system of PD linked to evaluation results, as recommended in these steps. But in case studies conducted in six high-performing schools, Reinhorn, Johnson, and Simon provided an indication that at least some aspects are possible:

> Each of these principals had an integrated strategy for improving teachers' practice across the school. Evaluation did not stand alone, but rather, was coordinated with other professional learning opportunities, such as instructional coaching, teacher teams, whole school professional development, and peer observation.

---

added, and feedback for principals on their leadership. No stakes were attached to teachers' performance.

> . . . In explaining the support they received, teachers often did not distinguish between practices that were part of the evaluation system and others; they considered them all as part of an ongoing, integrated improvement process. However, many identified classroom observations and feedback as the most valuable component of the process. (Reinhorn, Johnson, and Simon, 2017, pp. 25–26)

Intensive Partnership PD Policies

In all seven sites, a primary purpose of the teacher-evaluation system was to identify areas for teacher learning and development.[6] This was expected to occur in two somewhat different ways. First, as described in Chapter Three, teachers were provided with in-person feedback after each formal observation conducted as part of the teacher-evaluation process, and sometimes after informal observations.[7] As part of the feedback, principals and other observers were encouraged to provide ideas about areas on which teachers might focus to improve performance and to suggest specific changes they might make in their teaching. Thus, the feedback sessions might be viewed as a form of PD in themselves.

In addition, principals and others were expected to draw on teachers' overall evaluation ratings, as well as scores and other information from the evaluation components (e.g., observations and student achievement growth) to derive recommendations for PD activities in which teachers might engage. This chapter focuses on the policies and practices the sites implemented to link teachers' evaluation results to PD. This section briefly outlines the types of PD activities that the

---

[6] The information discussed in the text and reported in Tables 6.1 and 6.2 on the PD policies adopted and implemented in the sites comes primarily from annual interviews with central-office staff and reviews of policy documents from each of the sites.

[7] Chapter Three describes in greater detail the observations conducted as part of teacher evaluation. Informal observations were not scheduled in advance, were shorter in duration than formal observations, and were focused on a subset of indicators. Although informal observations generally did not include pre- and postobservation conferences, teachers nevertheless might have received brief feedback.

districts and CMOs provided and the extent to which they used evaluation results to guide teachers' PD participation.

District PD Policies

In all three districts, principals were encouraged to draw on individual teachers' observations and TE results (e.g., composite scores) to recommend PD programs and develop PD plans. These recommendations took varied forms. In some cases, principals recommended specific PD programs or activities—for example, attending a particular workshop, completing a reading, watching an online video, observing another teacher, or working with a mentor or coach.[8] In other cases, principals recommended a particular topic (e.g., PD on classroom management or questioning strategies) but not a specific activity. In addition to drawing on evaluation results to develop plans for individual teachers (including teachers identified for improvement), principals and district staff also drew on TE results to identify needs that many teachers shared, to inform school-wide or district-wide PD.

All three districts offered a combination of coaching, district-provided workshops, workshops at individual school sites, and resources for individual use (e.g., online PD; see Table 6.1.) In general, the initiative did not expand the types of PD available or, in most cases, the focus of the PD: Most of the types of PD listed were available in the districts prior to the initiative. However, the districts did add PD offerings during the initiative. For example, they developed workshops focused on the observation rubric to clarify the dimensions of practice and the standards expected; HCPS, for instance, offered a seven-hour workshop that could be completed in various settings (in person on a Saturday, in separate sessions, or online). In addition, all three districts expanded the use of coaching, often assigning coaches to work with teachers identified as struggling based on their evaluation results.

---

[8] We define *mentoring* as providing guidance on career development and professional goals and *coaching* as working with teachers in the classroom, which could include feedback on instructional practice or demonstrating lessons. In practice, the two activities can overlap. For example, mentors for new teachers might provide a mix of guidance on career development and direct feedback on classroom practice.

Table 6.1
District PD Policies Through 2015–2016

Policy Area	HCPS	PPS	SCS
Use of TE data to identify teacher PD needs	**2010–2011 and ongoing:** Principals, peer evaluators, and mentors use observation rubric component scores to recommend PD programs for teachers.	**2010–2011 and ongoing:** Observation scores are used as a basis for feedback from the observer.	**2010–2011 and ongoing:** Observation scores are used as a basis for feedback from the observer.
		2013–2014 and ongoing: The TE score is used as a basis for the PD plan.	**2013–2014 and ongoing:** The TE score is used as a basis for the PD plan.
PD provided	**Prior to initiative and ongoing:** Coaching; district-provided large-group sessions; PD provided at the school site; resources for individual use (see last section of table)	**Prior to initiative and ongoing:** Coaching; district-provided large-group sessions; PD provided at the school site	**2010–2011 and ongoing:** Coaching; district-provided large-group sessions; PD provided at the school site
		2011–2012 and ongoing: Resources for individual use (see last section of table)	**Prior to initiative and ongoing:** Resources for individual use (see last section of table)
PD for new teachers			
What	Mentoring program	Two-week summer induction program	New teachers participate in new-teacher induction before the start of the school year and receive a stipend to do so.
When	2010–2011 and ongoing	2010–2011 through 2013–2014	Summer 2013 and ongoing

Table 6.1—Continued

Policy Area	HCPS	PPS	SCS
For whom	Teachers in their first two years who had not previously taught elsewhere	Teachers with level I certification[a]	Teachers new to teaching or to SCS
Frequency	Throughout the year	Summer before starting at PPS	Summer before starting at SCS
What	Courses designed specifically for new teachers, aligned with the FFT and FEAPs. Required part of TIP.	New-teacher induction throughout the school year	Each new teacher is assigned a veteran mentor for his or her first year; mentors work with mentees face-to-face every month for coaching and support and discuss their mentees' progress with the principal. As of the fall of 2015, coaching was no longer part of the mentor teacher's role.
When	2011–2012 and ongoing	2014–2015 and ongoing	2012–2013 and ongoing
For whom	Any teacher in his or her first year who has not previously taught	Teachers with level I certification	Teachers new to teaching or to SCS
Frequency	Throughout the year	Throughout the year	Throughout the year
What		Ongoing coaching and support	
When		2014–2015 and ongoing	
For whom		Any teacher in his or her first or second year	
Frequency		Throughout the year	

Table 6.1—Continued

Policy Area	HCPS	PPS	SCS
New resources for individual use introduced during initiative	Website with PD programs listed by observation rubric component. The primary online platform before and during the initiative was Moodle.	Repository of training videos, readings, example lesson plans, Tripod resource guide, and other resources suggested by principals or accessible to teachers on LearningBridge (2011–2013) or BloomBoard (2013 and ongoing)	Catalog of PD options called the TEM Resource Book; Teachscape video capture and self-reflection; repository of training videos; earbud coaching, use of which was encouraged by central office; My Learning Plan (online repository of PD resources)
When	April 2012 and ongoing	2011–2012 and ongoing	2011–2012 and ongoing
For whom	All teachers	All teachers	All teachers
Frequency	As needed	As needed	As needed

SOURCES: Annual interviews with central-office staff in each of the IP districts.

NOTE: TIP = Teacher Induction Program.

[a] In Pennsylvania, state law requires teachers with level I certification to participate in an induction program as a condition of earning a level II certification. The majority of PPS teachers who participate in the induction program are in their first or second year of teaching.

All three districts also adopted new mentoring programs for novice teachers over the course of the initiative. For example, in SCS, starting in the fall of 2012, every teacher new to the district and to teaching was paired with a veteran teacher mentor for the new teacher's first year. Mentors worked with mentees face-to-face every month for coaching and support and discussed the progress of their mentees with the principal.

Finally, all three districts offered various types of resources for individual use, including videos illustrating classroom instruction consistent with specific dimensions of the observation framework. The districts also offered online access to lesson plans, readings, and other materials related to the observation rubric. In addition, HCPS and SCS developed "catalogs" to allow teachers to identify PD opportunities linked to specific dimensions of the observation rubric. In PPS, teachers could use various online systems to identify PD opportunities, but PPS did not have an explicit catalog.

Although these activities and systems were publicized and made generally available, none of the districts required teachers to pursue PD specifically linked to evaluation results (except for teachers identified as needing improvement). Nor did districts make substantial progress in developing systems to track teacher participation in PD, to assess the extent to which teachers followed up on the PD that was recommended, or to examine whether teachers' classroom behavior changed as a result. Prior to the initiative, all three districts had systems available for teachers to record participation in some courses and workshops, but these generally did not track teacher participation in coaching or school-based professional learning communities (Casabianca and Engberg, 2013). As part of the initiative, the districts expanded these systems in some ways—for example, to make it possible for teachers to indicate completion of online PD—but only SCS made an effort to track coaching district-wide.[9] The absence of fully developed tracking systems made it difficult for districts to determine how many teach-

[9] In SCS, TNTP made use of these data on coaching participation to study whether participation in coaching improved teachers' TE scores in the subsequent year. In other sites, some coaches might have maintained logs or spreadsheet records of their coaching.

ers took advantage of PD aligned with the new teaching rubrics, how intensively the PD focused on the rubric, or whether participating in the PD in fact helped teachers improve their practice, as measured by the evaluation system.

CMO PD Policies

During the initiative, the CMOs generally offered types of PD similar to those offered in the districts, including coaching, CMO-wide workshops, school-based workshops, and resources, such as videos for individual teacher use (see Table 6.2). However, prior to the initiative, only Aspire provided coaches; all four CMOs developed extensive coaching capacity during the initiative. As a PUC administrator looking back over the initiative observed, "It jump-started us on this coach thing. . . . And now everybody has embraced the value of coaches." Online resources were also greatly expanded during the initiative. Previously, the CMOs posted some articles and sample lesson plans, but, with the introduction of the new observation rubric, they developed videos and instructional guides linked to the rubric.

In the early years of the initiative, much PD in the CMOs focused on the instructional strategies that were embodied in the observation rubric. As in the districts, principals drew on teacher-evaluation results to make PD recommendations, and SLs and central-office staff looked at trends in the evaluations to determine school- and CMO-wide PD content. But with the advent of the Common Core State Standards, which were adopted in California (the location of almost all of the schools in the four CMOs) in 2012–2013, attention turned to supporting the implementation of the standards.

The CMOs are relatively new organizations and attract many young teachers. Because the CMO schools are located mostly in California, their new teachers were required to meet state requirements that teachers participate in induction programs to "clear" their preliminary

Table 6.2
CMO PD Policies Through 2015–2016

Policy Area	Alliance	Aspire	Green Dot	PUC
Use of TE data to identify teacher PD needs	**2011–2012 and ongoing:** Observation scores are used as a basis for feedback from the observer; SLs and central-office PD providers use trends in observation and assessment scores to plan PD offerings at the school and CMO levels.	**2011–2012 and ongoing:** Observation scores are used as a basis for feedback from the observer and to focus teachers' growth goals; SLs and central-office PD providers use trends in observation and assessment scores to plan PD offerings at the school and CMO levels.	**2011–2012 and ongoing:** Observation scores are used as a basis for feedback from the observer and to focus teachers' growth goals; SLs and central-office PD providers use trends in observation and assessment scores to plan PD offerings at the school and CMO levels.	**2011–2012 and ongoing:** Observation scores are used as a basis for feedback from the observer and to focus teachers' growth goals; SLs and central-office PD providers used trends in observation and assessment scores to plan PD offerings at the school and CMO levels.
PD provided	**Prior to initiative and ongoing:** CMO-provided large-group sessions; PD provided at school site; limited resources for individual use (see last section of table)	**Prior to initiative and ongoing:** CMO-provided large-group sessions; PD provided at school site; coaching	**Prior to initiative and ongoing:** CMO-provided large-group sessions; PD provided at school site	**Prior to initiative and ongoing:** CMO-provided large-group sessions; PD provided at school site
		2012–2013 and ongoing: Extensive resources for individual use (see last section of table)	**2011–2012 and ongoing:** Coaching	**2011–2012-and ongoing:** Coaching
	2013–2014 through 2015–2016: Coaching		**2013–2014 and ongoing:** Resources for individual use (see last section of table)	**2013–2014 and ongoing:** Resources for individual use (see last section of table)

Table 6.2—Continued

Policy Area	Alliance	Aspire	Green Dot	PUC
PD for new teachers				
What	Orientation prior to start of school	Orientation prior to start of school; follow-up PD sessions; BTSA induction coaching	Orientation prior to start of school; observation and debriefing	Orientation prior to start of school; induction workshops and observations of other teachers, BTSA induction coaching
When	2011 and ongoing	Prior to initiative and ongoing	2011–2012 and ongoing	Prior to initiative and ongoing
For whom	First-year teachers	Any teacher in his or her first or second year	Any teacher in his or her first or second year	Any teacher in his or her first year
Frequency	Two days (2012–2013 orientation was four days)	Coaching about one hour per week; one-week summer training, monthly follow-up sessions	One-week orientation; weekly observation and debriefing during one (first or second) quarter	One-week orientation; three induction days; two hours coaching per week
What	Teachers coach peers one to three periods per day			
When	2012–2013 through 2015–2016			
For whom	First- and second-year teachers			

Table 6.2—Continued

Policy Area	Alliance	Aspire	Green Dot	PUC
Frequency	The teacher receives 90 minutes of coaching per week.			
Resources for individual use	Limited number of lesson plans, videos, readings	More than 200 video clips of teachers implementing observation rubric indicators; Doug Lemov videos of effective teaching practices; Relay teacher-training videos; instructional guides for implementing observation rubric indicators. All resources are accessible on the BloomBoard internet platform.	Online instructional guides and videos for implementing observation rubric indicators, accessible on the BloomBoard internet platform	Online instructional guides and videos for implementing observation rubric indicators. Some resources are on the BloomBoard internet platform and some on the PUC intranet.
When	Prior to initiative and ongoing on internal website; 2012–2013 placed on BloomBoard internet platform	2012–2013 and ongoing	2013–2014 and ongoing	2013–2014 and ongoing
For whom	All teachers	All teachers	All teachers	All teachers
Frequency	At teacher's discretion	At teacher's discretion	At teacher's discretion	At teacher's discretion

SOURCES: Annual interviews with central-office staff in each of the IP CMOs.

NOTE: BTSA = Beginning Teacher Support and Assessment Program.

credentials—such as California's BTSA system.[10] Because the CMOs had a much higher proportion of new teachers than the districts, PD for new teachers played a larger role overall in the CMOs than in the districts.

Like the districts, the CMOs generally did not develop systems that permitted them to track or monitor teacher participation in PD or assess whether teachers who participated improved on measured aspects of teaching.

Implementation Challenges

As described earlier, all seven sites began the initiative with many types of PD already in place. The focus of their implementation efforts was in reshaping the existing PD system to align with the evaluation system and support teachers in improving on the measured dimensions of practice. Informed by our interviews with central-office staff, as well as with teachers and leaders, we found that the sites faced some challenges in accomplishing this work.

One of the chief challenges administrators noted was individualizing PD to address performance problems identified in a teacher's evaluation. A central premise of the IP theory of action was that sites would individualize PD based on teachers' evaluation results. But this proved difficult in practice. As one HCPS central-office staff member said,

> We have not been able to move in our district from the evaluation component to now, "how do we deliver the kind of professional development that that teacher needs, based on what we now know about their practice, so they can improve on their practice and become better instructors of students?"

[10] BTSA is a statewide California program offered to new teachers to provide support and mentoring to teachers in moving from a preliminary to a clear credential. It includes coaching for an average of one hour each week, observing in other classrooms, formative assessments, and reflections.

One contributing factor, cited by interviewees at all the sites, was the burden placed on the principal to both develop and offer appropriate PD and to match that PD with each individual teacher's needs. One way HCPS addressed this issue was to update its database of PD offerings to indicate the evaluation rubric indicator or component that each course or resource was designed to address. Adding this information allowed teachers to choose PD options on their own, based on the feedback they received on their classroom observations. PPS and SCS adopted similar systems.

Often, sites relied on coaches to individualize PD. For example, an administrator in Aspire said,

> most of the personalized PD [teachers] get is done through coaches, or using BloomBoard resources, or when they have their one-on-one debriefs with their principals—so they may have a data talk, or watch a video. It varies by teacher and by school.

Although coaches could, in theory, provide support aligned to teachers' needs as identified by the evaluation system, coaches did not always have access to teachers' evaluation results. In addition, available coaching time was constrained. In some sites, coaches were also full-time teachers. For example, one teacher in SCS said,

> I was a learning coach earlier this year, and it was difficult to see the teachers as often as [I] would like to, to give them the support that they need. It's a great thing to have someone on-site to assist you with improving, but, I think, to be effective, you [a learning coach] need more time.

Also, because coaches worked independently, it was hard to monitor their work and assess the quality and consistency of the support being provided. Asking coaches to document their work added burden and may have made the coaching seem evaluative rather than supportive.

In the three districts, central-office leaders struggled to develop a coherent system of PD offerings, although this was less of a challenge in some of the CMOs. All the sites sought PD systems in which the focus of the PD was aligned with the dimensions incor-

porated in the observation rubric and in which different PD activities built on one another, covered consistent content, and were not duplicative. Achieving this proved challenging, at times requiring coordination among different organizational units. For example, in SCS and Alliance, the TLE team—the group that administered and implemented the effectiveness measures—provided PD about the evaluation system and measures. Meanwhile, the PD department provided content-based PD and PD linking content with pedagogy. A separate coaching team coordinated the coaches and coaching support. In interviews throughout the initiative, central-office staff told us that PD support in SCS was not "cohesive" or "coherent." Central-office staff also reported that they struggled with developing resources that coaches could use to support teachers, as well as how to make sure coaches followed up to ensure that recommendations were implemented, and what data to collect to determine whether change was occurring in classrooms.

PPS reported a similar fragmentation of offerings and PD leadership and development during the initiative. In the summer of 2014, a new district leader charged the director of PD to work "across departments to make sure that the logistics are taken care of, but also to make sure coherence is met."[11] PPS partnered with Learning Forward to improve the coherence of PD offerings.

In HCPS, variation across schools in the amount of funding to support PD and the level of school administrator involvement led to inconsistencies in the PD offerings available to teachers at different schools. For example, some schools had access to grant funds to support extra PD offerings while other schools did not.

Although the districts struggled with coherence, these challenges were less of an issue in the CMOs, except perhaps in Alliance, which had a very decentralized management structure.

All the sites implemented online PD systems (including courses, online videos and other resources, and PD registration tools) but found it difficult to do so effectively. In the CMOs, SLs reported that teachers had technical difficulties accessing online materials, and central-office staff reported that usage was very low.

[11] District leader interview, 2014.

PPS started with a homegrown system, then adopted BloomBoard. However, staff we interviewed at the central office reported low usage among teachers and principals. SCS also experimented with several different systems during the initiative, but, according to central-office staff, usage was low. In HCPS, online courses were offered via Moodle, but central-office administrators said that the course options grew too quickly without enough effort to check the content for quality (e.g., whether the courses worked properly and were a good match to the curriculum and the TE rubric). Eventually, HCPS reduced its Moodle offerings, partly to better align course options to the rubric and partly because teachers expressed a preference for face-to-face sessions. HCPS began to prioritize blended offerings, with both online and in-person components.

Teacher and SL Reactions to PD Policies

To examine educators' perceptions of the sites' PD policies and offerings, we explore their input, from surveys and interviews, on four broad topics:

- awareness: Did teachers know what PD was available?
- endorsement: Did teachers think that the content of their PD was relevant?
- fairness: Did teachers perceive sufficient access to and support for their participation in PD?
- perceived effects: Did teachers report that their PD had been useful for improving their instruction?

Awareness

In this section, we examine teachers' awareness of the forms of PD available to them and their understanding of the links between PD and teacher evaluation. We give particular attention to two forms of PD that were particularly central to following up on evaluation results: coaching and support for new teachers.

Availability of Forms of PD

Teachers generally knew what PD was available to them. Each spring from 2013 through 2016, we asked teachers about the availability of various forms of PD.[12] Reflecting their awareness of the sites' PD offerings (as described earlier), nearly all teachers said that the following three types of PD were available:

- workshops or in-services for teachers at their school (reported by 95 to 99 percent of teachers in the districts and 86 to 97 percent in the CMOs)
- workshops, in-services, institutes, or conferences organized by the site for teachers from multiple schools (reported by 95 to 100 percent of teachers in HCPS, SCS, and PUC and 90 to 95 percent in the other sites)
- school-based teacher collaboration, such as grade-level or subject-area teams, professional learning communities, or study groups (reported by 95 to 100 percent of teachers in HCPS, SCS, Alliance, Aspire, and PUC; 90 to 95 percent in Green Dot; and 85 to 92 percent in PPS).

In addition, most teachers in the districts, but fewer in the CMOs, reported that online PD offered by or through their site was available. In HCPS and SCS, about 90 to 100 percent of teachers reported the availability of online PD in the four years in which we asked (2013–2016); the corresponding figures in other sites were 70 to 80 percent in PPS, 55 to 65 percent in Aspire, and 40 to 50 percent in PUC. In both Alliance and Green Dot, the percentage increased over time, rising from 32 percent in both sites in 2013 to 55 percent in Alliance in 2016 and 45 percent in Green Dot. Although the CMOs had many online resources for teachers, some teachers might not have considered these resources to constitute PD even if they were aware of the resources.

There was also variation across sites in the percentage of teachers reporting that videos of sample lessons were available to them in the

[12] The spring 2011 survey did not include items about the availability of specific forms of PD, except coaching.

four years in which we asked: about 90 percent in SCS and Aspire, about 70 percent in HCPS, about 65 percent in PPS, and about 60 percent in Alliance. In Green Dot and PUC, the percentages fluctuated between 60 and 80 percent from year to year. The high percentages for SCS and Aspire are consistent with these sites' efforts to develop a robust online video library, as noted in Tables 6.1 and 6.2 and as further elaborated in Appendix F.

Perceived Relationship Between Teacher Evaluation and PD

Teachers reported that their evaluations influenced what PD they received, but they reported other influences as well, including their own interests, standards, and the curriculum. In all seven sites, in each year from 2013 through 2016, majorities of teachers indicated that their evaluation results from the current year would be used to a moderate or large extent for each of the following development-related purposes:[13]

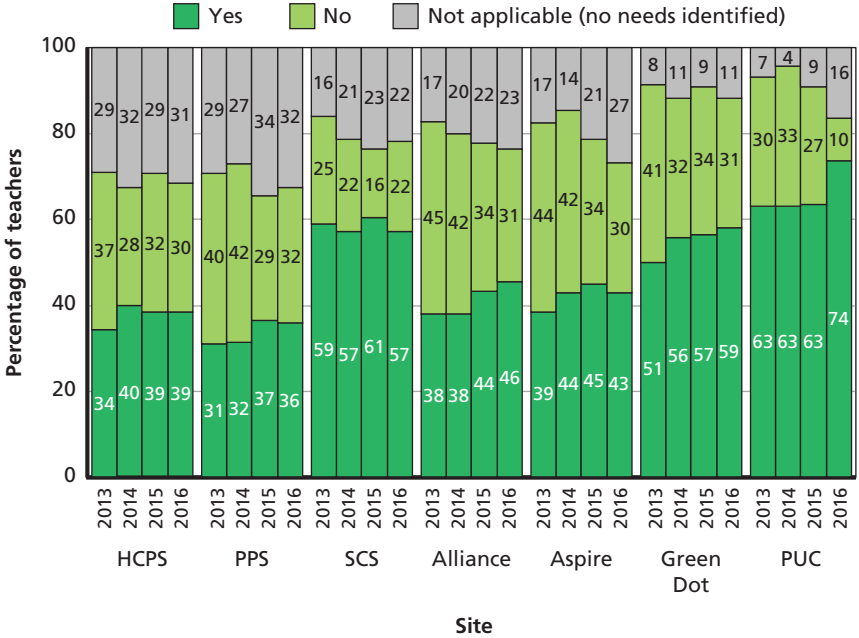
- to provide them with feedback they could use to improve their instruction (70 to 90 percent)
- to identify areas in which they needed PD (60 to 80 percent)
- to determine whether they needed additional support, such as from an instructional coach (50 to 80 percent).

In fact, these were among the most often-cited uses of evaluation results. (See Figure G.1 in Appendix G for more detail.)

As Figure 6.1 shows, lower percentages of teachers indicated that support had been made available to address needs identified in their previous year's evaluations. It is also worth noting that sizable percentages of teachers indicated that *no* needs had been identified in the previous year's evaluation, despite their expectations that the current year's evaluation would be used for purposes related to development. Not surprisingly, teachers with lower effectiveness ratings were generally

[13] We asked this of all teachers who said that their performance was being evaluated in the current school year (that is, the year during which the survey was administered). We asked about potential uses of the current year's evaluation results because we were interested in teachers' perceptions of the purposes of the evaluations they were currently undergoing.

Figure 6.1
Teachers' Responses to the Survey Question, "Has Support (Coaching, Professional Development, etc.) Been Made Available to You to Address the Needs Identified by Your [Previous Year's] Evaluation Results?" Springs 2013–2016



NOTE: We did not ask this question in 2011.
RAND *RR2242-6.1*

more likely than teachers with higher effectiveness ratings to say that needs had been identified.[14] Lower-rated teachers were also typically more likely than higher-rated teachers to say that support had been available to address the identified needs.[15]

---

[14] Novice teachers were also more likely than more-experienced teachers to say that needs had been identified by their previous year's evaluation, but these differences were less pronounced than those by effectiveness rating and were less likely to be significant (perhaps because only second-year teachers were included in the experience comparisons, given that first-year teachers did not have previous-year ratings).

[15] There were, however, some exceptions to this finding in some sites and years, and observed differences between lower-rated and higher-rated teachers were often not statistically signifi-

Although many teachers reported that their evaluation results influenced their PD, they reported other influences as well. Teachers were more likely to say that the PD in which they had participated had been influenced by needs and interests they had identified themselves or by priorities set by their schools or sites for multiple teachers than they were to say that it had been influenced by needs identified as part of a formal evaluation of their teaching or from informal feedback they had received. The survey did not specify the potential sources of informal feedback, but teachers might have received feedback from peers, mentors, coaches, or other staff outside of the formal evaluation process. Figure 6.2 shows results for the spring of 2016. Results for earlier years (shown in Figure G.2 in Appendix G) were generally similar, although the percentage indicating the influence of needs identified as part of formal evaluation increased over time in SCS and the CMOs, as did the percentage for informal feedback.[16] This is consistent with policy changes in these sites, such as the introduction of the coaching model in SCS. In PPS, the percentage of teachers indicating that needs identified as part of formal evaluation influenced their PD was notably lower than the percentage in the other sites starting in 2013. This is not because the PPS percentage declined after 2011 but rather because it held steady while the percentages in other sites increased, particularly (as just noted) in SCS and the CMOs.

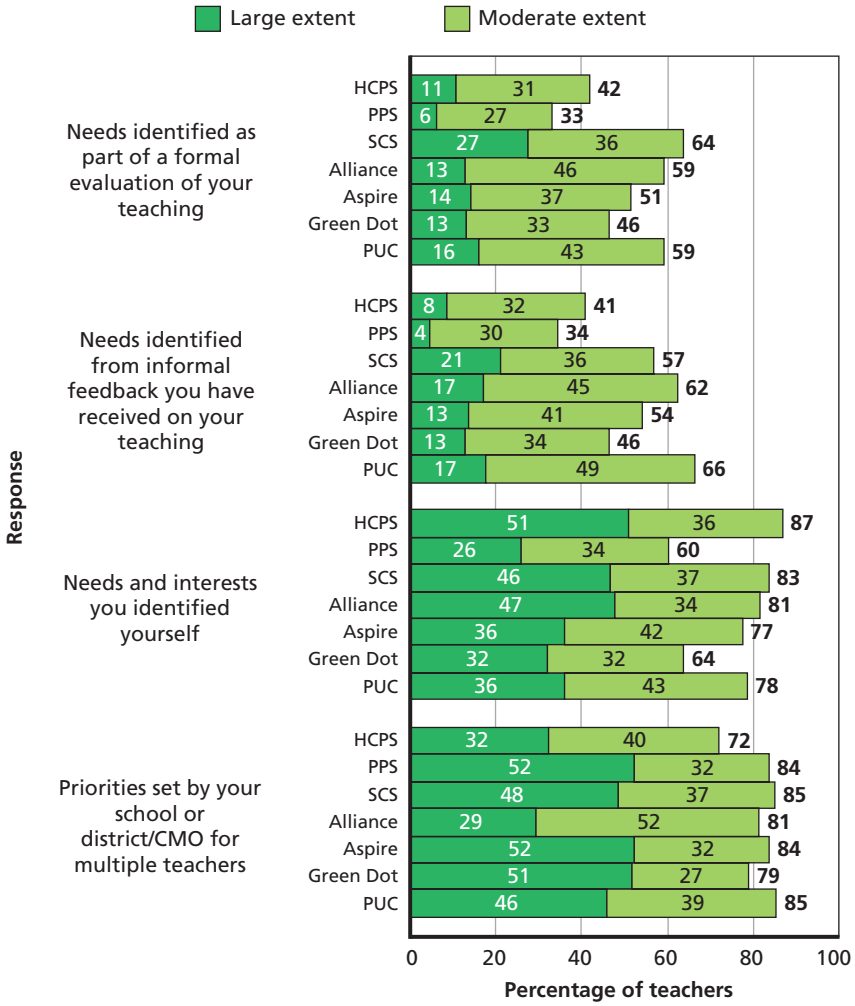
In addition to asking teachers what influenced their choice of PD, we asked them about the focus of the PD in which they participated. Teachers were more likely to report that their PD had been aligned with standards (such as the Common Core State Standards) or curriculum than they were to say that their PD had been aligned with elements of their sites' teacher-observation rubrics. Figure 6.3 illustrates this finding for 2016; the pattern in other years was similar (see Figure G.3 in

---

cant, perhaps because of the small number of low-rated teachers.

[16] For needs identified as part of formal evaluation of teaching, the increase from 2011 to 2016 in the percentage reporting influence on PD (to a moderate or large extent) was statistically significant ($p < 0.05$) in SCS and the four CMOs. For needs identified from informal feedback received on teaching, the increase from 2011 to 2016 in the percentage reporting influence on PD (to a moderate or large extent) was statistically significant ($p < 0.05$) in every site except HCPS.

Figure 6.2
Teachers' Responses to the Survey Question, "To What Extent Did Each of the Following Influence What Professional Development You Participated in This Year (2015–2016, Including Summer 2016)?"

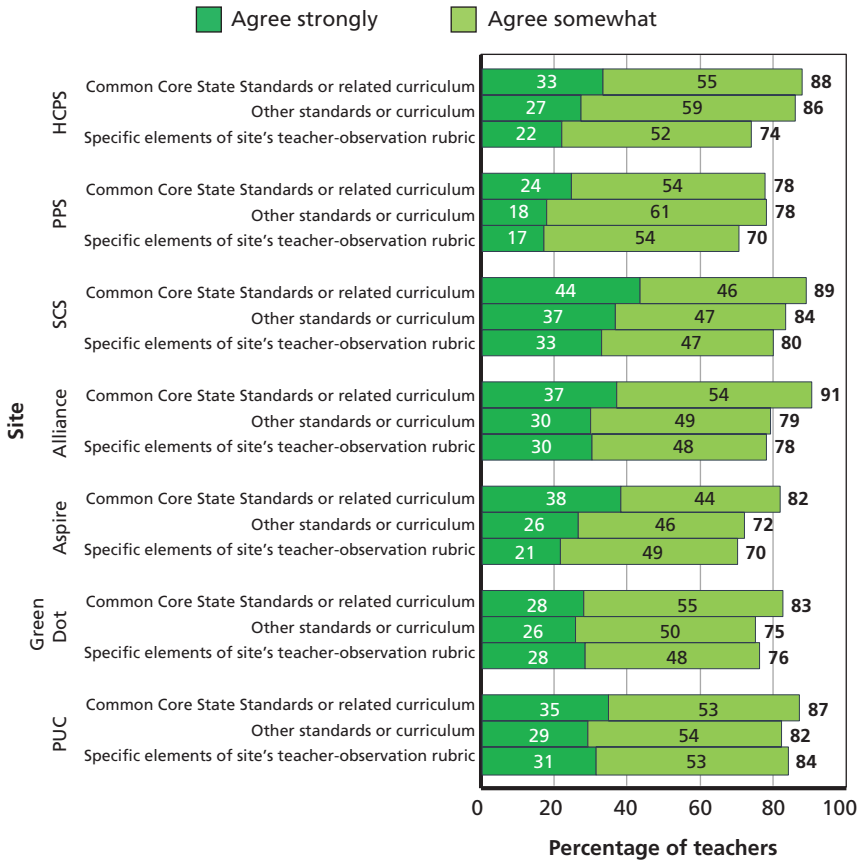


NOTE: Omitted response categories are "small extent" and "not at all."

RAND *RR2242-6.2*

Appendix G). Even so, large majorities of teachers—typically 70 to 80 percent—indicated alignment of PD with rubric elements.

Figure 6.3
Teachers' Agreement That Their PD During the 2015–2016 Year Was Aligned with Various Sources, Spring 2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly."

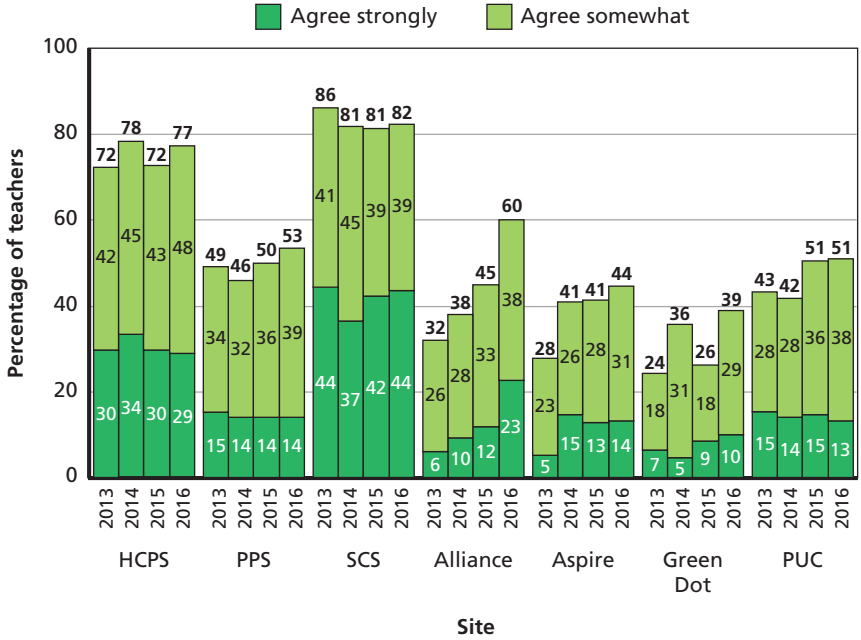
RAND *RR2242-6.3*

One strategy some sites used to help teachers locate PD that might help them address needs identified through evaluation was to develop a catalog of PD opportunities aligned with the observation rubric. Only in HCPS and SCS did a large majority of teachers agree that they had had "easy access" to a catalog of this kind. This is consistent with the fact that both HCPS and SCS took this approach (see Table 6.1). How-

ever, the percentage of teachers agreeing with this statement increased over time in the other five sites, most notably Alliance (see Figure 6.4).[17]

As another way of examining whether PD was linked to teachers' evaluations, we asked SLs how often they provided teachers with suggestions for PD based on their classroom observations.[18] As Figure 6.5

Figure 6.4
Teachers' Agreement That They Had Had Easy Access to a Catalog of PD Opportunities Aligned with Their Site's Teacher-Observation Rubric, Springs 2013–2016



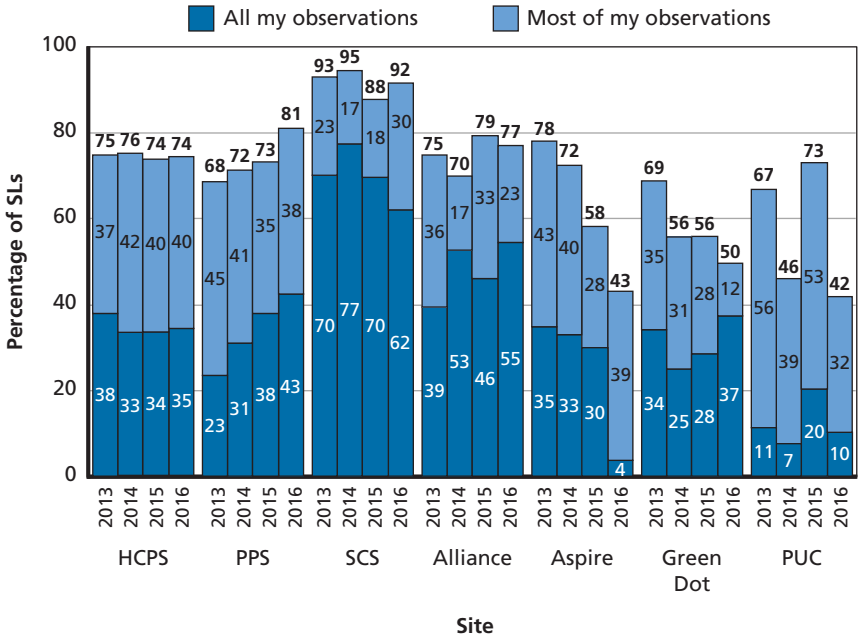
NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly." We did not ask this question in 2011.

RAND *RR2242-6.4*

[17] The increase in the percentage agreeing (strongly and somewhat) from 2013 to 2016 was statistically significant ($p < 0.05$) in HCPS and all four CMOs. In PPS, the increase from 2014 to 2016 was significant.

[18] The survey instructed SLs to "consider all observations that counted toward teachers' evaluations, regardless of length."

Figure 6.5
SLs Reporting That They Provided Teachers with Suggestions for PD Based on Their Observations of the Teachers' Instruction, Springs 2013–2016



NOTE: We asked this question only of SLs who reported that they observed teachers' instruction as part of the teachers' evaluations in the current school year. Omitted response categories are "some of my observations" and "none of my observations."

RAND *RR2242-6.5*

shows, SCS stands out as having particularly high percentages of SLs reporting that they provided suggestions for PD in most or all of their observations: above 85 percent in all four years the question was asked (2013 through 2016). One possible explanation for the high percentages in SCS is that SCS principals were expected to document their suggestions for PD in an online system. In HCPS, PPS, and Alliance, the cross-year average was about 75 percent, although the percentage increased over time in PPS. The percentages were lower in Aspire, Green Dot, and PUC, and, in Aspire, the percentage decreased from year to year, ending in 2016 with less than half of the SLs reporting that they had provided suggestions for PD in most or all of their obser-

vations. The decline in Aspire might have been a result of improvement in TE, reducing the need for PD, but other evidence suggests that this is unlikely (see Figure 10.1 in Chapter Ten and Figure 13.3 in Chapter Thirteen).

Coaching

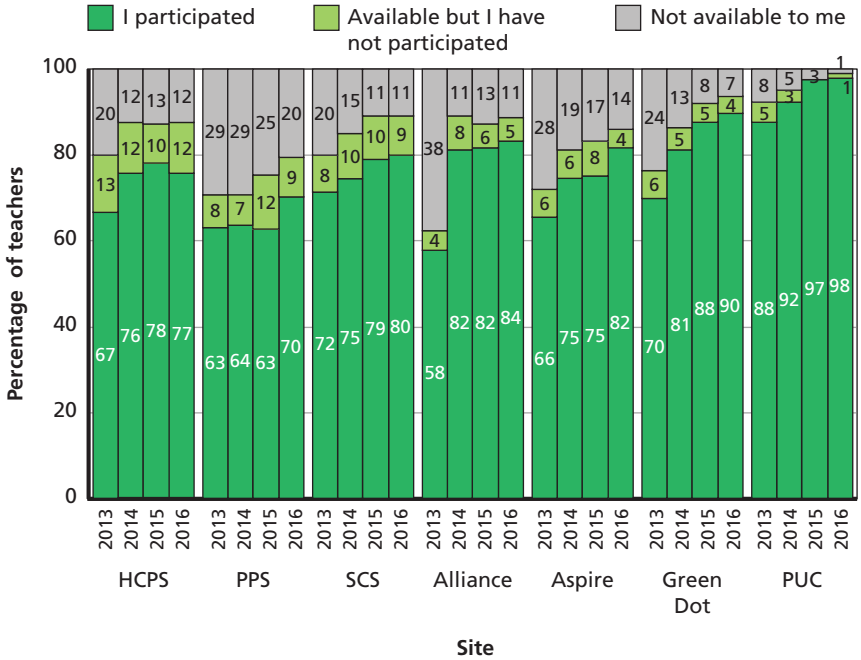
Most teachers had access to some form of coaching during the initiative, and the percentage of teachers reporting access increased over time. As discussed earlier in this chapter, most of the sites increased the amount of coaching they offered during the initiative, particularly to novice teachers or to teachers who were struggling. For example, HCPS implemented its peer mentoring program for new teachers, which was very popular; SCS offered coaching focused on new teachers and struggling teachers. And most of the CMOs vastly increased their coaching capacity during the initiative (often having started with little or no coaching capacity prior to the initiative).

Demonstrating awareness of these policies, majorities of teachers in all the sites reported that instructional coaching was available to them, and, in all sites, the percentage of teachers reporting the availability of such coaching increased from 2013 through 2016 (see Figure 6.6). An increasing (and high) percentage of teachers also reported receiving this kind of coaching.

However, in a separate survey question that asked teachers whether they had received "individualized coaching or mentoring (that is, from a staff member designated by your school or [site] to provide you with coaching or mentoring)," the percentage of teachers reporting receiving such coaching was substantially lower, particularly in the three districts (see Figure 6.7).[19] We are not certain why the percent-

[19] The full survey item began, "This school year (2015–16, including summer 2016), have you received any *formal, individualized coaching or mentoring* (that is, from a staff member designated by your [district or CMO] or school to provide you with coaching or mentoring)?" The survey went on to say, "By 'individualized,' we mean coaching or mentoring that is targeted toward specific, individual teachers or groups of teachers, but not a school's teaching staff as a whole." The item offered respondents three response options: "Yes, I have received one-on-one coaching or mentoring," "Yes, I have received coaching or mentoring as part of a group of teachers," and "No, I have not received any formal individualized coaching or mentoring this year." The "no" option was exclusive, but respondents could check one or

Figure 6.6
Teachers' Responses About the Availability of, and Their Participation in, "Instructional Coaching (Provided by School-Based Coaches *or* District Coaches)," Springs 2013–2016



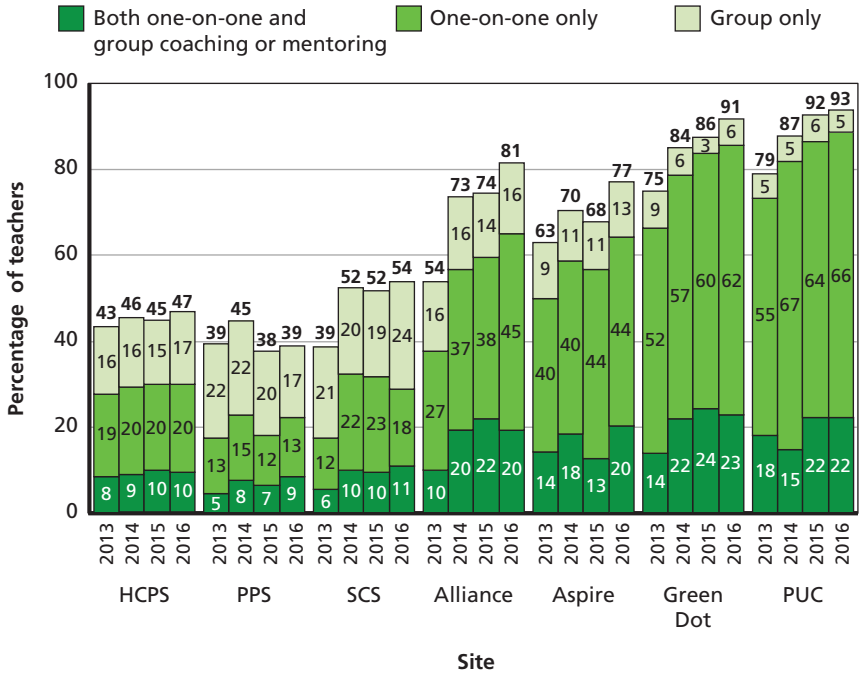
NOTE: We did not ask this question in 2011.
RAND *RR2242-6.6*

ages on this item were lower than on the former item, but it seems that some survey respondents interpreted "individualized coaching or mentoring" as having a more restricted definition than "instructional coaching," perhaps because the item guided teachers to include coaching "targeted toward specific teachers or groups of teachers" rather than all teachers in the school.[20]

both of the first two options. In Figure 6.7, the omitted response category is the "no" option, which means receipt of neither one-on-one nor group coaching or mentoring.

[20] The two survey questions also differed in the way they were structured, with the "instructional coaching" question being one subitem in a large block listing several forms of PD (and which included availability, participation, and perceived usefulness in a single set of response

Figure 6.7
Teachers' Responses About "Receiving Instructional Coaching (Provided by School-Based Coaches *or* District Coaches)," Springs 2013–2016



NOTE: Omitted response category is "neither." We did not ask this question in 2011.
RAND *RR2242-6.7*

Novice teachers were much more likely than other teachers to report receiving such individualized coaching or mentoring, as were teachers with low effectiveness ratings. For example, in HCPS in 2016, 98 percent of novice teachers and 84 percent of teachers with low effectiveness ratings reported receiving this coaching, compared with only 47 percent of teachers generally.

options), whereas the "formal, individualized coaching or mentoring" question was a stand-alone item that asked only about the receipt of such coaching. We cannot rule out that the differences in the two sets of responses could have been caused by the different structures of the questions.

New-Teacher Support

As noted in Tables 6.1 and 6.2, all the sites implemented or continued supports (in addition to coaching) for new teachers. In both 2015 and 2016, in all the sites, majorities of novice teachers reported the availability of a "[site-] or state-sponsored induction program, mentoring, or other professional development activities for new or beginning teachers." The percentage of novice teachers reporting the availability of this form of support was highest in HCPS (99 percent in 2015; 98 percent in 2016), but it was also quite high in PPS (about 95 percent in both years). The percentages tended to be lower in the CMOs (for example, 76 percent in Green Dot in 2015 and 72 percent in Aspire in 2016). Not all of the CMOs offered their own induction programs—teachers in these CMOs had to participate in external induction programs, such as at a university—so that could explain the lower percentages of novice teachers reporting availability of an induction program in some of the CMOs. In addition, because of the high proportion of novice teachers in the CMOs, there might not have been enough induction mentors to go around.

Endorsement

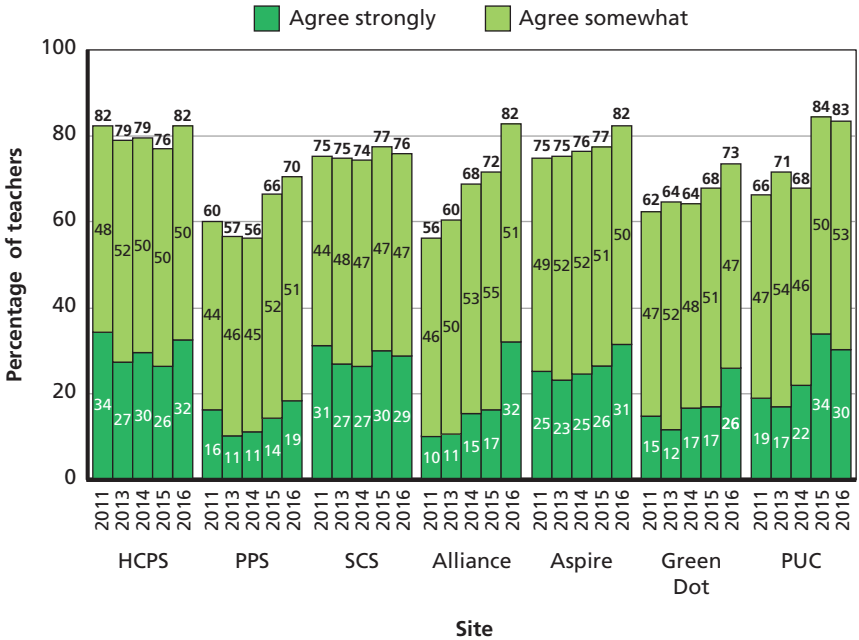
In all the sites, majorities of teachers thought that their PD was relevant and coherent. Teachers in all the sites agreed that their PD experiences had been relevant to the needs of their students. From 2011 through 2016, the percentage agreeing was highest in HCPS and SCS (75 to 80 percent); in the other sites, the percentage agreeing started out much lower but rose considerably over the years (see Figure 6.8).[21] The pattern was generally similar for teacher agreement that their PD experiences had been coherently related to one another, although this finding stands in contrast to the fragmentation noted by central-office staff, as discussed earlier.

Teachers' perceptions of the usefulness of their PD are presented under "Perceived Effects" later in this section.

---

[21] The increase from 2011 to 2016 in the percentage agreeing (strongly and somewhat) was statistically significant ($p < 0.05$) in PPS and all four CMOs.

Figure 6.8
Teachers' Agreement That Their PD Experiences in That School Year Had Been Relevant to the Needs of Their Students, Springs 2011–2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly." We did not survey teachers in 2012.

RAND *RR2242-6.8*

Fairness

On the face of it, the notion of fairness does not seem obviously applicable to PD (as opposed to, say, tenure or dismissal). For example, one seldom hears complaints about PD being "unfair" to teachers. Yet there are at least two respects in which fairness could, in fact, relate to PD. First, if teachers are not able to access PD in which they are required—or expected—to participate, or if some teachers for whatever reason have greater access than others, either of those conditions could be considered a fairness issue. Similarly, if teachers are required or expected to participate in PD but are not provided with support (e.g., time) to do so, that could also be considered a fairness issue. The teacher and SL

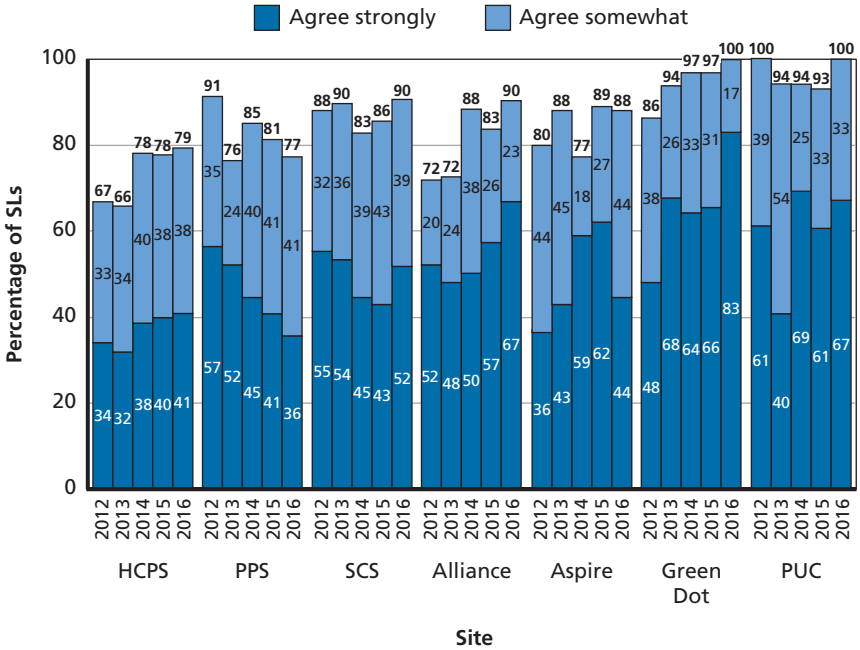
surveys included questions about access to and support for PD, and we present results in this section.

In all the sites, most SLs thought that teachers had adequate access to site-provided PD. We asked SLs whether site-provided PD offerings (e.g., workshops, in-services, online courses) had enough slots and availability to accommodate all teachers who wished or needed to participate. In all sites, a majority of SLs agreed at least somewhat. The percentage of SLs agreeing increased in HCPS and three of the CMOs and declined in PPS (see Figure 6.9).[22]

CMO teachers' perceptions improved in relation to scheduling flexibility and resources for PD, but this was not the case for teachers in the districts. We asked teachers whether they had had sufficient flexibility in their schedules to pursue PD opportunities of interest to them. The three districts experienced declines in the percentage of teachers agreeing from 2011 to subsequent years, while the CMOs experienced increases. There were also increases in the CMOs in the percentage of teachers agreeing that sufficient resources (substitute coverage, funding to cover expenses, stipends) had been available to allow them to participate in needed PD, but there was little change over time in SCS (with 56 to 60 percent of teachers agreeing), fluctuation in HCPS (44 to 53 percent agreeing with no discernible trend), and a decline in PPS (from 55 percent in 2011 to 46 percent in 2014, 2015, and 2016). (See Figure G.4 in Appendix G.) Interviews in PPS suggested a possible reason for the decline: In 2014–2015, the district decreased the amount of release time for PD so that teachers would spend more time in their classrooms. This change was not necessarily unwelcome to teachers, however, especially after the composite effectiveness measure was implemented with stakes attached: Some teachers with whom we spoke expressed reluctance about taking time away from teaching, even for PD, because of concerns that substitute teaching could have the effect of decreasing their VAM scores.

[22] The increase in the percentage agreeing (strongly and somewhat) from 2012 to 2016 was statistically significant ($p < 0.05$) in HCPS, Alliance, and Green Dot. The decrease in the percentage agreeing was statistically significant ($p < 0.05$) in PPS.

Figure 6.9
SLs' Agreement That Site-Provided PD Offerings Had Enough Slots and Availability to Accommodate All Teachers Who Wished or Needed to Participate, Springs 2012–2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly." The question as it appeared in the survey also had a "don't know" option, but we coded responses of "don't know" as missing in the calculation of the percentages shown. We did not ask this question in 2011.

RAND *RR2242-6.9*

Perceived Effects

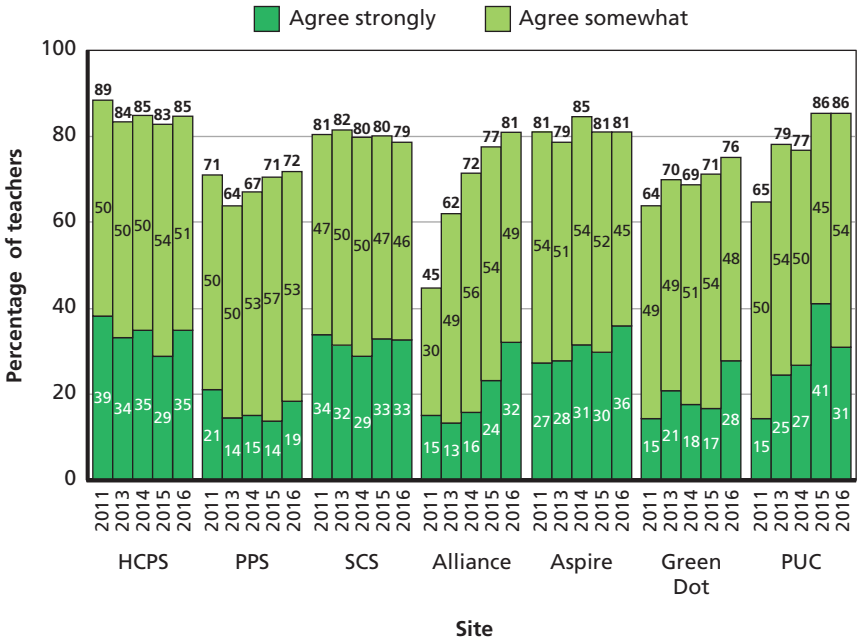
We asked teachers for their perceptions of the usefulness of the PD in improving their instruction and student learning, as well as the helpfulness of specific types of PD, the value of PD in meeting needs identified through the evaluation system, and the usefulness of coaching and mentoring. This section considers these issues.

Perceived Usefulness of PD in Improving Instruction and Student Learning

In every site, a high or increasing percentage of teachers said that PD had helped them improve their instruction and student learning. A consistently high percentage of teachers in HCPS, SCS, and Aspire and an increasing percentage in the other sites indicated that their PD experiences in the current year had been useful for improving their instruction and for enhancing their ability to improve student learning. (Figure 6.10 shows the former; the picture for the latter is almost identical.)

We also asked teachers about the extent to which, as a result of the PD in which they had participated, their knowledge and skills

Figure 6.10
Teachers' Agreement That Their PD Experiences in That School Year Had Been Useful for Improving Their Instruction, Springs 2011–2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly." We did not survey teachers in 2012.

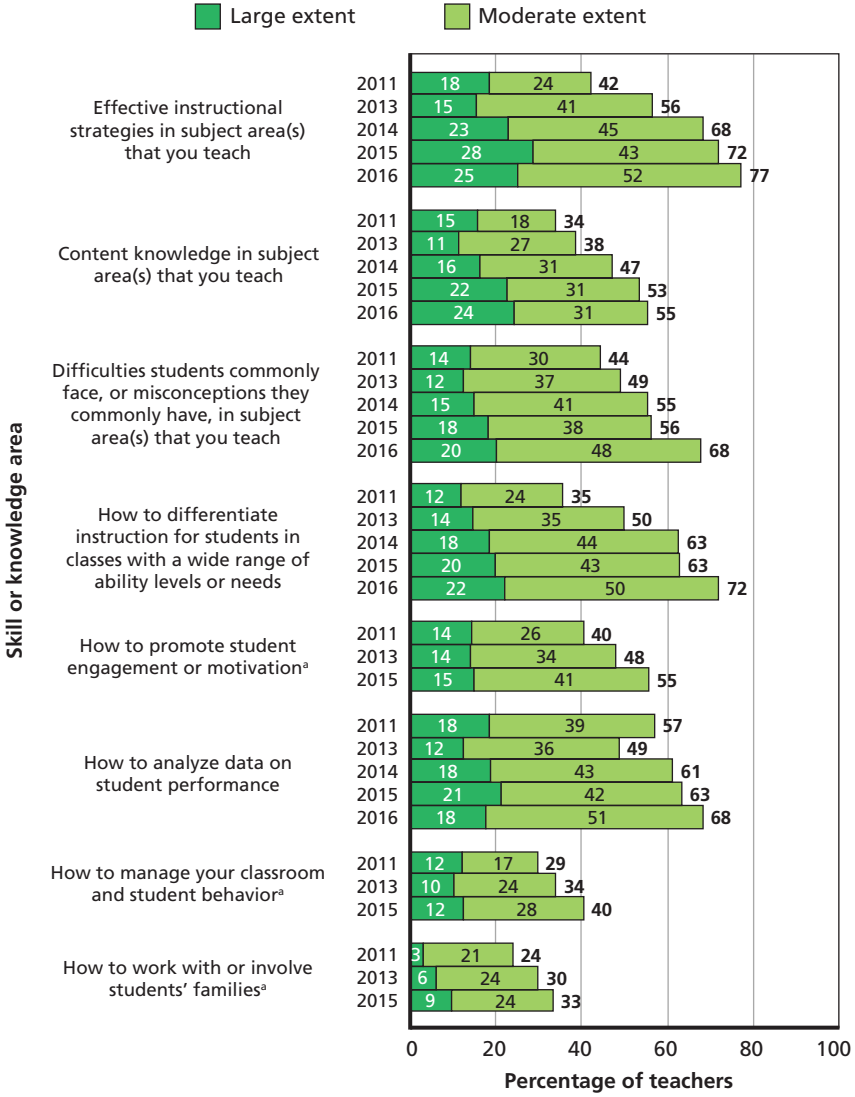
RAND *RR2242-6.10*

had been enhanced in specific areas, including instructional strategies, content knowledge, and differentiation of instruction. Figure G.5 in Appendix G presents results for each site in each year and each area. The results indicate the following:

- In most of the areas, the percentage of CMO teachers who said that their knowledge and skills had been enhanced to a moderate or large extent increased from 2011 to 2016. For example, in Alliance, the percentage of teachers indicating that PD had improved their ability to "differentiate instruction for students in classes with a wide range of ability levels or needs" increased from 35 percent in 2011 to 72 percent in 2016. In Aspire, it increased from 46 to 53 percent; in Green Dot, from 42 to 52 percent; and, in PUC, from 43 to 73 percent.
- In most of the areas, the percentage of teachers saying that their knowledge and skills had been enhanced to a moderate or large extent was lower in PPS than in HCPS and SCS; in the later years, it was also lower than the percentage in the CMOs. For example, in 2016, 34 percent of PPS teachers indicated that their ability to differentiate instruction had been enhanced by their PD, a much lower proportion than the 65 percent in HCPS and 66 percent in SCS.
- Of the areas listed, the highest percentage of teachers (across years and sites) tended to report enhancement of knowledge and skills related to their "familiarity with effective instructional strategies" in subject areas they teach. The two areas that tended to have the lowest percentage of teachers reporting enhanced knowledge and skills as a result of PD were classroom and behavior management and working with students' families.

Increases in the percentages of teachers reporting PD as useful were particularly notable in Alliance. The increase over time in the percentages reporting enhancement is particularly striking in Alliance, as shown in Figure 6.11. No other site experienced such a systematic increase (see Figure G.5 in Appendix G).

Figure 6.11
Alliance Teachers Reporting Enhanced Skills and Knowledge, in Various Areas, Resulting from PD, Springs 2011–2016



NOTE: Omitted response categories are "small extent" and "not at all."
[a] Not asked in 2014 or 2016. We did not survey teachers in 2012.

RAND *RR2242-6.11*

Helpfulness of Specific Types of PD

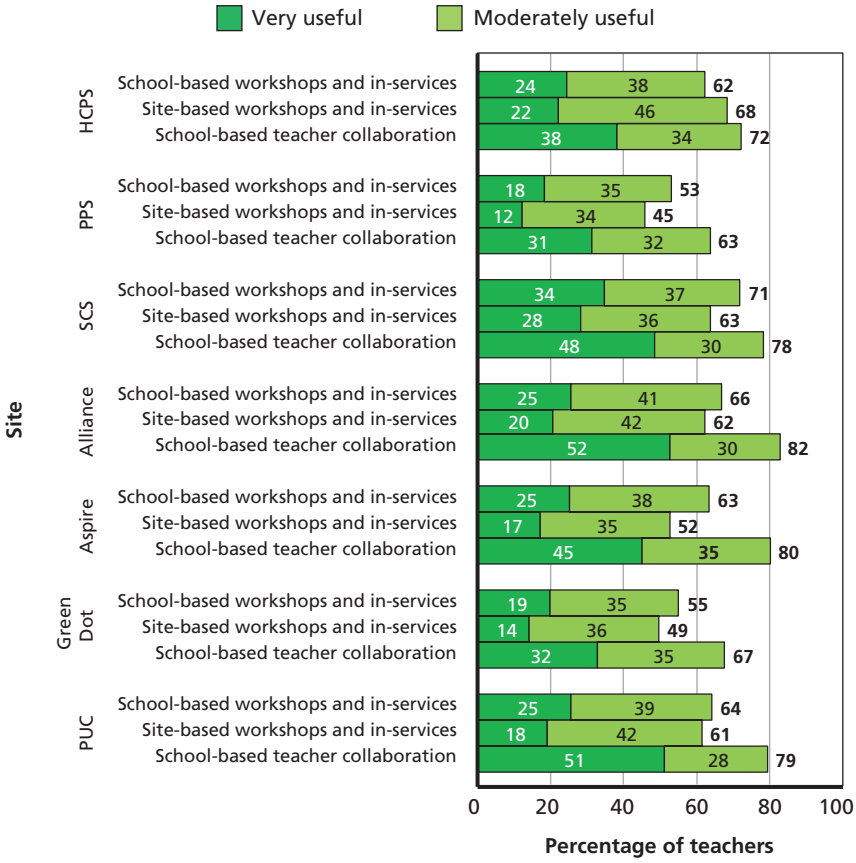
Teachers were more likely to say that teacher collaboration was useful than they were to say that workshops and in-services were useful. In addition to asking teachers whether the PD in which they participated was useful for improving specific types of skills, we asked teachers about the usefulness of specific forms of PD. Figure 6.12 shows the 2016 results for three forms of PD in which nearly all teachers participated: school-based workshops and in-services, district-organized (or, in the case of the CMOs, CMO-organized) workshops and in-services, and school-based teacher collaboration. In most years and most sites, school-based teacher collaboration was the form of PD rated as moderately or very useful by the highest percentage of teachers. The full set of results for all years and forms of PD is in Figure G.6 in Appendix G.

Interviews also indicated that teachers perceived collaboration with other teachers to be valuable, and they often wished for more of it. Example quotes include the following:

> Working in small groups is the most useful PD because we can put together the experienced and new teachers and talk about a topic and find ways to collaborate. If that PD were more common, I would respect PD more. The tone I get from other teachers is [that] PD is a waste of time. We want to improve, but what we get isn't helpful. We need fewer lectures. We work hard to know what we're doing—give us space to collaborate. That would be more useful. (PPS teacher, 2014)

> [There is] not enough [collaboration], but it's the best professional development we get. So whenever we do have collaboration time with other school sites within the organization, those tend to be the most effective, at least for the [teachers in my subject area], because you're able to trade ideas and strategies and content. At the school-site level, it's difficult because there are only two [teachers in my subject area], and each teacher also teaches multiple grade levels, so it's hard to do grade-level collaboration or content-level collaboration at the school-site level, but it is more possible at the PUC-wide level. (PUC teacher)

Figure 6.12
Teachers' Perceptions of the Usefulness of Three Forms of PD, Spring 2016



NOTE: Only teachers who reported having participated in a given form of PD rated the usefulness of that form. Omitted response categories are "slightly useful" and "not at all useful."

RAND *RR2242-6.12*

> What I would like to see—which does not happen as much as I would like to see—is more curriculum-based teams where all the [teachers from a particular subject or grade level] can get together at the same time and talk about what we're doing. Right now, the only time that happens is when we're passing or I go out of my way to go find the other . . . teacher who's teaching the same thing I am and saying, "How did you do this? Let's talk about this."

> "This is what I did that worked really well. Do you want to try it?" We don't get enough of that. (HCPS teacher)

However, some teachers also noted that, although collaboration is valuable, scheduling effective collaboration was not always easy. One Green Dot teacher, for instance, in recalling a cross–subject area collaboration with two other teachers, commented, "It was great," but that "it takes a perfect storm of you all to align your curriculum."

Of teachers who reported participating in online PD, typically no more than half rated the PD as moderately or very useful. The same was true for videos of sample lessons, although, for videos, the percentage rating them moderately or very useful had a cross-year average exceeding 50 percent in two of the seven sites: SCS and Aspire, the two sites that prioritized the development of a video library. Even so, some SCS teachers we interviewed complained that the videos were not sufficiently applicable to their situations because the videos did not take place in Memphis and the students did not resemble their students or did not match their subject area or grade level. One teacher commented,

> A lot of time in the PD, they will have a session where you'll watch a video of a teacher teaching and you look at the rubric and have to rate the teacher. It helps you learn the rubric and helps you see what an actual teacher is doing wrong or right. The only huge problem is that the examples are always elementary teachers usually. . . . We had one a couple weeks ago, and it was an elementary teacher. A lot of time, I don't know how realistic [the video] is. A lot of times, you'll see a classroom with seven kids in there and I've got 35.

For Alliance, the same pattern of increases over time we saw earlier about the content of PD was also present for the form of PD. For each of seven forms of PD, the percentage of Alliance teachers rating that form of PD as moderately or very useful increased each year from 2013 through 2016 (see the Alliance column of Figure G.6 in Appendix G).

Perceived Usefulness of PD in Meeting Needs Identified Through the Evaluation System

We asked teachers who reported having support available to address needs identified by their prior-year evaluation results to rate the usefulness of that support—that is, the extent to which the support helped them address the identified needs. In SCS and Alliance, about 70 percent of teachers said that the support had helped them to a moderate or large extent. The corresponding percentages in HCPS and Green Dot were about 55 and 60 percent, respectively. Aspire and PUC both had increasing percentages of teachers reporting that the support had helped to a moderate or large extent: In Aspire, the percentage rose from 61 percent in 2013 to 76 percent in 2016, and, in PUC, it rose from 70 percent 2013 to 91 percent in 2015 but then fell somewhat to 81 percent in 2016. In PPS, there was no clear trend: The percentage was 64 percent in 2013 and fell to 47 percent in 2014 but was above 60 percent in the later years.[23]

In interviews, some teachers noted that the feedback they received as part of their evaluations—especially their observations—was a type of PD, and frequently a useful one. As one Alliance teacher commented in 2015, "When we have our preconference evaluations and our postconference evaluations, that type of professional development helps me the most. And that's where I see improvement in my overall teaching performance."

Other teachers we interviewed commented that they found the observation rubric itself to be a useful resource to draw on in self-

[23] In PPS, 100 percent of novice teachers reported in 2013, 2014, and 2015 that the support had helped them address the identified needs (significantly differing, at $p < 0.001$, from the percentage of more-experienced teachers reporting that the support had helped). In other sites, especially HCPS, novice teachers were typically more likely than more-experienced teachers to say that the support had helped them address the identified needs, but not all of these differences were statistically significant ($p < 0.05$), perhaps because only second-year teachers were included in the experience comparisons, given that first-year teachers did not have prior-year ratings.

directed efforts to improve instruction. One PPS teacher described it thusly:

> I rely mostly on RISE [information to inform my practice]. That tool helps me to reflect on what I'm doing. After a one-year period, I revisit it and honestly think about it [and] compare it with other years. I focus most on getting better as a teacher and on meeting the needs of students who can't keep up with the rest of the students in class. [Component] 3f or VAM, that pretty much shows the administration that I'm moving my kids at an adequate pace, which is not so valuable to me [because it reflects more on the children than on the teaching]. Tripod results—I think that's more confirmation for me of students' perspective. . . . [So] it's really the RISE tool that I use for being reflective about my craft.

Teachers with lower effectiveness ratings were generally no more likely than higher-rated teachers to say that the support they received had helped them. In fact, it was more often the case, across sites and years, that the highest-rated teachers were significantly more likely than lower-rated teachers to say that the support had helped them address the identified needs (to a moderate or large extent). In our interviews, we heard occasional comments from principals and central-office staff in PPS and SCS that teachers' motivation to pursue and utilize PD tended to be positively correlated with TE scores (that is, teachers who were more effective to begin with found PD more valuable).

Perceived Usefulness of Coaching

Among teachers who reported receiving instructional coaching, the coaching was rated as moderately or very useful by 50 to 80 percent of teachers in each CMO (with the percentage rising over time in Alliance, Green Dot, and PUC), 55 to 65 percent of teachers in SCS, 50 to 60 percent in HCPS, and 37 to 47 percent in PPS. The increases in the CMOs again illustrate the improvement of teacher perceptions about PD in these sites.

Among teachers who reported receiving *individualized* coaching or mentoring, typically 60 to 80 percent rated it as moderately or very useful. In Aspire and Green Dot, novice teachers were consistently

significantly more likely than more-experienced teachers to rate their individualized coaching and mentoring as moderately or very useful. This was generally true in the other sites as well, but, in the other sites, the difference between novice and more-experienced teachers tended to be smaller and was often not statistically significant. Particularly in PPS and SCS, novice teachers were less likely than more-experienced teachers to say that they had received a sufficient *amount* of individualized coaching or mentoring. Overall, about 70 to 80 percent of teachers who received individualized coaching or mentoring thought that they had received a sufficient amount of such support.

Perceived Usefulness of New-Teacher PD

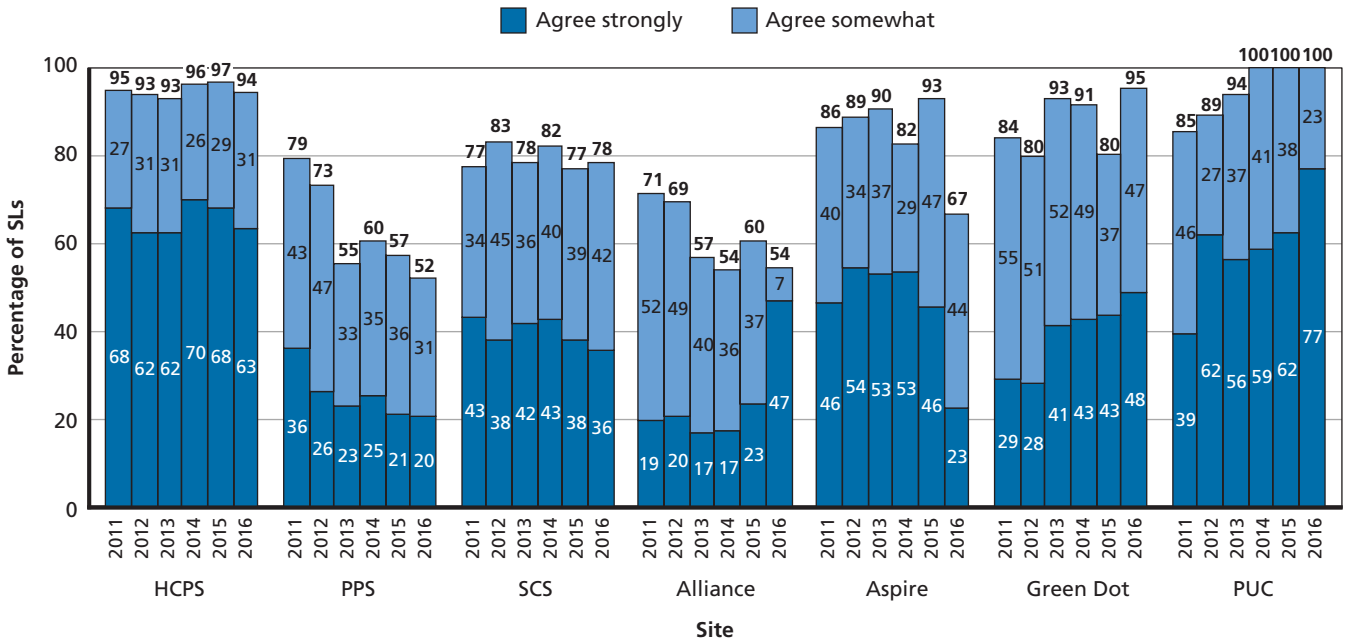
Among novice teachers who reported having participated in induction or PD for new teachers, there was considerable variation by site, and in some sites by year, in the percentage who thought it had been moderately or very useful. Sites with relatively consistent percentages in 2015 and 2016, when the question was asked, were HCPS (about 70 percent), SCS (about 60 percent), and PUC (just under 85 percent). In the other sites, differing percentages in the two years make the results difficult to interpret.

In addition to asking new teachers about their PD, we asked SLs about PD for new teachers. As shown in Figure 6.13, there was considerable variation across sites in the percentage of SLs who agreed that their site did "a good job of meeting the professional development needs of teachers *new to the profession*." A high percentage of SLs in HCPS and PUC agreed, whereas the percentage was smaller—and declining—in PPS and Alliance. Where there were declines, this could reflect decreasing satisfaction with PD opportunities for new teachers, but it could also (or instead) reflect increasing awareness of new teachers' needs.

Summary

Advocates of teacher-evaluation reform have argued that teacher evaluation can serve as a valuable source of information to guide teachers'

Figure 6.13
SLs' Agreement That the Site Did a Good Job of Meeting the PD Needs of Teachers New to the Profession, Springs 2011–2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly." The question as it appeared in the survey also had a "don't know" option, but we coded responses of "don't know" as missing in the calculation of the percentages shown.

RAND *RR2242-6.13*

PD. The basic premise is that, by identifying a teacher's strengths and weaknesses, evaluation results can help determine areas in which that teacher should invest his or her development efforts. To set the groundwork for linking evaluation and PD, proponents believe that several tasks must be carried out, including inventorying the PD available in a district or CMO, assessing its quality and alignment with the evaluation system, fostering embedded forms of PD (such as coaching and collaborative teams), and establishing the infrastructure to support SLs and teachers in identifying appropriate PD and tracking participation.

All seven of the IP sites offered multiple types of PD: coaching and mentoring, district- or CMO-wide workshops, school-based workshops, school-based teacher collaboration, access to online materials (lesson plans and readings), and access to videos of lessons illustrating teaching practices aligned with the site's observation rubric. In general, the sites succeeded in inventorying their PD and making catalogs available to teachers and leaders identifying PD aligned with the site's observation rubric, although, except in HCPS and SCS, many teachers did not report having access to a catalog. Principals and other staff recommended PD based on teachers' evaluation results. However, sites did not require teachers to participate in recommended PD, nor did they put systems in place to monitor teachers' participation or determine whether teachers' effectiveness improved after participation.

The sites confronted several challenges in linking PD to evaluation results. One of the chief challenges was the burden placed on the principal to identify appropriate PD and to match that PD with each individual teacher's needs. In addition, in each of the three districts, central-office leaders struggled to develop a coherent system of PD offerings aligned with the district's vision for teaching and learning as articulated in the observation rubric; this was less of a challenge in some of the CMOs. Finally, although all the sites tried to implement online PD systems (including courses, online videos and other resources, and PD registration tools), they encountered technical difficulties in doing this effectively.

Data from surveys and interviews indicate that teachers in all sites generally knew what PD was available to them. Furthermore, as intended, teachers reported that their evaluations influenced the

choices of PD in which they participated, although they reported other influences as well, including their own interests, as well as standards and curriculum. Not surprisingly, teachers with lower effectiveness ratings were generally more likely than teachers with higher ratings to say that their evaluations had identified needs for PD. Lower-rated teachers were also typically more likely than higher-rated teachers to say that support to address the identified needs had been available, which is consistent with the initiative's theory of action.

Most teachers had access to some form of coaching, on which the sites often relied to individualize PD, and the percentage of teachers reporting access to coaching increased over time. Teachers with lower effectiveness ratings were much more likely than other teachers to report receiving individualized coaching or mentoring, another finding that is in accordance with the theory.

In every site, a high or increasing percentage of teachers said that PD had helped them improve their instruction and promote student learning. But teachers with lower effectiveness ratings were generally no more likely than higher-rated teachers to say that the support they received had helped them.

Overall, the sites made progress in implementing evaluation-linked PD, and teachers with lower ratings were more likely than higher-rated teachers to report needs for PD, as expected, and were more likely to receive PD. But because the sites did not fully implement systems to record information on the PD recommended for teachers or track their actual participation, we could not assess whether participation led to improved effectiveness.

CHAPTER SEVEN

Compensation Policies

> [The bonus] is really great; who's going to complain? But we're not in the job because it's a lucrative field. We do it because we love what we do. It's nice to have that fiscal validation, but it's not the deal breaker for motivating whether I'm going to stay as a teacher or not.
>
> —Green Dot teacher, 2014,
> interviewed about compensation policies

In Brief

Did the sites modify their compensation policies to reward effective teaching, and did these policies lead to a more-equitable distribution of effective teaching?

The districts implemented bonus programs that awarded modest bonuses to effective teachers in certain schools on a year-to-year basis, but only PPS adopted an effectiveness-based salary schedule (however, it applied only to teachers hired after 2010). The CMOs all adopted bonus programs, which were discontinued after two to three years; in Aspire and Alliance, they were replaced by effectiveness-based salary schedules for all teachers.

Survey responses suggest that most teachers were aware of their site's compensation policies and endorsed the idea of additional compensation for teachers demonstrating outstanding teaching. More than half of the teachers in PPS and the CMOs thought that the programs were reasonable, fair, and appropriate. However, although most teachers supported the

idea of effectiveness-based bonuses, they reported that such incentives did not motivate them to improve their teaching (although teachers in Aspire and Alliance were exceptions). In addition, some teachers objected to compensation decisions being made on the basis of an effectiveness measure they perceived to be flawed.

More-effective teachers earned more in total salary than less effective teachers, but we do not have sufficient evidence to say whether this was due to the revised compensation policies or to other factors. In addition, effective teachers in high-LIM classrooms in HCPS received more compensation than effective teachers in other classrooms, but we could not assess this in the other sites.

Introduction

This chapter reviews the IP sites' efforts since the start of the initiative in 2009–2010 to enact effectiveness-based compensation policies, educators' reactions to these changes, and the new policies' effects on teacher compensation. Data come from annual surveys and interviews of teachers and SLs and from site compensation data (described in Chapter One). In Chapters Eleven and Twelve, we examine the combined effects that compensation and other policies had on the retention of effective teachers and on LIM students' access to effective teaching.

Effectiveness-based compensation policies link teachers' measured effectiveness with their salaries, either through a salary schedule with effectiveness as a built-in dimension or through supplemental bonuses contingent on teachers' effectiveness ratings. Effectiveness-based compensation policies bear some similarity to CL policies, which we discuss in Chapter Eight. Both involve additional pay, but CL positions reward teachers for taking on additional responsibilities and leadership, while effectiveness-based compensation rewards effective teachers who remain in their current roles.

To a smaller extent, this chapter also examines the use of compensation policy to encourage and reward effective teachers for teaching in classrooms with higher percentages of LIM students. Several of the policies discussed in this chapter offer additional compensation to

effective teachers in high-need schools or to groups of teachers most likely to be teaching LIM students.

The Gates Foundation did not offer specific guidelines for the structure of effectiveness-based compensation policies, but its application process alerted the sites to specifically call out the outcomes or behaviors that their proposed compensation structures were designed to incentivize. The RFP also mentioned demonstrated effectiveness in teaching low-income and minority students as one potential trigger for salary increases and bonuses (Bill & Melinda Gates Foundation, 2009). All the sites included performance-based compensation in their proposals, and each implemented one or more forms of performance pay.

Best Practices for Teacher Compensation

Traditionally, U.S. teachers' salaries are based on two factors: educational preparation and teaching experience. This approach makes sense if teaching quality improves with education and experience, but the evidence about these relationships is mixed. Although math and science teaching improves with subject-specific education (Monk, 1994), this does not seem to be the case in other subjects. With this exception, there is little evidence that obtaining more education (e.g., higher degrees or additional credits for continuing education) leads to improved teaching (Harris and Sass, 2011).

On the other hand, there is clear evidence that teaching quality improves with experience, but findings are mixed about how long this improvement continues. One review of evidence found that improvements to teaching skills seem strong only in the first few years on the job and that improvement tends to plateau after about five years in the classroom (Rice, 2010). Another review suggests that, although gains in effectiveness are greatest in a teacher's early years, performance continues to improve into the second and third decades of experience (Kini and Podolsky, 2016). There is also some evidence that more-experienced teachers demonstrate lower quality; for example, second-

ary math teachers with 25 years of experience were found to be less effective than their less experienced peers (Ladd, 2008).

As a result of such findings, many people have argued that the traditional teacher salary schedule does not serve education well. Instead, researchers have called for stronger connections between salary and measured effectiveness—that is, paying effective teachers more and ineffective teachers less (Hannaway, 2006; Hanushek, 2016). They argue that this could accomplish multiple goals, including increasing the retention of effective teachers, reducing the retention of ineffective teachers, and improving outcomes for students.

Recent evidence suggests that tying compensation to effectiveness does not always produce these effects but can achieve many of them under the right circumstances. For example, researchers found that effectiveness-based financial incentives awarded as part of the Washington, D.C., IMPACT program led to improved performance among effective teachers (Dee and Wyckoff, 2013). Similarly, an evaluation of ten districts participating in the TIF program found that pay-for-performance bonuses had small positive effects on student performance after three years but no improvement in measured TE (Wellington et al., 2016). On the other hand, Springer and his colleagues found that large bonuses for the teachers with the highest VAM scores in Metro Nashville Public Schools did not improve student outcomes overall (Springer et al., 2012). Similarly, Fryer found that incentives for New York City teachers who met a performance target did not raise school performance (Fryer, 2013). A meta-analysis of merit pay experiments in the United States and abroad that included most of the studies mentioned here concluded that merit pay programs overall had a small positive effect on student achievement but noted that the effects were sensitive to the design and implementation of the program. For example, group incentives were more effective than individual incentives (Pham, Nguyen, and Springer, 2017), although evidence was lacking to identify other critical features of these programs.

These various studies demonstrate that education research has not yet identified the optimum approach to teacher compensation, but they (and others) do offer some useful insights into three features of effectiveness-based compensation: how effectiveness should be mea-

sured, what aspects of effectiveness should influence salary, and how challenging and substantial effectiveness-based awards should be.

How Effectiveness Should Be Measured

In Chapter Three, we summarized the literature on measuring TE. Although it is far from definitive, the research suggests that the best approach to measuring effectiveness—in terms of both reliability and validity for predicting student achievement growth—is one that combines multiple types of information, including direct observation of teaching practice and evidence of a teacher's impact on student achievement growth (Kane, Taylor, et al., 2010; Kane and Staiger, 2012). It is important that the measure distinguish among teachers so that outstanding teachers can be clearly identified. A recent study of the performance pay policy that Florida implemented in 2011 noted that the pay policy was hampered by the fact that 98 percent of teachers were rated as E or HE by the state's teacher-evaluation system, providing no basis on which to identify the most-effective teachers (Walsh, Lubell, and Ross, 2017).

What Aspects of Effectiveness Should Influence Salary

In contrast, compensation experts suggest that different measures should be used to determine different aspects of compensation, rather than a combination of measures (as recommended for TE). Specifically, scholars (Odden and Kelley, 2002; Odden and Wallace, 2007) recommend that base salary—the largest portion of compensation—should be centered on how well a person performs his or her job's tasks relative to established best practices. In education, this might be measured through supervisory or peer evaluations. The researchers also recommend basing additional compensation on the production of desired outcomes, assessed at the appropriate level. In education, outcomes are typically measured in terms of the attainment and achievement of students and assessed at the classroom, school, or district level. These recommendations would suggest a base salary linked with observations of teacher performance and bonuses based on student achievement results.

How Challenging and Substantial the Effectiveness-Based Awards Should Be

Pay-for-performance experiments in education have examined bonus programs that varied in the size of bonuses and the performance level required to earn them and found mixed results (e.g., Springer et al., 2012; Dee and Wyckoff, 2013). The TIF proposal guidelines capture the best current wisdom: Bonuses should be differentiated (i.e., varying dollar amounts based on effectiveness), challenging to earn (i.e., a threshold that not all teachers could easily achieve), and substantial (i.e., at least 5 percent of salary). However, these criteria might be harder to achieve in practice than in theory. For example, although these features were part of the TIF program guidelines, the evaluators found that they were not fully implemented in most TIF programs (Wellington et al., 2016). Although the bonuses were almost always differentiated (with an average amount of $1,850 and a maximum of $7,700), they were neither as challenging nor as substantial as called for. Similarly, a recent study criticized Florida's performance pay program because the districts awarded small amounts for teaching excellence. The authors of that study reported that it would take a teacher an average of four years being consistently rated as HE to earn the same salary as he or she would earn teaching for one year with a master's degree (Walsh, Lubell, and Ross, 2017).

IP Compensation Policies

As part of the IP initiative, each site was to redesign its compensation system to emphasize demonstrated effectiveness in salary determinations. Foundation staff and site leaders hoped that these incentives would reward effective teachers, increase the likelihood that they would continue teaching, and motivate all teachers to improve. In these ways, the compensation reforms were to contribute to the overall goals of improving LIM students' access to effective teachers and improving student outcomes overall.

Before the IP initiative, all the sites had traditional step-and-column compensation systems, in which a teacher's base salary was

based on years of experience and education credits.[1] HCPS also had a performance-based bonus policy in place before the initiative, and legacy MCS (which merged with legacy SCS during the initiative) had experimented with a bonus system in a few schools from 2004 through 2009.

Under the IP initiative, the sites modified their compensation systems in a variety of ways so that teachers could earn additional pay as a function of their effectiveness. As described in Chapter Three, TE was measured by a single composite measure that combined at least two kinds of information—structured classroom observations to measure the quality of a teacher's practice and growth in student achievement.[2] Most of the compensation systems used the composite measure to trigger increases in pay.[3] Thus, the compensation policies that the sites adopted did not differentiate rewards based on measures of practice from rewards based on measures of outcomes, as some scholars have advocated (Odden and Kelley, 2002; Odden and Wallace, 2007).

Some sites proposed *effectiveness-based bonuses* in the form of annual or biennial supplemental awards that would be determined by a teacher's effectiveness rating for the preceding period.[4] In contrast, other sites planned to implement *effectiveness-based salary schedules*, in which a teacher's effectiveness rating and experience would be the only determinants of base salary (although, like in a traditional step-and-column schedule, once a teacher's salary reached a particular level, it would not revert in the future). Thus, the compensation policies that the sites adopted also did not conform to the TIF guidelines of being

---

[1] *Base salary* refers to the amount a teacher is paid for performing his or her normal job, excluding extra duties or supplemental responsibilities.

[2] Some sites also included information gathered from students via surveys and other kinds of evidence, although these sources carried much less weight in the composite measure than observations and student achievement growth.

[3] Two of the SCS bonus systems were based on student outcomes only.

[4] Not all sites use the term *bonus* to describe this supplement to salary, but it is the most common. HCPS, in particular, eschews the term because the payment takes effect the year after the year on which the award is based, so it is available only to teachers who remain in the district. The term *bonus* might convey that high-performing teachers should receive the payment whether or not they remained HCPS teachers.

differentiated, challenging, and substantial. Although all were differentiated, they varied considerably in the extent to which they set high thresholds for performance and awarded adequate amounts of money to individual teachers.

Some sites also encouraged greater access by LIM students to effective teaching by making the performance-based component of pay greater for teachers most likely to teach LIM students. By increasing the rewards for effectiveness among these teachers, sites tried to encourage effective teachers to relocate from low-LIM schools to high-LIM schools and to create larger rewards for improvement by teachers in high-LIM schools.

Tables 7.1 through 7.4 summarize the effectiveness-based salary policies enacted by the IP sites, and Appendix H provides additional details. In completing the tables, we found that it could be quite difficult to designate the school year in which a particular compensation policy went into effect because the teaching that earned a financial reward usually occurred in the school year before the reward was actually paid. In fact, some sites calculated the student achievement growth component of the effectiveness measure based on two or more years of teaching, making it harder to be consistent in labeling the policy enactment date. For clarity and consistency, in this chapter, we indicate that a new compensation policy started in the school year during which the relevant formal teacher observation occurred.[5]

---

[5] Consider the simplest case, in which a district computed its composite measure of TE annually. For example, let's say classroom observations occurred during 2013–2014, and the achievement measure included student test results from the spring of 2014 and the prior year. The observation scores (and the student survey scores, if they were part of the composite measure) would be available during 2013–2014, but the test would not be scored and the results returned to the district until the summer or early fall of 2014. VAM scores would be computed at some point during 2014–2015, as would the composite TE score. If the teacher earned a bonus as a result of these 2013–2014 performance measures as part of a new compensation policy, we would say that *the policy began in 2013–2014*, and the bonus would be awarded at some point during the 2014–2015 school year. The situation can be even more complicated: For example, in PPS, to smooth annual variation in VAM estimates, the VAM score included in a teacher's TE rating for 2013–2014 would actually be based on three years of data (2010–2011, 2011–2012, and 2012–2013).

Similarly, it is difficult to predict when changes to compensation policies are likely to have an effect on teacher behavior and student outcomes. Compensation policy changes might influence teachers in the year they are announced, the year that observations and other data collection begin, the year in which the additional payments are actually paid, or even in years following bonus payments to peers. Further, this chain of actions and reactions might differ for individual teachers. This uncertainty makes it more difficult to associate changes in policies with changes in teacher behaviors and student outcomes, and we try to acknowledge this limitation when we discuss the results of our analyses.

Although, in general, the IP initiative represented a major change in how sites managed their teacher workforces, some sites had already enacted certain elements of the initiative before it began but in different ways. To understand the extent of the IP initiative changes and make inferences about their impact, it is important to be aware of the following preinitiative compensation programs:

- HCPS had MAP, a district-wide bonus program going back to 2006–2007 that was based on Florida standardized test scores, district-created tests for subjects not tested statewide, and rudimentary tables of value added. This program was modified (as indicated in Table 7.1) to become the IP bonus program.
- Legacy MCS (which merged with legacy SCS during the initiative) had an experimental pay-for-performance program called the Effective Practice Incentive Fund, which applied to principals and teachers in 17 high-poverty schools from 2004 through 2009. A teacher could earn up to $7,500 based on school and classroom achievement, plus additional awards if he or she agreed to share best practices with an online network of educators. The program ended before the IP initiative began.
- Prior to 2011–2012, Alliance awarded bonuses to all teachers in schools that met school-level targets on the California Academic Performance Index.
- HCPS, SCS, and Green Dot offered a bonus, salary supplement, or higher placement on the salary schedule for any teacher who

earned National Board certification. In some cases, the amounts were quite substantial (e.g., $4,000 to $10,000, depending on the site). Giving bonuses for certification can certainly be considered "effectiveness-based" compensation, but we do not include it in our tables because it is not based on the sites' own effectiveness ratings.

District Compensation Policies

This section describes the compensation policies that the three districts adopted as part of the IP initiative. The next section describes the compensation policies that the four CMOs followed. Tables 7.1 and 7.2 summarize the districts' policies.

As part of the IP initiative, the three districts adopted new compensation policies that included effectiveness-based bonuses. One district also adopted an effectiveness-based salary schedule for teachers hired after 2010. All three districts adopted bonus schemes that rewarded teachers for performing above a certain threshold. In most cases, the threshold was a particular score on a site's effectiveness measure over a designated period (either one or two years); in some cases, a student outcome measure independent of the TE measure was used.

In addition, a relatively small number of teachers in PPS received salary increases as part of the district's effectiveness-based salary schedule that applied only to teachers hired after July 2010. Forty-three of the 44 eligible teachers (among the district's roughly 1,700 teachers) received performance-based salary increases of $8,000 or more in 2015–2016.[6] If this policy increases effectiveness among new teachers, it could improve LIM students' access to effective teaching because new teachers are more likely to be in high-LIM schools.

The compensation policies varied across the districts with respect to the timing, award criteria, eligible pool, size of the bonuses, and the proportion of teachers earning a bonus. To determine whether the changes in compensation policies had any effect, we

[6] Eligibility for the effectiveness-based salary increments is explained further in Appendix H.

Table 7.1
District Effectiveness-Based Bonus Policies Through 2015–2016

Name of Bonus Program	When Active?	Eligibility	Criterion for Award (among eligible teachers)	Amounts (per teacher)	Penetration in the Most Recent Year in Which the Bonus Was Awarded (and for which data were available)
HCPS					
Performance-based salary adjustment	**2006–2007 through 2009–2010:** Based on MAP; **2010–2011 through 2012–2013:** Based on TE; **2013–2014 and ongoing:** Based on revised TE	Any teacher in at least the fourth year of teaching	**2009–2010:** Top quartile on MAP; **2010–2011 through 2012–2013:** Top quartile on TE; **2013–2014:** TE rating of 4 or 5	**2013–2014:** $2,000 for level 4 and $3,000 for level 5; **2014–2015:** $1,900 for level 4 and $2,900 for level 5; **2015–2016:** $1,400 for level 4 and $2,100 for level 5	50% of teachers in HCPS in 2014–2015; 55% of teachers in HCPS in 2015–2016
TIF grant POWER1	**2007–2008 through 2011–2012**	116 schools	Top quartile on TE, revised to TE rating of 4 or 5 in 2013–2014	5% of base salary	No information available
TIF grant POWER2	**2010–2011 through 2014–2015**	All teachers in 35 designated schools	Top quartile on TE, revised to TE rating of 4 or 5 in 2013–2014	5% of base salary	No information available

Table 7.1—Continued

Name of Bonus Program	When Active?	Eligibility	Criterion for Award (among eligible teachers)	Amounts (per teacher)	Penetration in the Most Recent Year in Which the Bonus Was Awarded (and for which data were available)
TIF grant POWER3	**2012–2013 through 2016–2017**[a]	All teachers in 30 designated schools	Teachers with TE ratings of 4 or 5	**2015–2016:** $3,800	45% of all teachers at POWER3 schools; 12% of all teachers district-wide
PPS					
PRC Cohort Award	**2010–2011 and ongoing**	PRC schools only (six schools in 2010–2011; five schools in 2011–2012; three schools from 2012–2013 through 2014–2015; six schools starting in 2015–2016) Experienced ninth- and tenth-grade teachers in the PRC cohort or who teach at least 60% ninth- or tenth-grade students in PRC schools	VAM scores of 51 or higher and TE rating of satisfactory	$1,000–$8,800 for PRC cohort teachers in 2015–2016 based on VAM and TE scores; $64–$1,500 for non-PRC teachers, based on students taught and VAM and TE scores, prorated for absences	About 80% of PRC cohort teachers and 26% of non-PRC teachers in 2015–2016 (134 teachers in five schools in 2015–2016; ~8% of teachers in PPS)

Table 7.1—Continued

Name of Bonus Program	When Active?	Eligibility	Criterion for Award (among eligible teachers)	Amounts (per teacher)	Penetration in the Most Recent Year in Which the Bonus Was Awarded (and for which data were available)
STAR award (school level)	**2011–2012 and ongoing**	All schools	Up to eight schools; must be in top 15% or 25% of the state in student achievement growth.[b] In 2015–2016, four of 50 traditional schools and three of four special schools were eligible for STAR awards. Teachers in these schools who have satisfactory performance	$1,000–$6,000 based on the number of days worked in the school	All teachers in STAR schools (~9% of teachers in PPS)
AYP award	**2010–2011 through 2014–2015**	PFT members hired before July 2010 who were on the top step of their salary schedule	Satisfactory rating in years when the district made AYP. PPS made AYP and awarded bonuses only once—in 2010–2011—during the years in which this bonus was active.	$1,000	About half of PPS teachers

Table 7.1—Continued

Name of Bonus Program	When Active?	Eligibility	Criterion for Award (among eligible teachers)	Amounts (per teacher)	Penetration in the Most Recent Year in Which the Bonus Was Awarded (and for which data were available)
SCS					
Effectiveness bonus	**2011–2012 only**	All teachers	Any teacher with a score of 5 on each TEM component; any teacher with a TEM score in the top 10%; or any teacher with a composite TEM score of 5	$500–$2,000 based on TEM score	About 25% of teachers in SCS
Bonus for TVAAS gains (school level)	**2011–2012 only**	High-priority schools	Teachers with gains in TVAAS scores in all subjects	$3,000	No information available
Bonus for achievement on state tests (school level)	**2013–2014 only**	All schools	Schools that met or exceeded achievement goals on state tests in 2013–2014; all teachers in these schools	$2,000–$3,000	14 of about 200 schools (~7% of teachers in SCS)

Table 7.1—Continued

Name of Bonus Program	When Active?	Eligibility	Criterion for Award (among eligible teachers)	Amounts (per teacher)	Penetration in the Most Recent Year in Which the Bonus Was Awarded (and for which data were available)
Reward status bonus	**2012 and ongoing**	iZone schools that were in the top 5% of growth or proficiency in the state	All teachers	$3,000	No information available

SOURCES: Annual interviews with central-office staff in each of the IP districts.

NOTE: STAR = Students and Teachers Achieving Results. AYP = adequate yearly progress, a key metric established by the No Child Left Behind Act (NCLB) (Pub. L. No. 107-110, 2002).

[a] The POWER1 grant was about $15 million; the POWER2 grant was about $10 million; the POWER3 grant was about $60 million and offered larger bonuses.

[b] If no PPS schools were in the top 15 percent of schools in the state on the student achievement growth metric, the eligibility criteria were expanded to include PPS schools in the top 25 percent of schools in the state while still limiting the total to no more than eight schools.

Table 7.2
District Effectiveness-Based Salary Schedule Policies Through 2015–2016

Characteristic	HCPS	PPS	SCS
Name	None[a]	Performance-based salary increase	None
When active		2010–2011 and ongoing	
Eligibility		Teachers hired after July 2010	
Criterion for salary		Teachers rated D at least once in the three years since the previous level decision	
Amounts (per teacher)		$8,070–$11,310 based on years of service in 2015–2016	
Penetration in the most recent year for which data were available		98% of eligible teachers (43 out of 44) in 2015–2016 (about 2.4% of teachers in PPS)	

SOURCES: Annual interviews with central-office staff in each of the IP districts.

[a] HCPS did adopt a new salary schedule during the initiative; the new schedule eliminated degree-based supplements, and a teacher needed to maintain a rating of E (3) to be eligible for raises. The district also retained its previous schedule for a relatively small number of teachers who did not opt into the new one.

need to know when they went into practice, how large they were, and how many teachers earned them. The answers were different in each district:

- In HCPS, all experienced teachers were eligible for moderate bonuses based on effectiveness ratings beginning in 2010–2011 as part of the IP initiative. These bonuses replaced the existing MAP bonus program. MAP bonuses were similar in magnitude to the IP bonuses and were based on state and local test scores. In addition, HCPS had three bonus programs based on TIF that began prior to the IP funding and continued during the initia-

tive. Teachers in about 116 designated POWER1 schools could earn moderate-size, effectiveness-based bonuses. POWER2 and POWER3 offered similar bonuses but targeted a much smaller number of schools—35 and 30, respectively. These TIF programs focused on high-risk schools, thereby increasing the performance incentives for teachers of LIM students.

- PPS implemented three bonus programs and an effectiveness-based salary schedule in 2010–2011 and 2011–2012 as part of the IP initiative. The changes affected a limited number of teachers, and the bonuses received ranged from $64 to $11,000 (although $20,000 was the maximum for the district's PRC bonus).
- SCS adopted four bonus programs as part of the IP initiative. Three of them were short-lived and awarded bonuses from $500 to $3,000. The only remaining performance-based bonus program in SCS is limited to iZone schools, which are schools in the bottom 5 percent of the state achievement distribution. These schools tend to have large LIM populations, so this policy has the potential of increasing LIM students' access to effective teaching.

CMO Compensation Policies

The CMOs adopted revised compensation policies that included bonus payments for exceeding effectiveness thresholds; the bonuses were discontinued in all four CMOs. Two of the CMOs replaced the bonuses with effectiveness-based salary schedules. Tables 7.3 and 7.4 show the effectiveness-based bonus and salary policies that the CMOs adopted as part of the IP initiative. All four CMOs adopted bonus policies that applied to all but lower-performing teachers in 2012–2013 and 2013–2014 (and 2014–2015 in Alliance). The bonus policies were then discontinued in Green Dot and PUC because of teacher and administrator concerns about the fairness of the measure used to determine the bonus. In 2014–2015, both Alliance and Aspire adopted effectiveness-based salary schedules; experienced teachers at the highest level of performance (master) could earn $13,000 to $15,000 more than teachers rated at the middle level of performance (E).

Table 7.3
CMO Effectiveness-Based Bonus Policies Through 2015–2016

Characteristic	Alliance	Aspire	Green Dot	PUC
When active?	**2012–2013 through 2014–2015**	**2012–2013 and 2013–2014**	**2012–2013 and 2013–2014**	**2012–2013 only**
Eligibility	All teachers	All teachers	All teachers	All teachers
Criterion for award	Top four of five TE levels	Top four of five TE levels	Top three of five TE levels	Top three of five TE levels
Amounts (per teacher)	$750–$5,500 based on TE level	$500–$3,000 based on TE level	$500–$2,000 based on TE level	$1,500–$5,000 based on TE level for 2013–2014
Penetration in the most recent year in which the bonus was awarded	99% of teachers for 2014–2015	99% of teachers for 2013–2014	89% of teachers for 2013–2014	98% of teachers for 2012–2013

SOURCES: Annual interviews with central-office staff in each of the IP CMOs.

Table 7.4
CMO Effectiveness-Based Salary Schedule Policies Through 2015–2016

Characteristic	Alliance	Aspire	Green Dot	PUC
When active?	**2014–2015 and 2015–2016**	**2014–2015 and ongoing**	Not applicable	Not applicable
Eligibility	All teachers	All teachers		
Criteria for salary	Years of service and TE rating level for two years	Years of service and TE rating level		
Amounts (per teacher)	$15,001 additional for a master beyond what an E teacher with ten years of experience would receive	$13,050 additional for a master beyond what an E teacher with ten years of experience would receive		
Penetration in the most recent year for which data were available	No information available	28% of teachers reached the master level in 2015–2016		

SOURCES: Annual interviews with central-office staff in each of the IP CMOs.

The CMO compensation systems were similar to the district compensation systems with respect to award criteria, timing, size, and the proportion of teachers earning bonuses but did not specifically target teachers of LIM students. The CMOs took two general compensation approaches. They all adopted similar bonus strategies initially and discontinued them after one to three years. Two of the CMOs then adopted effectiveness-based salary schedules. The specifics of the compensation policies are as follows:

- In Alliance, bonuses were first awarded based on data collected in 2012–2013, and almost every teacher received $750 to $5,500. Alliance discontinued bonuses after 2014–2015, the year in which it adopted an effectiveness-based salary schedule.
- In Aspire, bonus policies began in 2012–2013, and almost every teacher received $500 to $3,000. Aspire discontinued bonuses after 2013–2014. It adopted an effectiveness-based salary schedule in 2014–2015, and many teachers reached the higher levels of effectiveness.
- In Green Dot, bonus policies began in 2012–2013, and almost every teacher received $500 to $2,000. Green Dot discontinued bonuses after 2013–2014.
- In PUC, bonus policies were implemented in 2012–2013, and almost every teacher received $1,500 to $5,000. PUC discontinued effectiveness-based bonuses after one year.[7]

Implementation Challenges

Changes in compensation policy evoked concerns in the IP sites. Our interviews with central-office administrators, principals, and teachers highlighted three issues that sites had to address when designing their new compensation policies.

[7] Every PUC teacher received a bonus for 2013–2014 for participating in the research and development of the evaluation system.

Teachers wanted assurances that base salaries would not be reduced under the new compensation system. This "hold-harmless" principle was part of the public discourse about the initiative in many sites. For example, when Alliance and Aspire adopted effectiveness-based salary schedules, they had to include a provision known as "no backsies" that guaranteed that no teacher would be moved backward on the schedule even if his or her effectiveness rating declined in the future. People accepted the argument that teachers needed to know that their salaries would be stable or increasing so they could meet their financial commitments, although this meant that teachers at the highest level might not feel as much pressure to maintain or improve the quality of their teaching. In PPS, to win teacher approval for a new effectiveness-based salary schedule, the district had to agree that it would apply only to new teachers hired after a certain date. Existing teachers remained on the traditional salary schedule.[8]

Some sites faced uncertainty about their overall levels of state funding, which made it difficult to change policies that affected salary. Volatile state funding made it difficult to commit to bonuses or salary schedules that increased the total compensation for the teaching staff. Sites could not adopt new compensation systems that cost more in total than the existing system if they were unsure that future funding would support it. The best example of this is the economic downturn in California that led to decreases in state funding for public education. During the lean years, the CMOs suspended all raises for three years. They also became more cautious about making salary commitments that could lead to higher costs in the future. This was one reason that Green Dot and PUC did not join Alliance and Aspire in adopting effectiveness-based salary schedules, along with concerns about the validity of the effectiveness ratings.

Funding uncertainty also led to many of the sites having an initial preference for bonuses. The CMOs liked bonuses because their

---

[8] The collective bargaining agreement that included the new salary schedule also included a performance-based salary schedule for existing teachers, called *voluntary incentives for earnings at work*. The agreement specified that teachers would have to opt into this incentive program and that it would be implemented more slowly during the following two or three years. However, it was never implemented.

sizes could be adjusted annually based on the overall financial situation. HCPS was concerned about the difficulty of estimating how many teachers would earn bonuses in any given year and about the possibility that the number might exceed the district's ability to pay. To make the program sustainable, beginning in the 2016–2017 school year (using the 2015–2016 effectiveness scores), HCPS divided the total salary adjustment pool across all eligible teachers, rather than giving a set amount to each teacher like it had previously.

In sites with teacher unions or professional organizations, these groups were very concerned about salary issues and focused lobbying efforts on policies that affected salary. In HCPS, the union made retaining the salary adjustments a priority, and the adjustments were incorporated into the 2016 negotiated teacher contract. In SCS, in which the professional organization did not represent all SCS teachers and therefore did not have a strong voice on most matters, it raised its concerns with the district leadership, and those concerns were addressed in the final compensation policy. In PPS, the district worked closely with PFT leadership to craft the performance-based salary schedule and bonus programs.

Teacher and SL Reactions to Compensation Policies

Revised compensation policies were designed to boost student outcomes by encouraging teachers to become more effective and motivating effective teachers to continue teaching. This section describes the reactions of teachers, and, in some cases, SLs, to the changes and their perceptions of the policies' impact on teacher practice.

To examine teachers' perceptions of the new compensation policies, we explore their input, from surveys and interviews, on four broad topics:

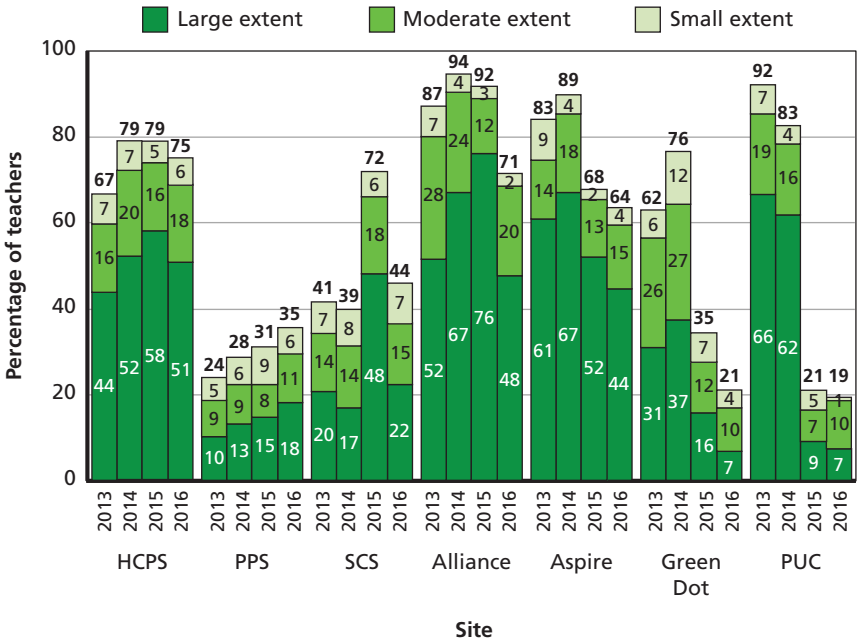
- awareness: Did teachers know about the compensation policies?
- endorsement: Did teachers approve of the compensation policies?
- fairness: Did teachers think that the compensation policies were fair?

- perceived effects: What types of effects did the compensation policies have, according to reports from teachers and SLs?

Awareness

Most teachers' responses to questions about compensation were consistent with their site's policies, suggesting awareness of the policies. In all sites except HCPS (with respect to salary) and SCS (with respect to bonuses and salary), most teachers seemed to be familiar with their site's compensation policies. Figures 7.1 and 7.2 show the percentage of teachers in each site reporting on the extent to which evaluation results would play a role in determining one-time bonuses or

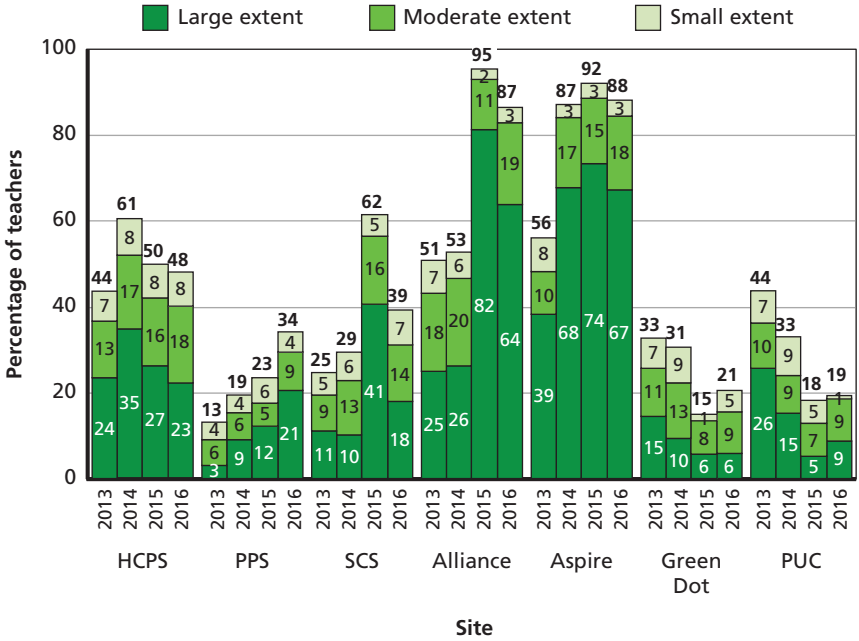
Figure 7.1
Teachers Reporting That Their Evaluation Results Would Be Used to a Large, Moderate, or Small Extent to Determine Whether They Would Receive a Monetary Bonus on Top of Their Salary, Springs 2013–2016



NOTE: Omitted response categories are "not at all" and "don't know." We did not ask this question in 2011.

RAND *RR2242-7.1*

Figure 7.2
Teachers Reporting That Their Evaluation Results Would Be Used to a Large, Moderate, or Small Extent to Determine How Much of a Salary Increase They Would Receive for the Subsequent Year, Springs 2013–2016



NOTE: Omitted response categories are "not at all" and "don't know." We did not ask this question in 2011.

RAND *RR2242-7.2*

permanent salary increases, from the spring of 2013 through the spring of 2016. Their responses provide a partial indicator of how well teachers understood the overall compensation system in place each year.

For example, 70 to 80 percent of HCPS teachers reported that their evaluation results would play a role in whether they received bonuses (Figure 7.1), and, in fact, all teachers with four or more years of experience were eligible for bonuses at that time. On the other hand, about half of HCPS teachers reported that their evaluation results would affect their future salaries (Figure 7.2), despite HCPS not having a salary schedule in which different TE ratings were linked to different increase amounts. One possible reason for this relatively high percent-

age linking evaluation results to salary might be that the temporary performance-based increases were called *salary adjustments* rather than *bonuses*. Another reason might be that, at about the same time that HCPS instituted new criteria for the temporary performance-based adjustments, the district also adopted a new opt-in salary schedule. Although the new salary schedule was not primarily effectiveness-based, some teachers might have conflated it with the performance-based salary adjustments. For example, one HCPS teacher interviewed in 2016 said, "Since I was more of a senior employee, I didn't have to opt into the new system, and I never did. So I'm still under the old pay system. So I wasn't going to get bonuses or anything like that, above and beyond what I would normally get." Although this teacher, who was hired before the 2013–2014 school year, correctly understood the option to opt in to the new salary schedule, opting in was not required to be eligible for bonus pay.

Responses from PPS teachers were also consistent with site policies. Few reported that their evaluation results would influence their future salaries or bonuses, which is consistent with the fact that only a small share of teachers were in schools (PRC cohort schools or STAR-eligible schools) that offered substantial effectiveness-based bonuses. Similarly, most PPS teachers were hired prior to 2010 and thus were not on the effectiveness-based salary schedule.

Interpreting the SCS results is more difficult because of the array of short-lived bonus policies (except for the reward status bonus, which was in place for longer but only in iZone schools). About 40 percent of teachers reported that evaluation results would play some role in their bonuses in 2013, 2014, and 2016 despite the fact that an effectiveness bonus was in place only in select schools. The large jump in the percentage of SCS teachers indicating that effectiveness would play a role in compensation—both bonuses and salary—in 2015 is perplexing because we are not aware of any new bonuses awarded in that year, and SCS did not have an effectiveness-based salary schedule during this period. A new salary schedule including pay for performance was discussed publicly in 2015 but was not implemented; teachers might have been responding to the prospective policy change. Interviews in SCS suggested that the rules governing the different programs were not

well communicated. Many teachers were unaware that such programs existed, which suggests that these programs were not well understood throughout SCS.

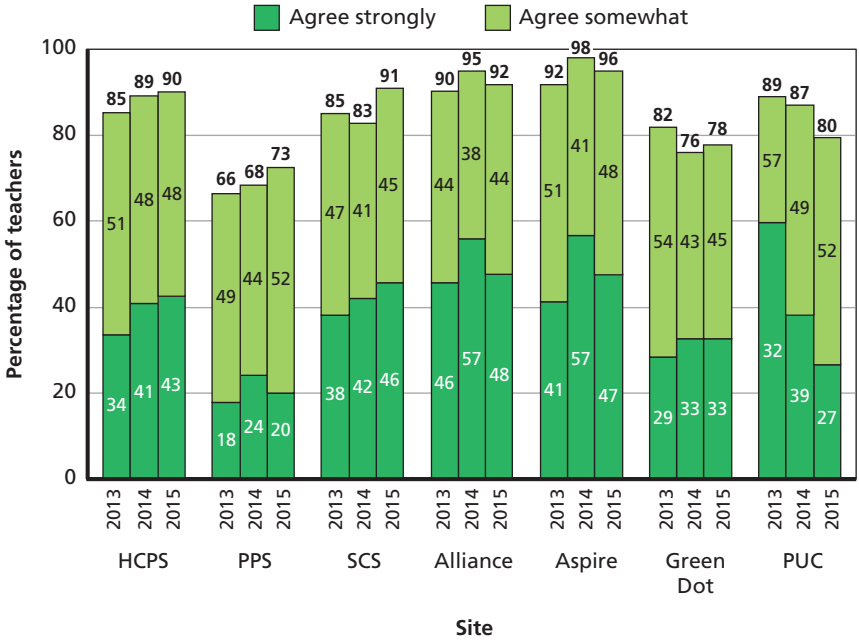
Responses from teachers in the CMOs indicated that they were generally aware of their site's policies related to effectiveness-based compensation. In Alliance, Aspire, and PUC, more than 80 percent of teachers reported in the springs of 2013 and 2014 that their evaluation results would be used to determine whether they received bonuses. Those percentages dropped somewhat in Alliance and Aspire in 2015 and 2016, when those two sites changed to effectiveness-based salary schedules. The percentages of teachers agreeing also dropped in PUC, which discontinued bonuses after the 2012–2013 school year. Only about 60 percent of teachers in Green Dot reported that evaluation results would play a role in determining bonuses in 2013 and 2014, and the Green Dot percentages declined sharply in the two years after the bonuses were removed. In Alliance and Aspire, the percentages of teachers agreeing that evaluation results would be used in salary determinations rose when those policies were put in place, as one would hope.

Endorsement

In all the sites, most teachers endorsed the idea of additional compensation for effective teaching. Teachers generally endorsed the idea of additional compensation based on outstanding teaching skills. Figure 7.3 shows that there was broad but not universal support for this idea, with high percentages in all the sites agreeing that "teachers should receive additional compensation for demonstrating outstanding teaching skills." In PPS in both 2014 and 2015, nearly all novice teachers agreed, compared with only about 70 percent of more-experienced teachers.[9] Interestingly, however, there were no large differences—particularly ones consistent across years and across sites—in agreement between teachers of different effectiveness ratings (low versus middle versus high).

[9] This difference was statistically significant at $p < 0.001$ in both years.

Figure 7.3
Teachers' Agreement That Teachers Should Receive Additional Compensation for Demonstrating Outstanding Teaching Skills, Springs 2013–2015



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly." We did not ask this question in 2016.

RAND *RR2242-7.3*

Although teachers generally supported the idea of additional compensation for effective teaching, large majorities in the three districts (especially PPS) thought that base pay should be based on seniority, suggesting a belief that effectiveness-based compensation should supplement but not supplant the traditional system.[10] Our interviews with PPS teachers confirmed this finding; over the six years of the study, most PPS teachers we interviewed endorsed the idea that base

[10] Not surprisingly, in every site and every year (2013 through 2016), experienced teachers were more likely than novice teachers to agree that base pay should be based on seniority. This difference was significant ($p < 0.05$ or lower) in all three districts and at least two CMOs each year.

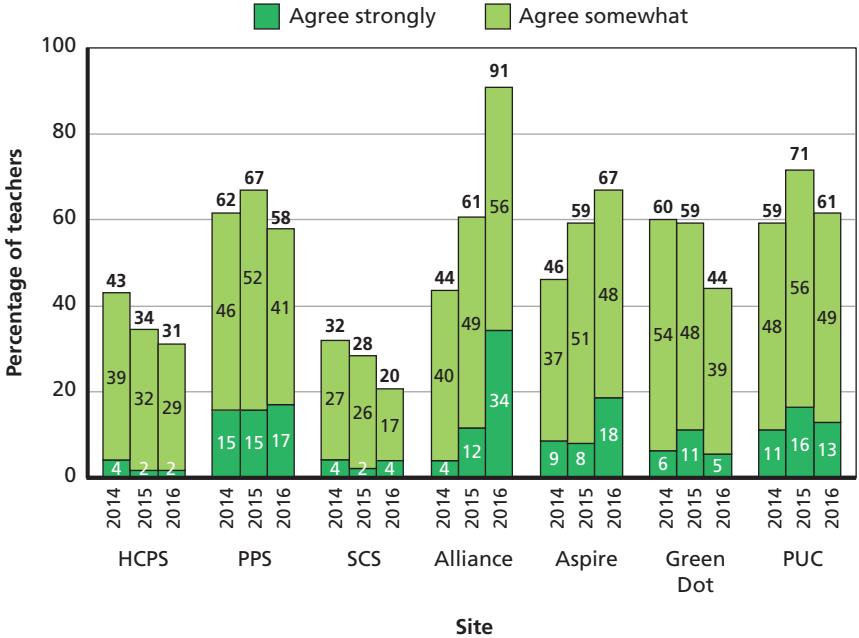
salary should be based on seniority, but they allowed for the possibility of monetary rewards for effectiveness, provided that the methods and measures for awarding such rewards were fair and transparent. As one teacher said, "There should be a bonus maybe, not a pay scale. Things happen—sometimes, somebody has something happen with their family and they can't be distinguished that year—but bonus for the extra effort would be nice."

Similarly, most teachers in the CMOs supported the idea of additional compensation for effective teaching. In interviews, teachers generally commented that any opportunity to earn extra money is good, although financial rewards are not their main reason for teaching. As an Aspire teacher noted in 2014, "Making our salary competitive, which they're trying to do, is where it should start. Having stipends for other roles is great as well. I think there should be both. The more effective you are as a teacher, I think you should be paid more."

Fairness

In most years, a majority of teachers in PPS and the CMOs agreed that their site's compensation system was "reasonable, fair, and appropriate," but agreement was lower in HCPS and SCS. There was considerable variation among the sites in teachers' views on the fairness of the compensation system. Figure 7.4 shows that majorities of teachers in PPS and in the CMOs tended to agree somewhat or strongly that their site's teacher compensation system (e.g., salary structure, opportunities for bonuses) was "reasonable, fair, and appropriate." In HCPS and SCS, however, the percentages agreeing were lower and declined each year. In HCPS, the decline in positive responses about the compensation system's fairness could reflect a general decline in teachers' attitudes toward the evaluation system as a whole (see Chapter Three). It could also be because teachers did not receive step increases in salary for three years because of budget constraints, which might have led to a general feeling that compensation was not reasonable or appropriate.

Figure 7.4
Teachers' Agreement That Their Site's Teacher Compensation System Was "Reasonable, Fair, and Appropriate," Springs 2014–2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly." We did not ask this question in 2013.

RAND *RR2242-7.4*

In interviews, HCPS teachers indicated mixed sentiment toward the compensation system. In 2014, one teacher expressed suspicion about the system:

> It seems that everyone got lower evaluation scores this year because of the pay scale. Seems kind of coincidental that, all of a sudden, everyone got lower scores at the same time that we heard [that,] if you don't get high scores, you don't move up in the pay scale. It just seems odd to me and other teachers.

On the other hand, another HCPS teacher stated, "Great, we're getting the money that's due to us."

In PPS, it appears that the relatively high percentages of agreement indicated by the survey results were driven mainly by experienced teachers who were not subject to the new performance-based salary schedule. In 2016, for example, although 59 percent of PPS teachers with at least three years of experience agreed that the system was fair, only 35 percent of teachers with two years or less of teaching experience (who were all subject to the new schedule) agreed.[11] Indeed, interviews with PPS teachers who were hired after 2010 and were subject to the salary schedule suggested that many did not believe that the performance measures on which it was based were fair. They thought that the observations of their practice were subjective and the student growth measures were unreliable and based on inappropriate or inaccurate tests. Many teachers we interviewed were stressed and anxious about the possibility that raises would not be distributed fairly. As one teacher put it in 2016,

> I can speak for a number of people hired after 2010 under the new salary scale who feel [that] it is extremely subjective. . . . Our principal doesn't give out many distinguished ratings. That's really stressful. I think about when I get to that point or whether I can relax into my career financially or worry constantly about my finances and slogging along to get there.

Some teachers we interviewed were worried that they would not receive raises at all, reflecting an apparent misunderstanding; central-office staff told us that every teacher subject to the new salary schedule was to receive an annual base raise with a supplement based on his or her effectiveness rating.

Interviews in PUC also revealed concerns about linking effectiveness to compensation because of questions about the quality of the effectiveness ratings. Central-office staff told us in 2014, "We cannot give bonuses in good conscience based on a system [teachers] say to us is flawed." The newly developed career path positions (described in Chapter Eight) for which teachers had to qualify were seen as an alternative way to reward TE.

---

[11] This difference was statistically significant at $p < 0.01$.

Not surprisingly, teachers who agreed that "the amount of compensation I receive as a teacher allows me to live reasonably well" were far more likely than teachers who disagreed to think that the compensation system was reasonable, fair, and appropriate. In 2016, for example, in every site, more than half the teachers who thought that their own compensation was adequate agreed that the system was fair. But among teachers who disagreed that their compensation was adequate, percentages saying the system was fair were much lower (see Figure 7.5).[12]

Perceived Effects

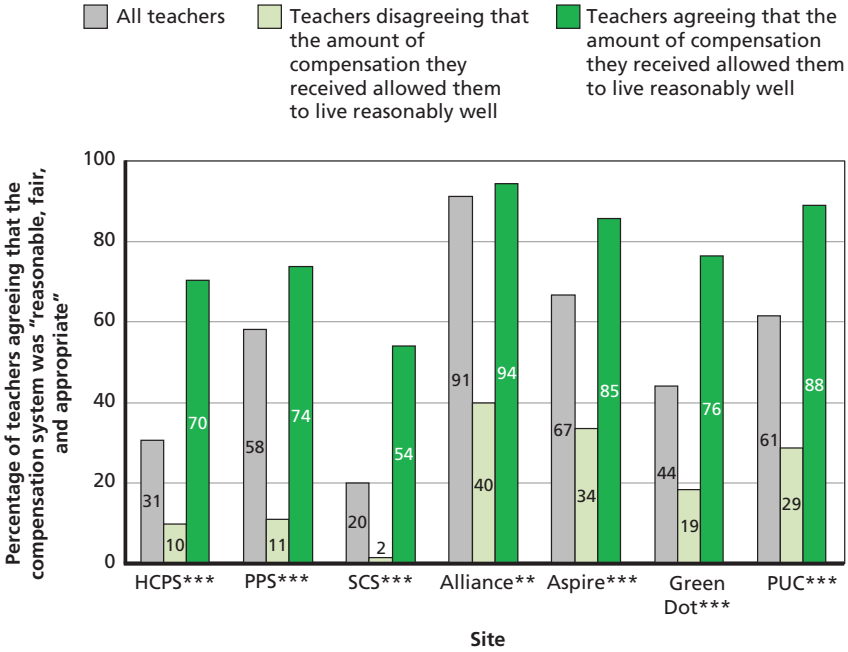
In every site except Alliance and Aspire, less than half the teachers reported that their site's compensation system motivated them to improve their teaching. Compared with the percentages of teachers who thought that outstanding teaching should be rewarded with additional compensation (see Figure 7.3), far fewer teachers reported that their site's compensation system motivated them to improve their teaching (see Figure 7.6).[13] The exceptions were Alliance and Aspire, in which majorities of teachers in 2015 and 2016 reported that their site's compensation system motivated them to improve their teaching (although that was also true in Aspire in 2014). These two sites modified their salary schedules for all teachers to directly reward effectiveness, and the additional salary that master teachers could earn was substantial. In 2015, more than half of teachers in Alliance and more than 40 percent in Aspire were in the two highest effectiveness levels earning larger salaries. PUC discontinued bonuses in 2013–2014,

---

[12] It is worth noting that teachers in PPS were more likely than teachers in the other sites to agree that the amount of compensation they received allowed them to live reasonably well. In the years we asked this question (2014, 2015, and 2016), at least three-quarters of PPS teachers agreed, much more than the 40 to 65 percent in the CMOs and 35 to 50 percent in the other two districts. This could be another reason the percentage of teachers who agreed that the compensation system was fair was higher in PPS than in the other sites.

[13] In HCPS, teachers rated as HE were more likely than low-rated teachers to say that the compensation system motivated them, but the opposite was true in SCS. These differences could be related to features of the bonus system and the distribution of teachers across schools in each site (e.g., bonuses were available only to teachers in certain schools).

Figure 7.5
Teachers' Agreement That Their Site's Teacher Compensation System Was "Reasonable, Fair, and Appropriate," by Perception of Salary Adequacy, Spring 2016



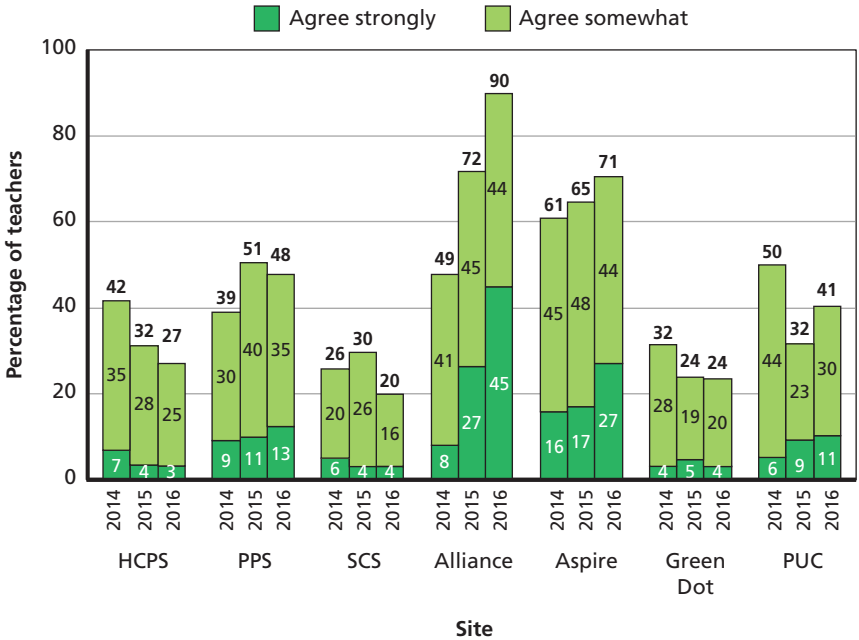
NOTE: We have combined the original response categories of "agree strongly" and "agree somewhat." Omitted response categories (for the vertical-axis variable) are "disagree somewhat" and "disagree strongly." ** denotes that the difference between agreeing and disagreeing percentages (legend variable) is significant at $p < 0.01$. *** denotes that the difference between agreeing and disagreeing percentages (legend variable) is significant at $p < 0.001$.

RAND *RR2242-7.5*

and the proportion of PUC teachers agreeing that the compensation system motivated them to improve declined after the spring of 2014. We also asked SLs whether the compensation system motivated teachers at their school to improve their teaching, and most did not strongly believe that it did.

Interviews reinforced the survey findings about motivation. In PPS, most teachers we interviewed said that performance-based compensation did not motivate them to improve their teaching; rather,

Figure 7.6
Teachers' Agreement That Their Site's Compensation System Motivated Them to Improve Their Teaching, Springs 2014–2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly." We did not ask this question in 2013.

RAND *RR2242-7.6*

most cited their love for teaching or a desire to help children learn. As one PPS teacher said, "It would be hard to say [that the current compensation system] has a significant impact on teacher performance. A good teacher wants to be good regardless; they're not motivated to perform based on salary increase." Teachers in other sites expressed similar sentiments.

Researchers have noted that performance-based incentives can lead to undesirable efforts to boost scores without improving underlying performance, and such efforts to "game" the system can even undermine the validity of the effectiveness measure (Campbell, 1976). Some teachers we interviewed commented on pressures of this type. For example, in PPS, eligible teachers reported that, to improve their

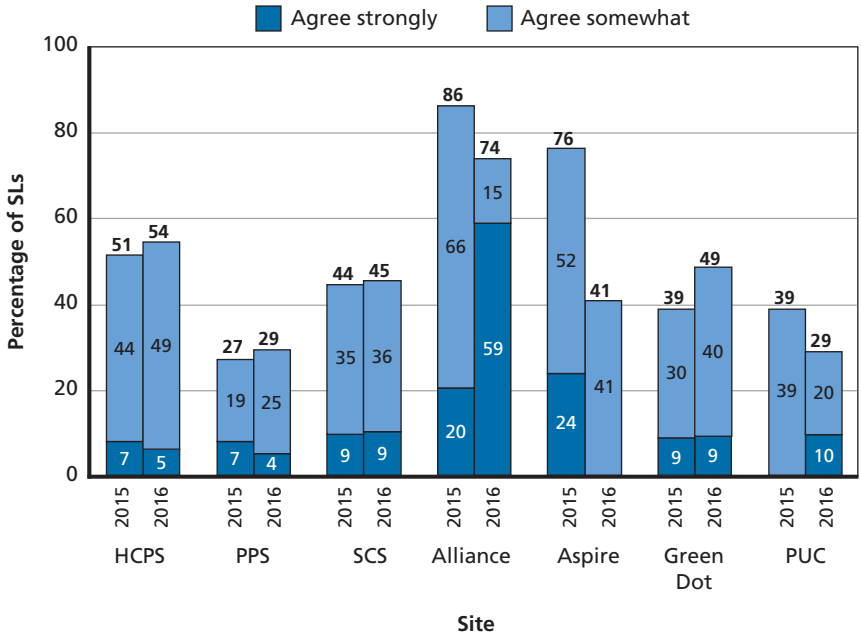
observation scores and receive salary increases, they were highly motivated to change their teaching for lessons that were going to be observed but that they were not so motivated to make lasting changes in their practice. Teachers in the CMOs made similar comments. For example, several Alliance teachers described being motivated to improve their observation scores but did not equate that pressure with improving their teaching. As one remarked in 2016,

> The language is about "if you want a four, here's what you do." So there's a disconnect between the purpose of why we're doing this, which should be student achievement, and then the compensation, which—you can't blame people for wanting to get a good score if it's compensation-based.

Survey responses from SLs, particularly in the districts, indicated that effectiveness-based compensation might also have had some negative effects on teacher collaboration and morale. In both 2015 and 2016, 40 to 50 percent of district SLs agreed that "rewarding teachers based on individual performance hurts teacher collaboration at my school." However, in Alliance, Green Dot, and PUC, less than 25 percent of the SLs agreed in either year. In Aspire, agreement grew from 22 percent in 2015 to 39 percent in 2016. The patterns were similar for SL agreement with the statement, "Since it was implemented, the compensation system has decreased teacher morale at my school." Agreement was highest in the districts (around 50 percent, except in PPS in 2016, when only 25 percent agreed) and in Aspire in 2016 (50 percent).

The majority of SLs in HCPS, Alliance, and Aspire thought that the compensation system improved their ability to retain effective teachers, but less than half agreed in the other sites. We asked SLs in 2015 and 2016 whether they agreed that the "compensation system is helping my school retain good teachers." As shown in Figure 7.7, just more than half of the SLs in HCPS agreed each year, and more than 70 percent of SLs in Alliance and Aspire agreed in 2015. (In Aspire, 76 percent of leaders agreed in 2015, but only 41 percent agreed the following year.) In PPS, SCS, Green Dot, and PUC, less than half of SLs agreed in either 2015 or 2016. (In PPS, the percentages were less than 30 percent.) The responses from Alliance and Aspire likely reflect

Figure 7.7
SLs' Agreement That Their Site's Compensation System Was Helping Their School Retain Good Teachers, Springs 2015 and 2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly."
RAND *RR1295-7.7*

these two sites' implementation of effectiveness-based salary schedules in which effective teachers could increase their salaries substantially. This change was also accompanied by general salary increases for all teachers in both CMOs in 2014–2015 and a second salary increase for all teachers in Aspire in 2015–2016. However, in the spring of 2016, the level of agreement fell somewhat in Alliance and dramatically in Aspire.[14] An Aspire staff member suggested that this might be because the effectiveness measures had not been adapted to reflect the site's emphasis on the Common Core curriculum.

[14] The decline in Aspire was statistically significant ($p < 0.001$), but the decline in Alliance was not ($p > 0.05$).

Did the IP Compensation Policies Increase Pay for Effective Teachers?

In addition to examining the sites' compensation policies and staff perceptions of these policies, we were interested in the extent to which these policies led to greater compensation for effective teachers. For two of the sites, HCPS and PPS, we were able to examine teacher salary data to see whether more-effective teachers or effective teachers with higher proportions of LIM students were, on average, receiving greater compensation. Ideally, we would have restricted our analyses to teachers who were eligible for a particular compensation regime and then examined whether teachers' actual compensation was consistent with their effectiveness ratings and the reward structure. Unfortunately, we could not conduct that analysis because of (1) the small numbers of teachers strictly eligible for bonuses or effectiveness-based salary schedules (see Tables 7.1 and 7.2 for information about eligibility and penetration) in the two districts and (2) our inability to identify which teachers opted into which alternative compensation schedules. As a result, we conducted a descriptive analysis on all teachers that provides indirect information about whether the compensation policies were rewarding the teachers they were designed to support. In Chapter Eleven, we examine teacher retention and its relationship to the set of relevant policies implemented as part of the IP initiative, including effectiveness-based compensation and tenure and dismissal policies.

The Relationship Between TE and Total Compensation

We were able to compare the total compensation that teachers with different effectiveness ratings (HE versus E versus U or NI) received in HCPS and PPS; we did not have adequate compensation data from the other sites. We can compare teachers based on two different measures of effectiveness: the site's composite TE measure and the study-calculated VAM measure, which is available only for mathematics and ELA teachers in grades 4 through 8. The site's TE measure has the advantage of being available for all teachers. The VAM measure has the advantage of being stable across years, whereas sites' own composite measures of TE differ from one another and are subject to potential

inflation over time. Using VAM also facilitates comparisons across districts. Appendix A describes the methodology we employ for estimating VAM scores.

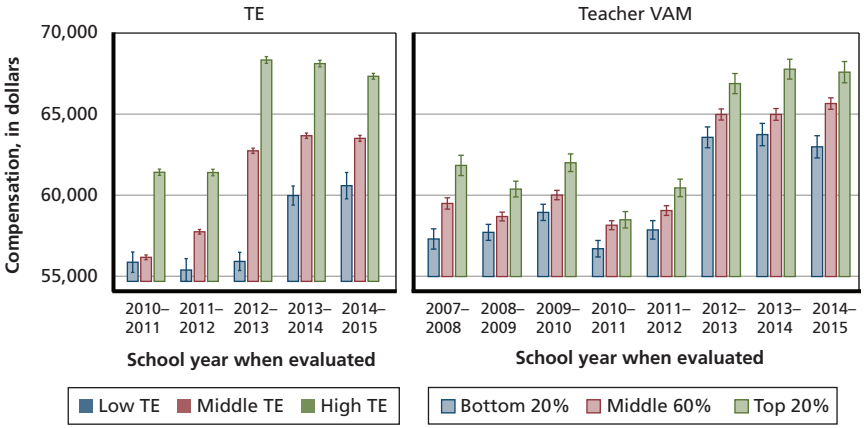
For the purpose of comparison, we combined teachers into three groups based on the site's TE measure and three groups based on the study-calculated VAM. In both HCPS and PPS, we collapsed the site's TE categories into three, which we labeled low, middle, and high. Thus, the five-category HCPS TE measure is condensed to three categories by grouping together the bottom two and the top two categories; in PPS, we grouped only the bottom two categories together to condense the site's four-category TE measure into three categories. In interpreting our TE analyses, it is also important to remember that the number of teachers is not evenly distributed across TE levels; typically, the lowest level has the fewest teachers. For the study-calculated VAM metric, we combined teachers into three groups representing the bottom 20 percent, middle 60 percent, and top 20 percent. In addition, we can look at trends in the comparison over time to see whether differences change in ways that are consistent with changes in compensation policies. We present the results for HCPS first and then the results for PPS.

The estimates we present in subsequent figures result from modeling teacher compensation as a function of TE (measured in terms of the site's TE score or the study-calculated VAM score), controlling for the teacher's age, teaching experience, educational attainment, gender, and race. Appendix I provides more details about the methodology. Because we conceptualize teacher compensation as responsive to effectiveness, the estimates reported in this section show the effect of TE and VAM measured in one year on compensation in the next year.[15] We ran the models separately for each site.

More-effective teachers in HCPS received significantly more total compensation than less effective teachers, and the difference did not change over time. Figure 7.8 shows the results for HCPS from 2010–2011 through 2014–2015. In the left-hand side of the figure,

[15] The figures show the years when TE and VAM were measured; compensation is from the next year: We modeled 2013–2014 teacher compensation as a function of 2012–2013 TE or VAM.

Figure 7.8
HCPS Total Compensation, by District Composite TE Level and Study VAM Score



NOTE: For the TE category, bottom = U and NI, middle = E, and top = HE 4 and 5. Error bars show 95-percent confidence intervals; estimates control for teacher characteristics.

RAND *RR2242-7.8*

showing the results by TE, high-TE HCPS teachers received more total compensation, on average, than teachers with lower ratings in each year, and the difference is statistically significant at $p < 0.01$.

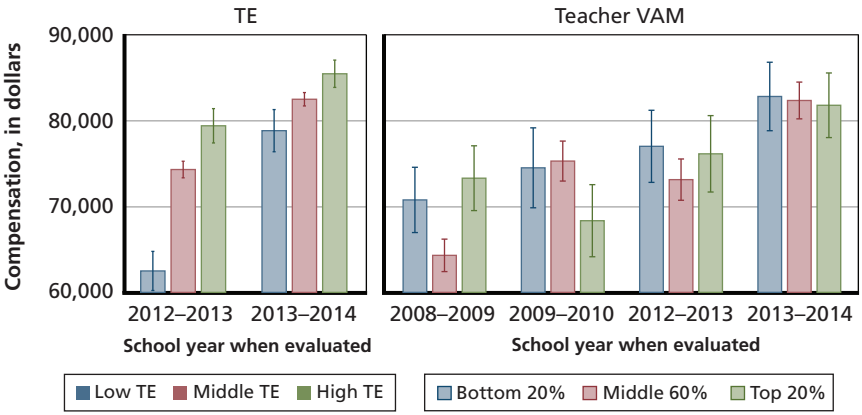
We observed similar patterns for HCPS by VAM (right-hand side of Figure 7.8). In all years for which we had data, teachers with top 20-percent VAM scores were more likely than middle and bottom 20-percent VAM teachers to receive higher salaries, with the differences between their salaries averaging about $4,000.

Although these data suggest that compensation policies were rewarding more-effective teachers more than less effective teachers, it is not clear what role the IP effectiveness-based bonus policies played in this relationship. HCPS had an existing bonus program that predated the IP initiative (see Table 7.1) and influenced teacher compensation. However, because these estimates account for differences in teacher characteristics, we know that this relationship is not due to greater pay for teachers with more experience or education. (Note that HCPS salaries increased for all teachers beginning in 2012–2013.) Nevertheless,

these findings demonstrate the desired relationship between TE and compensation.

PPS teachers with higher composite TE scores received greater total compensation than teachers with lower scores. In PPS, we received compensation data only for selected years, so our analysis by TE is limited to 2012–2013 and 2013–2014, and our analysis by VAM is limited to 2008–2009, 2009–2010, 2012–2013, and 2013–2014. High-TE PPS teachers received more total compensation than teachers with lower ratings in 2012–2013 and 2013–2014, differences that are statistically significant at $p < 0.01$ (left-hand side of Figure 7.9). Additionally, compared with what they received in 2012–2013, teachers received significantly more compensation following a high-TE rating in 2013–2014. Similarly, high-VAM teachers (those in the top 20 percent) in 2013–2014 received significantly more compensation than high-VAM teachers in 2012–2013. However, comparing across VAM levels, the only statistically significant difference in total compensation between low-VAM (those in the bottom 20 percent) and high-VAM teachers occurred in 2009–2010, when low-VAM teachers

Figure 7.9
PPS Total Compensation, by District Composite TE Level and Study VAM Score



NOTE: For the TE category, bottom = F and NI, middle = P, and top = D. Error bars show 95-percent confidence intervals; estimates control for teacher characteristics.

RAND *RR2242-7.9*

received greater compensation than high-VAM teachers. In other years, high-VAM teachers received compensation that was not significantly different from that of low-VAM teachers.[16]

On the one hand, the differences between compensation by TE are substantial, but we do not have sufficient data to see whether the difference between low-TE and high-TE compensation changed in response to specific policies. Teachers received significantly more compensation following a high TE rating in 2013–2014 than following a similar score in 2012–2013. One possible explanation is that eligible teachers on the new salary schedule were awarded performance-based salary increases in addition to step increases. Another possible explanation is that the STAR bonus payout more than doubled in 2014–2015. However, each of these programs affected relatively few teachers, so it is difficult to fully explain the differences in compensation across years.

The same is true for high-VAM teachers: Teachers received significantly more compensation following a high-VAM score in 2013–2014 than following a high-VAM score in 2012–2013. Although we have more years of VAM score data to analyze, the evidence does not support a claim that high-VAM teachers received greater compensation than low-VAM teachers.

Compensation for Effective Teachers of LIM Students

Next, we examine the associations between teachers' compensation and both TE and the percentage of their students who were LIM. Some of the sites had policies that rewarded teachers for being effective. Some had policies that rewarded teachers for teaching LIM students. The combination of these rewards is the minimum we might expect an effective teacher of LIM students to receive. In this section, we examine whether sites provided additional compensation over and above these separate rewards in order to provide extra incentive to increase the effectiveness of the teachers of LIM students.

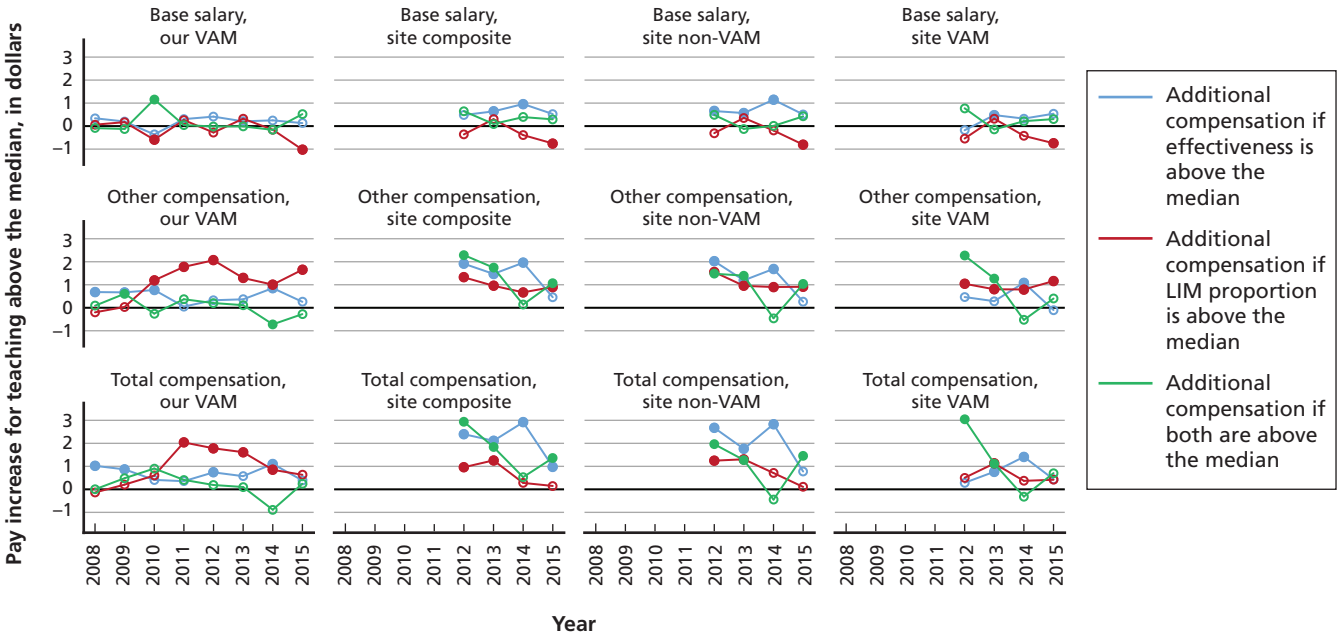
Because only HCPS provided the study team with sufficient compensation data, this analysis of incentives for effective teachers to teach

[16] For reasons explained earlier, we could not restrict this analysis to just teachers who were eligible for various performance-linked compensation programs.

in classrooms with high proportions of LIM students is limited to that district. For the data we did have available in PPS and SCS, we found no statistical evidence of a relationship between LIM-classroom assignments and increased pay. We examine the relationships with respect to four separate measures of TE: study-calculated VAM (our own estimates of TE), site composite TE rating (the composite, or overall, effectiveness scores of the teachers), site VAM (each site had its own VAM estimates of teacher productivity that were one part of the TE composite), and site non-VAM (site composite TE score minus site VAM score). Chapter Twelve provides more details about these measures, and more details about the analytic methods are found in Appendix I.

In HCPS, effective teachers who worked with LIM students were paid more than comparable teachers who did not, and, in certain years, the more-effective teachers who taught LIM students were paid even more. Figure 7.10 shows that HCPS rewarded effective teachers and teachers who worked with LIM students. The pay increase arises not from base pay but from other compensation, such as bonuses. There did not seem to be a relationship before 2010, whereas, for every year since, there is strong evidence that teachers who taught more than the median fraction of LIM students received, on average, between $1,000 and $2,000 additional pay in other compensation. Across the different measures, we generally find increased compensation for being above the median in effectiveness, above the median in LIM teachers' compensation, and above the median in both. We note that the effects are strongest when using the site metrics of effectiveness and weakest when using our own VAM measure. This is not surprising because the sites had access to the TE measures and not to our measure. The effects were the strongest in the 2012 and 2013 years in general, although there is evidence that an effect persisted in later years. This implies that HCPS was trying to encourage effective teachers to work with LIM students.

Figure 7.10
HCPS Change in Compensation



NOTE: A solid circle indicates that the estimated amount is statistically significant at $p < 0.05$. A hollow circle indicates that the estimate is not statistically significant.

RAND *RR2242-7.10*

Summary

All seven IP sites implemented effectiveness-based compensation reforms, although they varied in terms of timing, eligibility for bonuses, the size of bonuses, and the proportion of teachers earning bonuses—factors that are likely to affect the impact of the compensation policies. Each site used a composite measure of TE that included both direct measures of practice and measures based on the growth in student achievement. The use of multiple measures is consistent with research recommendations regarding teacher evaluation, but it is not consistent with recommendations regarding compensation; some researchers argue that direct measures of performance should be associated with base pay, while achieving desired outcomes should inform bonuses.

Our analysis found that teachers were generally aware of their site's compensation policies and endorsed the idea of bonuses for teachers demonstrating outstanding teaching. However, most teachers outside of Aspire and Alliance did not think that their site's compensation system motivated them to improve their teaching. In most years, a majority of teachers in PPS and the CMOs, but only minorities in HCPS and SCS, thought that their site's compensation policies were reasonable, fair, and appropriate. Although more-effective teachers in HCPS tended to receive greater total salaries, the salary gap did not increase after the implementation of the effectiveness-based awards, suggesting that the new compensation policies were not making large changes to how much money effective teachers earned compared with what less effective teachers earned.

More broadly, the initiative itself presented some challenges to interpreting the results. For one, its many moving pieces made it hard to separate the effects of one lever from the effects of the others. In other words, the IP initiative does not provide a very rigorous test of the specific effects that effectiveness-based pay had on teachers because the policy was enacted as part of a package of HR reforms that also included new effectiveness measures, changes in hiring and placement and dismissal policies, and new PD efforts. Although we can isolate the effects of compensation policies in a few instances (e.g., total salary in

HCPS), in general, we cannot disentangle the interactions among these multiple efforts.

Another challenge is that the initiative sometimes did not follow best practices. Our analysis reveals that the features of the enacted compensation policies were not entirely consistent with the features identified in the literature on performance-based compensation. As noted earlier, the TIF guidelines suggest that effective performance-based pay should be differentiated, challenging, and substantial. For the most part, the IP compensation reforms did not meet all three standards. In HCPS and SCS, the bonus awards spanned relatively small ranges ($1,400 to $3,800 and $500 to $3,000, respectively), not differentiating much among teachers. The range of awards was larger in PPS (although very few teachers were eligible) and the CMOs. The compensation policies also differed with respect to what proportion of teachers were eligible to receive additional compensation and how easy it was to earn—in most years, more than half of teachers in HCPS, about one-quarter in SCS, and less than one-tenth in PPS earned bonuses. In comparison, the vast majority of teachers earned bonuses in the CMOs during the years that bonuses were available; thus, the criteria were not very challenging. Although there were many more bonuses in the CMOs, the bulk of these awards were not very substantial relative to salary. Top bonuses were larger in PPS than in the other two districts. Effectiveness-based salary schedules in PPS, Alliance, and Aspire offered more-substantial salary increases to those teachers reaching the highest level of effectiveness.

When one considers all the aspects of the effectiveness-based compensation policies enacted by the sites, it might not be surprising that there were mixed reactions among teachers. None of the policies took a two-tiered approach that, as recommended by some researchers, used practice-based measures differently from outcome-based measures in determining compensation. Few if any of the sites enacted policies that were differentiated, challenging, and substantial.

CHAPTER EIGHT

CL Policies

> We made a very deliberate decision to change the way we thought about the "career ladder" because it implied that there was only one direction of movement, which ended out of the classroom (administrator or curriculum specialist). Instead, we chose to think about leadership positions as ways for teachers to broaden their experience and impact by taking on leadership positions . . . . Each position still has entrance requirements, but we try not to say one is more important than the other. They all serve in developing our teachers further.
>
> —Green Dot central-office administrator, 2015

In Brief

To what extent did sites implement CL policies offering teachers additional responsibilities and leadership opportunities, and how did staff react to these policies?

Our analysis found that the sites used CL positions as both a strategy to increase retention and a means to give teachers opportunities for professional advancement. The districts and CMOs took somewhat different approaches to CLs. The districts created a few positions that focused on coaching and mentoring new teachers in struggling schools, while the CMOs created several positions, with a wider range of specific responsibilities as needs shifted over time. Most teachers knew whether CL options were available in any given year, and, for the most part, teachers' judgments about CL policies and personnel were positive. For example, teachers

thought that the CL teachers with whom they worked were effective educators who had helped them improve their instruction and deserved the additional compensation they received. Most teachers in most sites also reported that the opportunity to become a CL teacher motivated them to improve their instruction and increased the chances they would remain in teaching.

Introduction

This chapter reviews the efforts of the IP sites to enact CL policies and teachers' reactions to these changes. Like effectiveness-based compensation, CLs are designed to improve the retention of the most-effective teachers by giving them more responsibility and more pay; they are also designed to take advantage of the expertise of effective teachers to help improve the performance of other teachers.

We use the term *CLs* loosely to refer to a wide range of specialized teaching positions with greater responsibility that are open to teachers who have demonstrated skills or expertise. Specifically, CLs offer new roles for effective teachers that provide leadership opportunities and additional salary in exchange for taking on additional responsibilities, including using their expertise to help improve other teachers' effectiveness.

In using the term this way, we are ignoring a distinction that researchers often make between two types of specialized teacher positions: differentiated roles and CLs (Johnson, Berg, and Donaldson, 2005). Differentiated roles give teachers new responsibilities outside the classroom and additional compensation, but neither the duties nor the remuneration is permanent. When the assignment is over, the teacher relinquishes the extra responsibilities and extra pay. Examples of differentiated roles include serving as department chair in an HS or being a member of a curriculum development team. In contrast, a CL establishes an ordered set of job positions with increasing responsibility and leadership, assigned based on teaching merit, with permanent advanced status and salary. An example of a CL is the sequence of positions at Brooke Charter Schools in Boston: associate teacher (first two years), classroom teacher (more than two years of experience), and

master teacher (rated exceptionally E for three years in a row and given additional responsibilities mentoring associate teachers, planning PD, and similar tasks) (McCann and Zuflacht, 2015).

Many of the IP sites originally planned to create CLs but instead established differentiated roles. Because the term *CL* was used in the initiative and because we incorporated it into our surveys and interviews, we continue to use it. But, in this report, *CL* refers to a broad array of specialized teaching positions and teacher leadership roles that are not necessarily hierarchical or permanent.

This chapter begins with a review of the research literature on CLs and what features make them more effective. Then we describe the CL policies put in place as part of the IP initiative. The next section discusses the implementation of those policies and the challenges the districts and CMOs faced. We then turn to educator attitudes toward sites' CL policies, before briefly discussing their effects on teachers' intentions to remain in teaching. We conclude with a summary of our findings related to CL policies.

Best Practices with Respect to CLs

CLs serve multiple purposes, but one of the main motivations for creating CLs is to increase the retention of effective teachers (McCann and Zuflacht, 2015). Recent research has found that 40 percent of teachers quit teaching within their first five years, and the attrition rate is higher at high-poverty, high-minority, urban, and rural schools (Ingersoll, 2014). Perhaps more discouraging is that half of those who leave are among the most-effective teachers (TNTP, 2012). TNTP refers to these teachers as *irreplaceables* and warns that "the real teacher retention crisis is not simply the failure to retain enough teachers; it is the failure to retain the right teachers" (p. 4). According to the TNTP analysis, half the teachers in the top 20 percent of effectiveness ratings leave within five years. Although teachers' salaries increase as they gain experience and education, in recent years, these traditional incentives have not been enough to motivate some of the most-effective teachers to remain in the classroom: "Because teaching has few possibilities for

career advancement, highly motivated teachers seeking more responsibility and a better salary may move into administration or leave the profession altogether" (Holley, 2008, p. 93). CLs might be an effective way to combat the potential loss of the most-effective and most-energetic teachers:

> In order to ensure that the most ambitious and effective teachers stay for 10 years and beyond, school systems should create career ladders that allow teachers to gradually earn leadership opportunities that meet their aspirations, provide them the time to serve, and compensate them appropriately. (McCann and Zuflacht, 2015, p. 14)

A second reason to enact CLs is to draw on the best teachers' expertise to raise the performance of other teachers and improve outcomes for students. For example, the American Federation of Teachers encourages districts to create CLs:

> By working with teachers to develop and implement a career ladder program, a district demonstrates its commitment to enable all staff to achieve their full potential, which in turn increases levels of personal satisfaction and improves job performance. But most importantly, a career ladder program builds the internal capacity of the school district to positively affect student achievement by using its most underutilized resource: its people. (American Federation of Teachers, 2013, p. 1)

PPS touts the benefits of its CL options in similar terms:

> Career Ladder positions are promotional opportunities for effective teachers to take on leadership responsibilities without leaving the classroom. Career Ladder teachers have an increased impact on . . . students because they elevate not only their own practice, but that of their peers as well. (PPS, 2012, p. 1)

These are important outcomes, and there is some evidence that CLs work to reduce teacher attrition, boost student achievement, and promote other positive teacher outcomes. For example, TNTP (2012)

reports that, in one district, "the percentage of teachers who would choose to work in a low-performing school doubled when the school offered teacher leader roles" (p. 30). Some smaller studies have found evidence of improved student achievement associated with CLs. For example, a reanalysis of data from a Tennessee class size experiment in which students were randomly assigned to teachers found that students whose teachers participated in the state's CL program had higher average achievement scores than students whose teachers did not (Dee and Keys, 2005). Student achievement gains were also reported in studies in Arizona (Dowling, Murphy, and Wang, 2007) and Missouri (Booker and Glazerman, 2008). Finally, a separate study of 12 Missouri schools that implemented CLs and 12 matching schools that did not found that the CL program had a positive impact on teachers' self-reported morale, commitment, and satisfaction (Ebmeier and Hart, 1992). Those authors found this impact among midcareer and new teachers in particular.

Advocates for CLs offer recommendations about the best way to design and enact such policies. For example, TNTP identified four sets of strategies districts could use to retain the most-effective and most-energetic teachers: (1) meaningful feedback and development, (2) recognition, (3) responsibility and advancement, and (4) resources (TNTP, 2012, p. 16). Effective strategies should recognize teachers' accomplishments publicly, identify opportunities for teacher leader roles, put them in charge of something important, and offer them additional resources they can use in their own classroom. Informed by surveys of teachers, TNTP reported that "[t]op teachers who experience two or more of these retention strategies plan to keep teaching at their schools for nearly twice as long (2–6 more years)" (TNTP, 2012, p. 16). Other scholars argue for flexibility in designing CL positions. Rather than a single hierarchy of jobs with increasing responsibility, argues Doyle (2015), an "ideal career structure is a lattice in which great teachers can pursue different paths that best fit their strengths and career interests." Doyle suggested that, in addition to leading peers, this might include "roles such as reaching more students using technology, serving as a multi-classroom teacher, and specializing in a particular subject or grade combination" (p. 4).

However, there is limited empirical evidence about the features of effective CL policies because such policies are not widely implemented and have not been subject to much formal research. Nevertheless, a recent summary of this research highlighted key features of effective CL policies. Teachers must perceive the roles to be "legitimate, accessible, and 'doable,' if they are to increase teachers' satisfaction or retention" (Johnson, Berg, and Donaldson, 2005, p. 96). Moreover, the process by which the roles are implemented and supported affects their effectiveness. Although early studies found some positive effects on teacher satisfaction, later research suggested that how the policies were enacted explained a lot of their effects. In particular, having "clear selection processes, committed leadership, and ongoing communication throughout the planning and implementation of a career ladder" led to better responses from teachers (Johnson, Berg, and Donaldson, 2005, p. 96).

In the next section, we draw from our interviews and document reviews to describe the CL policies that the IP sites enacted. Then, we report teachers' and SLs' survey and interview responses to questions about CLs. Finally, we briefly discuss the impact that these policies had on student outcomes.

IP CL Policies and How the Sites Defined Them

As part of the IP initiative, each site was encouraged to create a CL in which effective teachers could take on greater responsibility while still maintaining a focus on teaching in exchange for greater compensation. At the start of the initiative, some sites already had a few specialized positions that allowed teachers to earn extra pay for taking on additional responsibilities, but the number of such positions was relatively small, and teachers' roles were limited.

Although planned in their proposals to the Gates Foundation, none of the sites implemented a CL that embodied a structured, hierarchical pathway. Sites had good initial intentions, but practical realities affected the creation and awarding of teacher lead-

ership positions. The implementation challenges are best illustrated through the experience of the CMOs.

Prior to the initiative, the primary career advancement positions for teachers in the CMOs were department chair and mentor for new teachers. In their IP proposal, the CMOs proposed that HE teachers have access to three new pathway options—coach, master teacher, and administrator—and they delineated multiple positions within each pathway. For example, within the coach pathway, effective teachers could become lead teacher, California BTSA support provider, or coach/specialist. The CMO proposal (Partnership of California Charter Management Organizations, 2009, p. 14) explained, "Teachers progress along the career path as they show increasing effectiveness in advancing student achievement."

However, after the grant was awarded, the CMOs' first priority was not career pathways but creating teacher practice measures and developing the teacher-observation scoring process. The CMOs created new positions over time, but they were one-year positions not arranged in any order of skill or responsibility. In the fall of 2015, an Aspire central-office administrator described it as "not a [career] path but a menu of opportunities." Eventually, most of the CMOs created menus of one-year career opportunities for teachers, more like a lattice than a ladder.

Thus, what the CMOs created was more of a differentiated-role strategy than a CL strategy. This differentiated-role strategy met important goals for the CMOs—to engage many teachers productively in addressing the organization's immediate educational needs and to develop many teachers' sense of participation and belonging. However, this approach created a challenge for the research team when we set out to describe CLs and measure teachers' reactions to them. Some of the positions that the sites—the districts and the CMOs—created fall outside the field's usual definition of *CL* because they were temporary, did not focus on teaching, or lacked other features that are typically associated with CLs. Yet these positions should not be ignored, particularly if people in the site perceived them to be part of the site's CL. Moreover, the sites varied considerably in how they defined, described, and dis-

cussed specialized teacher positions, making classification of CL strategies even more challenging.

To address this challenge, we developed a set of features that define a CL position for the purpose of this report, and we tried to apply these consistently across sites. In Tables 8.1 and 8.2, we include teacher positions that meet the following criteria:

- **Eligibility is based on effectiveness.** The position targets top performers, as determined by TE score, professional judgment, or other criteria.
- **The position involves active support for improving teaching through mentoring or coaching of other teachers.** CL teachers might or might not continue as classroom teachers, but they retain a focus on instruction.
- **The position does not focus solely on school administration.** Although the position could build a teacher's qualifications to become an administrator, it is not an administrator position (e.g., AP).
- **There is additional compensation for the additional duties.**

In completing Tables 8.1 and 8.2 about the sites' implementation of CLs, we allowed some exceptions to these criteria. If site leaders thought of a position as part of the site's CL, we included it, even though it might not meet all our criteria. For example, we considered membership on a CMO's advisory panel to be a CL position, although teachers holding this position were not selected based on TE and did not have coaching or mentoring duties.

Tables 8.1 and 8.2 summarize the features of the CLs created in the districts and the CMOs, respectively; Appendix J provides additional details. They show the CL positions that existed in the districts and CMOs prior to the IP initiative and the positions the sites created during the initiative. The data clearly show that all the sites enacted new specialized positions for teachers, although the positions varied considerably in emphasis, permanence, and remuneration. In Tables 8.1 and 8.2, the term *coaching* means providing direct feedback on another person's teaching, usually based on classroom observation.

Table 8.1
District CL Positions Through 2015–2016

CL Position	Years in Effect	Minimum TE Level or Other Qualification	Stipend	Position Duration	Continues to Teach Students?	Number of Schools with This Position	Number of Positions per School	Percentage of Teachers in CL Positions in 2015–2016
HCPS								
CL positions prior to the initiative: team leader, subject-area leader, department head	Ongoing	No information available	Team leader: $758 Subject-area leader: $2,149–$2,887 Department head: $2,125–$3,286 (2015–2016)	No information available	Yes	No information available	No information available	No information available
Mentor: Mentor novice teachers; observe novice teachers who are not mentees.	2010–2011 and ongoing	Minimum five years as an effective teacher (by whatever definition was current at the time of application)	$5,750 ($5,000 before 2014–2015)	Initially two years; extended to up to six years (full IP grant)	No	Not school-based; about 90 total from 2010–2011 and ongoing		Mentors did not have teaching duties.

Table 8.1—Continued

CL Position	Years in Effect	Minimum TE Level or Other Qualification	Stipend	Position Duration	Continues to Teach Students?	Number of Schools with This Position	Number of Positions per School	Percentage of Teachers in CL Positions in 2015–2016
Teacher leader: Provide coaching to teachers in selected high-need schools, both individually and in groups.	2013–2014 and ongoing	HE (level 4 or level 5)	$1,000	Annual, with an option for reappointment at the principal's discretion	Yes	15 high-need schools in 2013–2014, expanded to 30 in 2014–2015		0.25% (30 teacher leaders out of 12,000 teachers)
PPS								
CL positions prior to the initiative: None	Not applicable							
CRI: Coach struggling teachers; help to implement school improvement plan; develop PD materials.	2011–2012 through 2014–2015	P or D on composite measure	$13,300	Three-year term	Yes	One in 2014–2015 Two in the earlier years	12 in 2014–2015	9% (152 teachers in leadership positions out of 1,700 teachers)
PRC: Support excellent teaching in grades 9–10 through daily meetings and teaching in multisubject teams.	2011–2012 and ongoing	P or D on composite measure and at least one year of experience	$9,300	Two-year term	Yes	Six in 2015–2016	Varied; 76 total in 2015–2016	

Table 8.1—Continued

CL Position	Years in Effect	Minimum TE Level or Other Qualification	Stipend	Position Duration	Continues to Teach Students?	Number of Schools with This Position	Number of Positions per School	Percentage of Teachers in CL Positions in 2015–2016
ITL2: Coach and mentor teachers in subjects in which the ITL2 has expertise.	2012–2013 and ongoing	P or D on composite measure and at least three years of experience	$11,300	Two-year term	Yes	2012–2013: Several of the district's highest-need schools Expanded in 2015–2016 to most (40) of the district's schools	Varied; 70 total across schools in 2015–2016	9% (152 teachers in leadership positions out of 1,700 teachers), continued
LES: Coach and mentor teachers who struggle with classroom management.	2011–2012 and ongoing	P or D on composite measure and at least three years of experience	$9,300	Two-year term	No	2011–2012: Seven of the district's lowest-performing schools 2015–2016: District-level positions added	Six (one per school in three schools; three others responsible for multiple schools) in 2015–2016	
SCS								
CL positions prior to the initiative: No information available	No information available							

Table 8.1—Continued

CL Position	Years in Effect	Minimum TE Level or Other Qualification	Stipend	Position Duration	Continues to Teach Students?	Number of Schools with This Position	Number of Positions per School	Percentage of Teachers in CL Positions in 2015–2016
Learning coach: Coach new and struggling teachers.	2013–2014 through 2014–2015	4 or 5 on the observation rubric and minimum composite score of 3, but final choice is at principal discretion	$3,000	One year	Yes	No information available	No information available	No information available
Master teacher: Support new and struggling teachers and learning coaches; evaluate teaching and provide PD.	2013–2014 through 2014–2015	4 or 5 on the observation rubric and minimum composite score of 3, but final choice is at principal discretion	$4,000	One year	Yes	No information available	No information available	

Table 8.1—Continued

CL Position	Years in Effect	Minimum TE Level or Other Qualification	Stipend	Position Duration	Continues to Teach Students?	Number of Schools with This Position	Number of Positions per School	Percentage of Teachers in CL Positions in 2015–2016
PLC coach: Observe teaching and give rubric-based feedback; evaluate teaching.	2013–2014 and ongoing	4 or 5 on the observation rubric and minimum composite score of 3, but final choice is at principal discretion	$6,000	No information available	No	No information available	No information available	No information available, continued
PAR CT: Coach veteran teachers with persistently low TE ratings.	2013–2014 and ongoing	4 or 5 on the observation rubric and minimum composite score of 3, but final choice is at principal discretion	$3,000	None	No	Not school-based		

Table 8.1—Continued

CL Position	Years in Effect	Minimum TE Level or Other Qualification	Stipend	Position Duration	Continues to Teach Students?	Number of Schools with This Position	Number of Positions per School	Percentage of Teachers in CL Positions in 2015–2016
PIT crew: Coach teachers with persistently low TE ratings.	2013–2014 through 2014–2015	4 or 5 on the observation rubric and minimum composite score of 3, but final choice is at principal discretion	$6,000	None	No	Not school-based		No information available, continued

SOURCES: Annual interviews with central-office staff in each of the IP districts.

NOTE: CRI = clinical resident instructor. PLC = professional learning community. PAR = peer assistance and review. CT = consulting teacher. PIT = performance improvement team.

Table 8.2
CMO CL Positions Through 2015–2016

CL Position	Years in Effect	Minimum TE Level or Other Qualification	Stipend	Position Duration	Continues to Teach Students?	Number of Schools with This Position	Number of Positions per School	Percentage of Teachers in CL Positions in 2015–2016
Alliance								
CL positions prior to the initiative: Math teacher trainer (train teachers in own and other schools)	No information available	No information available	$6,000	No information available	Yes	No information available	No information available	No information available
Advisory panel member: Disseminate central-office information and transmit staff feedback regarding the IP initiative.	2011–2012 through 2015–2016	E, HE, or master	$25 per hour	One year	Yes	All or nearly all	One or two	No information available
ALLI coach: Coach new teachers one to three periods per day.	2013–2014 through 2015–2016	HE or master	$2,500	Two years	Yes	All or nearly all	One or two	No information available
Resident mentor: Coach teacher trainees.	2012–2013 through 2014–2015	HE or master	$6,000	One year	Yes	No information available	No information available	No information available
Data fellow and Blended Learning for Alliance School Transformation (BLAST) facilitator: Help teachers interpret student data and integrate computers into instruction.	2013–2014 through 2015–2016	HE or master	$2,500	One year	Yes	No information available	No information available	No information available
Lab or demo teacher: Teach classes observed by new and veteran teachers.	2013–2014 through 2015–2016	HE or master	$2,500	One year	Yes	No information available	No information available	No information available

Table 8.2—Continued

CL Position	Years in Effect	Minimum TE Level or Other Qualification	Stipend	Position Duration	Continues to Teach Students?	Number of Schools with This Position	Number of Positions per School	Percentage of Teachers in CL Positions in 2015–2016
Instructional PD teacher: Design PD for Alliance-wide sessions.	2013–2014 through 2015–2016	HE or master	$25 per hour	One year	Yes	Not school-based		No information available
Aspire								
CL positions prior to the initiative: Lead teacher (lead department meetings and serve on school instructional team)	Ongoing	No information available	No information available	One year	Yes	All or nearly all	No information available	No information available
Advisory panel member and teacher effectiveness driver: Disseminate central-office information and transmit staff feedback regarding the IP initiative.	2011–2012 and ongoing	No minimum qualifications	$500 per year	One year	Yes	All or nearly all	One	50% (333 teachers out of 671 teachers in 38 schools in California and Tennessee)
Instructional drivers (ELA, math, history, NGSS, data, Common Core, equity, online resources, and electives): Facilitate PD; serve as expert at school; develop curriculum.	2013–2014 and ongoing (NGSS 2015–2016 and ongoing; Common Core through 2014–2015)	E, HE, or master	$1,500	One year (NGSS driver for three years)	Yes	No information available	No information available	

Table 8.2—Continued

CL Position	Years in Effect	Minimum TE Level or Other Qualification	Stipend	Position Duration	Continues to Teach Students?	Number of Schools with This Position	Number of Positions per School	Percentage of Teachers in CL Positions in 2015–2016
Video teacher: Have lessons video-recorded so that the recordings can be used for training.	2013–2014 and ongoing	HE or master	$250 per filming	One year	Yes	Not school based	No information available	50% (333 teachers out of 671 teachers in 38 schools in California and Tennessee), continued
Model teacher: Model instruction and provide feedback on other teachers' instruction (five mentees).	2012–2013 through 2015–2016	HE or master	No information available	One year	Yes	No information available		
Peer observer: Observe and rate instruction and provide feedback.	2013–2014 and ongoing	HE or master	$1,500 per year	One year	Yes	Not school based; 50 total in 2013–2014		
Site-based induction coach: Help teachers complete requirements for full credentials.	2013–2014 through 2015–2016	HE or master	$1,500	One year	Yes	No information available	No information available	50% (333 teachers out of 671 teachers in 38 schools in California and Tennessee), continued
Mentor teacher: Provide coaching to new teachers.	2013–2014 through 2015–2016	HE or master	$3,000	One year	Yes	No information available	No information available	
PLC leader: Deliver six training sessions on a specific topic.	2012–2013 and ongoing	E, HE, or master	$500 per year	One year	Yes	No information available	No information available	
Writing team member: Coordinate two writing assessment scoring sessions.	No information available	E, HE, or master	$750 per year	One year	Yes	No information available	No information available	

Table 8.2—Continued

CL Position	Years in Effect	Minimum TE Level or Other Qualification	Stipend	Position Duration	Continues to Teach Students?	Number of Schools with This Position	Number of Positions per School	Percentage of Teachers in CL Positions in 2015–2016
Training facilitator: Facilitate training sessions for new teachers.	2013–2014 and ongoing	E, HE, or master	Up to $600 per year	One year	Yes	No information available	No information available	50% (333 teachers out of 671 teachers in 38 schools in California and Tennessee), continued
Alternative growth measures instructional leader (Memphis only): Help develop growth measures for noncore subjects.	2015–2016 and ongoing	E, HE, or master	$500 per year	One year	Yes	No information available	No information available	
Green Dot								
CL positions prior to the initiative: ILT member (serve as department head and on school ILT); new-teacher mentor (provide orientation, peer observation, feedback, and support to new teachers)	Ongoing	No information available	No information available	No information available	Yes	No information available	No information available	No information available
Green Dot ILT member: Receive training to act as department head.	2012–2013 and ongoing	No information available	Department head stipend	No information available	Yes	All or nearly all	Six	17% (98 teachers out of 569 teachers in 21 schools)
Advisory panel member/site liaison: Disseminate central-office information and transmit staff feedback regarding the IP initiative.	2011–2012 and ongoing	No information available	No information available	One year	Yes	All or nearly all	One	

Table 8.2—Continued

CL Position	Years in Effect	Minimum TE Level or Other Qualification	Stipend	Position Duration	Continues to Teach Students?	Number of Schools with This Position	Number of Positions per School	Percentage of Teachers in CL Positions in 2015–2016
PD leader: Organize five training days in subject of expertise.	2013–2014 and ongoing	Three years of experience, excellent ratings, administrator recommendation	$3,500	One year	Yes	Not school-based. 11 positions total: ELA (MS, HS), history (MS, HS), special education, science (MS, HS), math (MS), visual and performing arts, Spanish, physical education		17% (98 teachers out of 569 teachers in 21 schools), continued
Demo class teacher: Teach four demonstration classes each semester.	2013–2014 and ongoing	HE rating and administrator recommendation	$3,500	One year	Yes	Not school-based. Four positions total: one ELA, one math, and two science, history, or electives		
TIP coach: Help new teachers complete state formative assessment system requirements to obtain credentials.	2013–2014 and ongoing	No information available	$1,500	One year	Yes			
English learner lead: Give three PD sessions at the school site.	2013–2014 and ongoing	No information available	No information available	One year	Yes	All or nearly all	One	
Special-education new-teacher support advisers: Hold biweekly office hours; design and facilitate two PD sessions.	2013–2014 and ongoing	No information available	$1,500	One year	Yes	Not school-based. Two teachers total: one MS, one HS		

Table 8.2—Continued

CL Position	Years in Effect	Minimum TE Level or Other Qualification	Stipend	Position Duration	Continues to Teach Students?	Number of Schools with This Position	Number of Positions per School	Percentage of Teachers in CL Positions in 2015–2016
National Expansion Leadership Collaborative: Plan summer training sessions for new teachers; mentor teachers; host visits from teachers (Tennessee and Washington teachers).	2015–2016 and ongoing	Two years of experience on ILT	$1,000	One year	Yes	Not school-based. Four teachers		17% (98 teachers out of 569 teachers in 21 schools), continued
Curriculum review and development positions: Curriculum specialist, core curriculum review team, sheltered ELA revision committee, special-education curriculum and assessment advisor and academic success working team, technology pathways review team	2014–2015 and ongoing	No information available	No information available	One year	Yes	Not school-based. 11 positions total		
PUC								
CL positions prior to the initiative: Learning lab demo teacher (teach demonstration classes); induction support provider (assist new teachers in fulfilling state credential requirements)	Ongoing	Learning lab demo teacher: principal and central-office approval; induction support provider: nominated by principal and screened by central office for achievement and performance management	No information available	One year	Yes	Not school-based. Four lab classroom and demo teachers in 2015–2016; 22 induction support providers in 2015–2016		No information available

Table 8.2—Continued

CL Position	Years in Effect	Minimum TE Level or Other Qualification	Stipend	Position Duration	Continues to Teach Students?	Number of Schools with This Position	Number of Positions per School	Percentage of Teachers in CL Positions in 2015–2016
Advisory panel member: Disseminate central-office information and transmit staff feedback regarding the IP initiative.	2011–2012 through 2015–2016	Good evaluation and principal's recommendation	No information available	One year	Yes	All or nearly all 15 in 2015–2016	One	20% (54 teachers out of 272 teachers in 15 schools)
Content coordinator and assistant content coordinator: Create sample lessons available online.	2014–2015 and ongoing	Three years of teaching experience	No information available	No information available	Yes	Not school-based. Four total in 2015–2016		
Alumni Teach Project mentor: Mentor trainees in residency program targeting PUC alumni.	2015–2016 and ongoing	Good evaluation and principal's recommendation	No information available	One year	Yes	Not school-based. Ten total in 2015–2016		
Common Core Pioneer: Model Common Core strategies for teachers at the school.	2013–2014 only	Good evaluation and principal's recommendation	No information available	One year	Yes	No information available	One	

SOURCES: Annual interviews with central-office staff in each of the IP CMOs.

NOTE: ALLI = Alliance Learning and Leadership Initiative. NGSS = Next Generation Science Standards. ILT = instructional leadership team.

The term *mentoring* is more general and means providing information and support on a variety of matters relevant to teachers, not limited to teaching practice.

The districts created new CL positions during the IP initiative; most focused on coaching or mentoring novice or struggling teachers. Table 8.1 shows that the districts implemented a mixture of full-time coaches, who did not have teaching responsibilities during their service as coaches (e.g., mentors in HCPS), and regular classroom teachers, who provided coaching or mentoring while still having teaching responsibilities (e.g., CRI in PPS). All three districts created both types of positions. Almost all the CL roles focused on providing mentoring support for first- or second-year teachers or on providing coaching to more-experienced teachers who were not performing well. In HCPS and SCS, teachers in these roles received stipends ranging from $1,000 to $6,000. In PPS, the stipends were larger, ranging from $9,300 to $13,300; this might reflect the fact that these CL positions were in struggling schools and the stipends were set at a level designed to attract effective teachers to these schools. The larger stipends might also reflect the value that PPS placed on coaching, the higher base salaries in PPS, or the presence of a strong teachers' union. In many cases, the new positions were in or targeted low-performing schools (teacher leaders in HCPS, ITL2s in PPS, and PLC coaches in SCS). In HCPS and PPS, many of the positions lasted for two or more years; in SCS, most were one-year positions. To the best of our knowledge, only a small percentage of teachers served in CL roles (less than 1 percent in HCPS and 9 percent in PPS).

The CMOs created new CL positions during the IP initiative, offering many different CL roles that teachers could fill for a year at a time. Table 8.2 shows the CL positions that the CMOs implemented. The CMOs worked collaboratively (together with TCRP) to apply for the IP grant and to implement their teacher-evaluation systems. One of the first steps each CMO took was to identify teachers to serve on advisory panels to engage in planning the initiative. They offered these teachers stipends for taking on additional responsibilities, which is consistent with the CMOs' approach to CL positions—they were primarily one-year assignments to address identified needs across

a wide range of educational areas, including curriculum and instructional support. A quick glance at Table 8.2 shows that each of the CMOs implemented many different CL positions, and many teachers held such roles. For example, every CMO continued to have an advisory panel to communicate between the central office and the schools. None of the CMO CL positions was full time; all involved extra duties in addition to teaching responsibilities.

Like in the districts, many of the CMO CL roles focused on providing coaching or mentoring support for first- or second-year teachers or on providing coaching to more-experienced teachers who were not performing well. In contrast to the positions in the districts, however, some CMO positions involved other sorts of responsibilities:

- Alliance data fellows helped teachers interpret student data.
- Aspire math instructional drivers facilitated PD.
- Green Dot English learner leads provided PD for other teachers.
- PUC content coordinators created sample lessons that were available online.

Almost all the CMO positions were for one year, and teachers received stipends ranging from about $1,000 to $3,000. Some positions were paid hourly. Unlike in the districts, CMO CL positions were awarded to many teachers—fully half the teachers in Aspire held CL positions in 2015–2016, and about 20 percent of teachers held them in Green Dot and PUC. Moreover, positions were established in every school. But, like in the districts, none of the CMOs implemented a structured, hierarchical pathway.

Implementation Challenges

Perhaps because the sites' limited approach to CLs affected relatively few teachers in most sites and reportedly had few negative consequences, CL policies were implemented without many problems. Nevertheless, our interviews with staff in the sites revealed a few challenges that were relevant to designing and implementing even these limited types of

CLs. These challenges related to the nature of the career options that were created, the timing of the process, the supply of applicants for the positions, and the influence of external factors.

For a variety of reasons, including costs and external factors, the sites did not implement hierarchical CLs. All sites except PPS included hierarchical CLs in their initial IP plans but then decided to not implement them. In SCS's original IP proposal, the district planned to implement hierarchical CL roles in which responsibilities and compensation would increase with experience and effectiveness. However, SCS never implemented the clear career pathway for teachers outlined in the proposal. Staff we interviewed mentioned several factors that contributed to this broad retreat from the plan, including the local teachers' association's opposition to performance-based pay and frequent turnover in the staff who were tasked with developing and implementing the CL program. Some administrators we interviewed indicated that they believed that the district was never really committed to implementing CLs. After the merger between legacy MCS and legacy SCS in 2013, the district implemented three coaching roles (learning coach, master teacher, and PIT crew) as CL positions, but these positions were not hierarchical (in terms of advancing from one role to another). In addition, they were discontinued after two years because TNTP found that the district's coaching program did not improve teachers' TEM scores (TNTP, 2014).

In HCPS, the site did not try to launch the full hierarchical CL that had been planned as part of the proposal until late in the grant period, and it had not yet been completed when the grant ended. Central-office interviewees reported that the delay was initially due to prioritizing other aspects of the initiative. Because the mentors were necessary to fully implement the TE measure, that role was launched right away, but the hierarchical ladder did not have a similar impact. By the time HCPS began to look at the CL, uncertainty about whether sufficient grant funds would be available to pay the consultant hired to assist in developing the CL hampered the district's efforts to get started.

The CMOs' plans for hierarchical career pathways for HE teachers evolved over time in response to local needs. Green Dot made a con-

scious decision to recast its original single pathway into a more diverse set of options for teachers to broaden their experience by taking on different leadership roles. Green Dot described it to us as "more horizontal than vertical." The other CMOs did not point to specific decisions to change direction but an evolution of thinking about broadening options for teachers. For example, Aspire delegated to regional leaders the responsibility to develop CL roles to meet local needs. Similarly, PUC changed its leadership opportunities in response to the needs and expertise it found among its teachers; as one respondent noted, "professional development and mentorship that is led by peers is particularly effective."

PPS implemented nonhierarchical CL roles, including LES, PRC, and ITL2, much as initially planned.

In many cases, it was difficult to coordinate the timing of the application and decision process with the timing of the TE ratings used to qualify for the CL positions. The logical cycle for new positions was to have applications in the spring and decisions before the school year ended so teachers could start in their new roles at the beginning of the new school year. Often, this did not match the timing of the evaluation cycle. For example, in the CMOs, the TE composite scores were not available until the fall after the school year in which observations and tests occurred. To accommodate the delay in TE scores, the CMOs used scores either from previous years or from measures available before the end of the school year, such as the observation scores and stakeholder survey scores. PPS accommodated the timing problem by using a "lagged" VAM score (described in Chapter Three) that was based on data from the prior year. SCS computed a "temporary" composite based on the prior year's achievement growth estimate for initial decisions and then updated it later in the year.

In many sites, few teachers applied for the CL positions. Rather than being highly competitive, the positions attracted limited numbers of teachers. The experiences of the CMOs illustrate several reasons teachers might not have found the positions attractive. In the CMOs that adopted a large number of diverse one-year positions, some teachers said that the jobs were not prestigious enough to warrant the extra work. Other teachers reported that their regular teaching posi-

tions required so much of their time that they did not have sufficient time to take on additional responsibilities. At PUC, new teachers did not yet qualify for CL positions; in all the CMOs, some of the positions required that teachers be rated as HE or master, and new teachers did not yet have ratings. Furthermore, many new teachers said that they were already working to their maximum to make their first year a success. In 2015, a Green Dot central-office administrator summed up a number of these points:

> There are so many new teachers [that] everyone [who is a] veteran has to take on all these school site roles, plus organizational tasks. And people are getting married, having families—that's the age of our organization now. [Teachers say,] "I just want to do my job and go home. I don't have time to come to the home office for three hours. . . ."

Finally, the adoption of the Common Core refocused many CMO teachers' attention on curriculum and instruction, and the supplemental duties associated with many of the CL roles were less germane to their interests.

Some of the districts, too, found that few teachers were attracted by the CL options. For example, initial interest was low in PPS because the positions were offered only in selected schools, which meant that accepting a position might require a transfer and a concomitant loss of seniority; if future furloughs occurred, teachers who transferred would be among the first to be displaced. Thus, experienced teachers with seniority in their current schools particularly stood to lose from accepting the transfer-requiring positions. After about four years, PPS and the teachers' union reached an agreement that teachers who transferred could keep their building seniority, which made the positions more attractive. When PPS expanded the CL roles to other schools, there was a further increase in interest among teachers, according to district staff. Even so, the initial implementation problems in PPS (e.g., some teachers' unwillingness to switch schools) dampened support among some teachers for CLs in that district in the early years of the program (as we also discuss in the section on reactions, next).

In contrast, a successful launch in HCPS built momentum for broadening the program. The mentor program was well received by new teachers, and multiple central-office staff cited its positive effects on new-teacher retention. Other teachers, particularly teachers who were new to HCPS but had prior teaching experience, expressed interest in receiving similar support. This created a positive environment for creating a new teacher leader instructional mentor position, piloted in select schools starting in 2013–2014 and expanded district-wide in 2016–2017. The pilot program allowed HCPS to better define the role and determine what qualities enabled teachers to be a good fit for the role.

Teacher and SL Reactions to CL Policies

To examine educators' perceptions of the sites' CL policies and positions, we analyzed their survey and interview responses, on four broad topics:

- awareness: Did teachers and SLs know about the CL policies and positions?
- endorsement: Did teachers and SLs approve of the CL policies and positions?
- fairness: Did teachers and SLs think that the CL policies and positions were fair?
- perceived effects: Did teachers find the CL policies motivating, and did they (and SLs) think that people in CL positions were helpful?

Awareness

Teachers and SLs generally knew whether their site had CL positions. Our analysis of whether teachers were aware of their site's CL

policies and positions was based largely on three survey questions that defined a CL as follows:

> In a career ladder, teachers may be promoted and are given additional pay to take on new or different responsibilities, such as mentoring other teachers, typically without having to give up teaching. The positions on a career ladder may vary, but the higher-level positions typically have titles like "teacher leader" or "master teacher."[1]

The three awareness questions were (1) whether the site had or was phasing in a CL and, if so, (2) whether the respondent's school had any CL teachers, including the respondent, and, (3) if the respondent held a CL position, what the position title was. (See Table K.1 in Appendix K for the exact wording of the questions, how respondents were routed to each question, and the response options.)

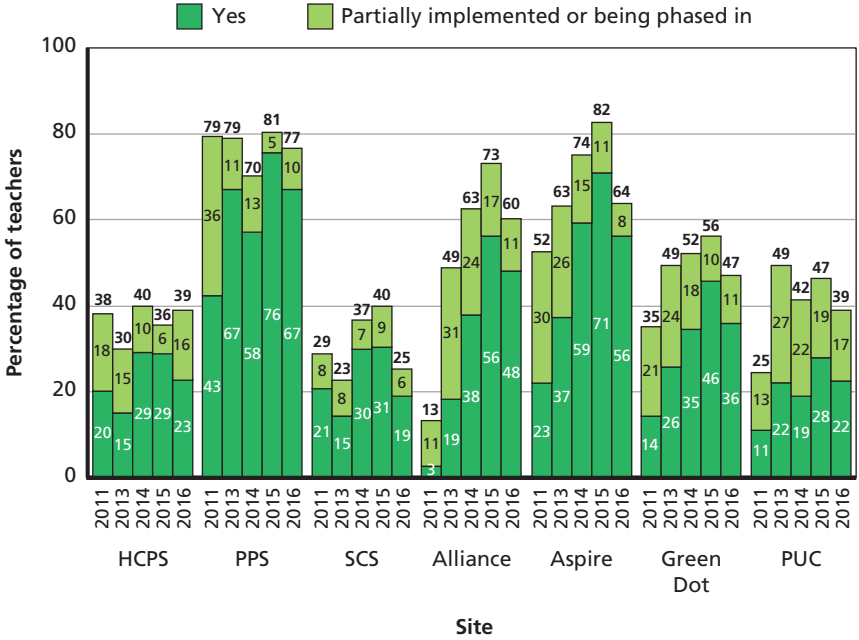
Analyses indicated that teachers' awareness of their site's CL policies and positions was generally consistent with the actual policies and positions implemented in any given year, as described in Tables 8.1 and 8.2. Figure 8.1 presents responses to the first survey question, and Figure 8.2 presents responses to the second survey question. The third was a write-in question, and the written responses were quite diverse.

As shown in Figures 8.1 and 8.2, teacher awareness and understanding of CL policies was particularly striking in PPS and Aspire. Majorities of teachers in each of these two sites correctly reported in every year that the site had (or was phasing in) a CL (Figure 8.1) and that *their school* had teachers in CL positions (Figure 8.2). In addition, among PPS and Aspire respondents who said that they held a CL position, nearly all wrote in titles corresponding to the positions listed in Table 8.1 (for PPS) and Table 8.2 (for Aspire). And in years after cer-

---

[1] This definition is not perfectly consistent with the definition and criteria set forth earlier in this chapter. In particular, whereas the CL positions described in Tables 8.1 and 8.2 include nonteaching positions, the survey definition's inclusion of the phrase "typically without having to give up teaching" might have led some respondents to exclude nonteaching CL positions from consideration in answering the survey questions about CLs.

Figure 8.1
Teachers Reporting That Their Site Had or Was Phasing in a CL or Specialized Instructional Positions, Springs 2011–2016



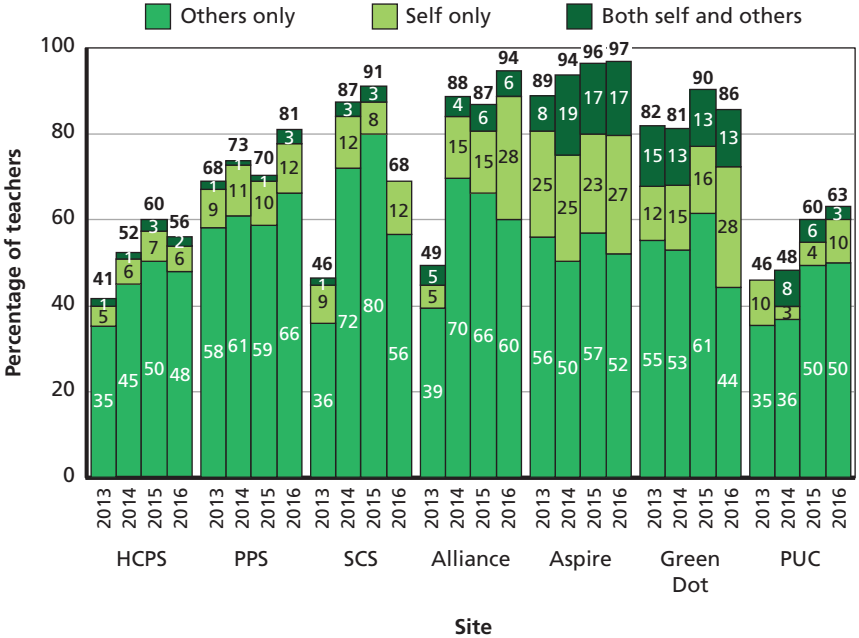
NOTE: Omitted response categories are "no" and "don't know." We did not survey teachers in 2012.

RAND *RR2242-8.1*

tain positions had been discontinued, those titles disappeared from the write-in responses.

Although, in every year in SCS, minorities of teachers reported that their site had in place a CL for teachers, the percentages reporting a CL were higher in the two school years (2013–2014 and 2014–2015) when the site offered the two school-based CL positions (learning coaches and master teachers) (see Figure 8.1). Moreover, in those two years, more than 85 percent of the teachers who reported that the district had a CL indicated that there were teachers who held higher-level CL positions *at their school*, in contrast with considerably lower

Figure 8.2
Teachers Reporting That There Were Teachers at Their School Who Held Higher-Level CL or Specialized Instructional Positions, Springs 2013–2016



NOTE: We asked this question for the first time in the spring of 2013. We asked it only of teachers who said that their site had a fully or partially implemented CL. Omitted response categories are "no [one]" and "don't know."

RAND *RR2242-8.2*

percentages in the other years (see Figure 8.2).[2] Accordingly, in the following discussion on attitudes toward CLs, we include SCS only for the two middle years, for which we have confidence that we know the positions to which the survey respondents were referring.

In Alliance, awareness of CLs grew after the site put positions in place during the 2013–2014 school year (see Figures 8.1 and 8.2); for this reason, we include Alliance in the discussion of attitudes toward

[2] In addition, among SCS respondents who said that they themselves held CL positions—of whom there were many more in school years 2013–2014 and 2014–2015 than in school years 2012–2013 and 2015–2016—about 80 percent wrote in "learning coach" or "master teacher" in the two middle years, far more than the very few in the two outside years.

CLs beginning with the 2013–2014 year. However, in none of the years was there a strong correspondence between the CL titles that Alliance respondents wrote in on the survey and the official CL titles that the CMO staff described to us. Thus, there appears to have been some confusion in Alliance about the CL positions—at a minimum, respondents were unclear about the titles of the positions, and perhaps they were also unclear about other aspects of the roles.

In Green Dot and PUC, minorities of teachers reported that CLs existed throughout the period (see Figure 8.1), which is consistent with how the CL policies operated in those two sites. In Green Dot, although the site created several CL positions, many of the positions were not school-based, nor were they widely available. PUC, meanwhile, did not have many CL positions, and of the positions it did have, few fit the survey definition well. Nevertheless, because the level of awareness was consistent with site policy, and because most respondents to the CL questions said that there were CL holders in their school (see Figure 8.2), we include all years for both sites in the subsequent discussion.

In HCPS, the percentage of teachers saying that the district had in place a CL for teachers was consistently low—40 percent or lower—in every year from 2011 through 2016 (see Figure 8.1). Most likely, HCPS respondents who said "yes" to the initial CL question were thinking of existing school-based positions, such as department heads, grade-level team leaders, and subject-area leaders; among respondents who said that they themselves held a CL position, typically about half wrote in such titles. However, there was apparently some ambiguity among respondents about whether such positions "counted" as CL positions. Despite these positions being widespread, few reported that HCPS had a CL or, as shown in Figure 8.2, that their school had CL teachers. For that reason, and because these positions were not created as part of the initiative, we exclude HCPS from the following discussion about attitudes.

Like the teacher survey, the SL survey asked respondents whether their site had in place a CL or specialized instructional positions for teachers and, if so, whether there were teachers in higher-level CL or specialized positions *at their school*. For the most part, SLs' responses

to these questions closely paralleled the teachers' responses (see Figures K.1 and K.2 in Appendix K). For instance, on the first question, clear majorities of SLs in PPS and Aspire said "yes," while fewer said "yes" in HCPS and SCS. On the second question, high percentages of SLs said "yes" not only in PPS and Aspire but also (as did the teachers) in SCS in the two middle years only and in Alliance after 2013.

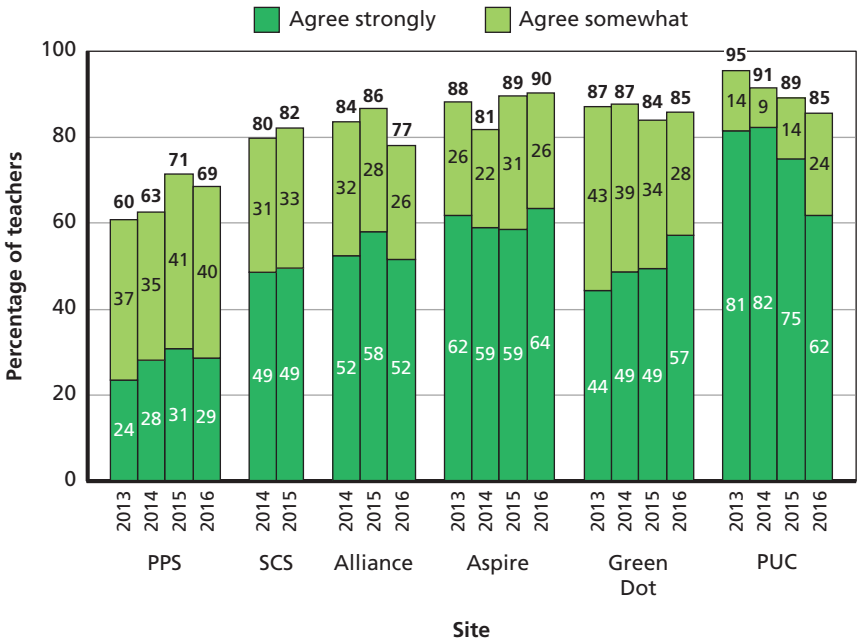
Endorsement

In most of the sites, majorities of teachers thought that the CL teachers at their school were effective educators who deserved the additional compensation they were receiving; agreement was lower in PPS than in the other sites.[3] For survey respondents who reported that their schools had CL teachers and that they themselves were *not* CL teachers, the survey then asked whether they thought that the teachers who held higher-level CL and specialized positions at their school (1) were effective educators and (2) deserved the additional compensation (bonuses or higher salaries) they were receiving. To facilitate interpretability, we present the results only for the sites and site/year combinations (site-years) in which, as discussed in the previous section, we found consistency between educators' understanding of CL policies (as indicated on the survey) and the sites' actual policies (as described in Tables 8.1 and 8.2).

In most sites, in most years, at least 80 percent of teachers agreed that the CL teachers at their school were effective educators (see Figure 8.3). The one exception was PPS, in which a smaller majority (60 to 70 percent) agreed with the statement. The percentages of teachers agreeing strongly were notably high in PUC but should be interpreted with caution because relatively few PUC teachers answered the survey question (because less than half of all PUC teachers reported

[3] Because most of the sites did not implement new CL positions until 2011–2012 or later, we present attitude findings from the spring of 2013 forward. (There was no teacher survey in the spring of 2012.)

Figure 8.3
Teachers' Agreement That the Teachers Who Held Higher-Level Career Ladder and Specialized Positions at Their School Were Effective Educators, Selected Site-Years



NOTE: Sites and site-years included are based on the analysis of awareness presented earlier. Omitted response categories are "disagree somewhat," "disagree strongly," and "don't know."

RAND *RR2242-8.3*

that the site had a CL and, of those, only half reported that their school had CL teachers other than themselves).[4]

Interviews suggested that the lower level of agreement in PPS might have been the result of the hiring process for CL positions in

[4] There were no consistent differences in judgments about the effectiveness of CL teachers when we compared responses from teachers in higher-LIM and lower-LIM schools. This was the case for both of the two types of LIM comparisons we made: comparing high-LIM schools (i.e., schools with at least 80 percent LIM students) with all other schools in the site and comparing schools above and below the site-wide median percentage of LIM students. (See "Survey Methods" in Appendix A.)

the first few years of implementation. In the first year of implementation, PPS received far fewer applications for CL positions than anticipated, so the district subsequently reduced the experience requirement. As a result, it awarded some of these positions to relatively inexperienced teachers. Some teachers we interviewed in 2011, 2012, and 2013 reported that some CL teachers were not effective because of their limited teaching experience.[5] In addition, each year, about half the non-CL teachers we interviewed in PPS reported that the CL roles—particularly the CRI and LES roles—were not well defined, and some described the positions as unnecessary or extraneous.

Most teachers in most sites agreed that CL teachers deserved their additional compensation, although the percentages agreeing with the compensation statement tended to be lower than the percentages agreeing with the effectiveness statement, especially in PPS and SCS. For example, in PPS, across the years, only 40 to 55 percent of teachers agreed that the CL teachers deserved their additional compensation.[6]

As noted in Tables 8.1 and 8.2, the stipend amounts that PPS CL teachers received were higher than the CL stipends in the other sites. Because the CL positions were so well known in PPS (recall the high levels of awareness there), the high stipend amounts that the CL teachers received might have created some resentment among non-CL teachers. Indeed, some teachers we interviewed said that there was resentment about CL teachers, who taught less and were paid more, particularly if it was not clear to others what they had done to earn those privileges. SLs in PPS also tended to be less likely than SLs in the other sites to agree with the two survey statements, particularly the

---

[5] On the 2013 survey, PPS novice teachers were significantly ($p < 0.001$) more likely than PPS experienced teachers to agree that their school's CL teachers were effective: 100 percent of novice teachers agreed but only 66 percent of experienced teachers agreed. The two groups' responses were not, however, significantly different in subsequent years.

[6] In all years, PPS novice teachers were more likely than PPS experienced teachers to agree that their school's CL teachers deserved the additional compensation (bonuses or higher salaries) they were receiving. The difference was significant in 2013 ($p < 0.001$) and 2015 ($p < 0.05$). Also, PPS teachers in the upper half of the school LIM distribution (i.e., schools with relatively more LIM students) were significantly more likely than teachers at lower-LIM schools to agree in 2014 ($p < 0.05$) and 2016 ($p < 0.001$).

one about compensation. In the next section, we present information about the perceived fairness of the selection process, which might have affected judgments about the effectiveness of the people in CL roles.

Fairness

Majorities of teachers in all sites agreed that the process for selecting CL teachers was fair, but few agreed strongly, and agreement was lowest in PPS. Like with the analysis of endorsement, we analyzed perceptions of fairness only in sites and site-years in which there was consistency between educators' understanding of CL policies and the sites' actual policies. Figure 8.4 shows that a majority of teachers who said that their site had implemented a CL (regardless of whether their school had CL teachers) agreed that the process by which teachers were selected for the various CL and specialized positions was fair, but less than one-third agreed strongly.[7] Again, agreement was lower in PPS than in the other sites. The lower agreement in PPS, particularly in the early years of CL implementation, might be related to the fact that some CL roles went to teachers with less experience, as described earlier.[8] In addition, the launching of the CRI role in 2011–2012 unfortunately coincided with a reduction in overall staffing levels and furloughs for about 180 teachers. It did not sit well that some relatively inexperienced teachers were promoted into CL roles while other, more-experienced teachers lost their positions.

Perceived Effects

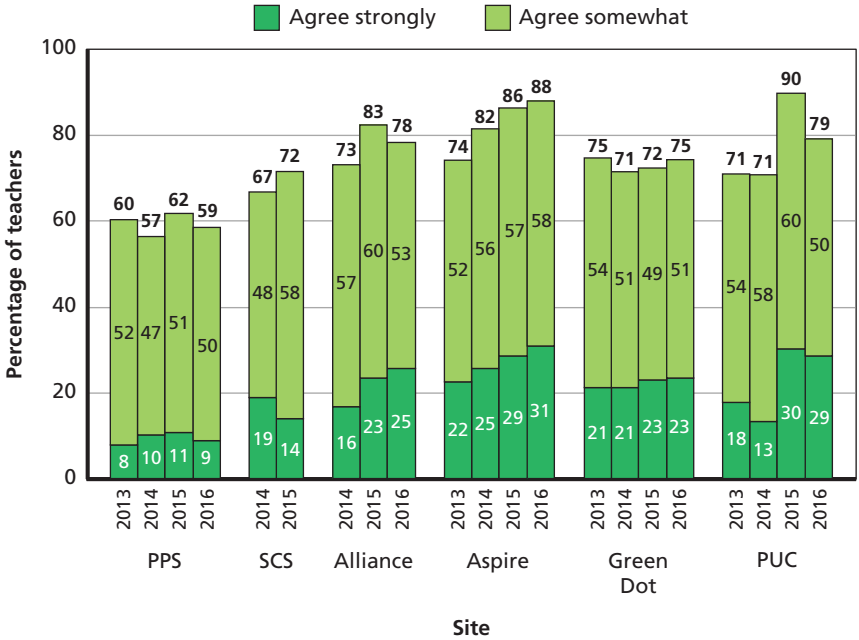
Except for experienced teachers in PPS, majorities of teachers said that they aspired to become CL teachers and that the opportunity

---

[7] Again, the percentage of teachers agreeing that the selection process was fair did not differ systematically between teachers in higher-LIM and lower-LIM schools. This was the case for both types of LIM comparisons we made: comparing high-LIM schools (i.e., schools with at least 80 percent LIM students) with all other schools in the site and comparing schools above and below the site-wide median percentage of LIM students.

[8] Interestingly, however, PPS novice teachers were no more likely than PPS experienced teachers to agree that the CL teacher selection process was fair. In fact, starting in 2014, they appear to have been slightly *less* likely to agree (but the differences were not significant in any year).

Figure 8.4
Teachers' Agreement That the Process by Which Teachers in Their Site Were Selected for the Various CL and Specialized Positions Was Fair, Selected Site-Years

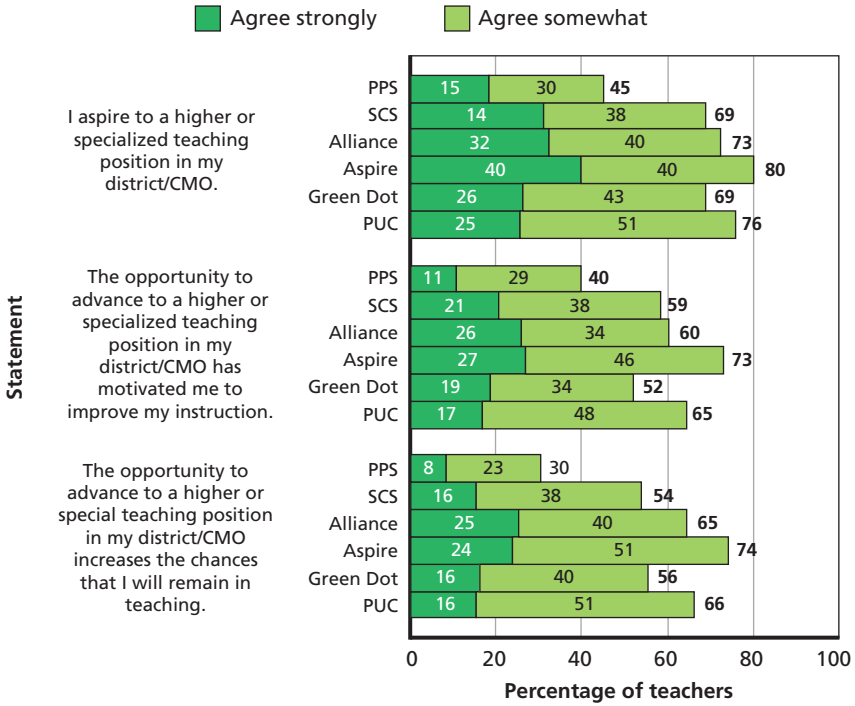


NOTE: Sites and site-years included are based on the analysis of awareness presented earlier. Omitted response categories are "disagree somewhat" and "disagree strongly."

RAND *RR2242-8.4*

to do so motivated them to improve their instruction and increased the chances that they would remain in teaching. Like with the analyses of endorsement and fairness, we analyzed perceived effects only in sites and site-years in which there was consistency between educators' understanding of CL policies and the sites' actual policies. Except in PPS, a majority of teachers—typically 60 to 80 percent—who said that their site had a fully or partially implemented CL or specialized instructional positions in 2014–2015 reported that they aspired to such positions. The top section of Figure 8.5 shows results for 2014–2015. We also asked teachers whether the opportunity to advance to higher or

Figure 8.5
Teachers' Agreement with Statements About Advancing to Higher or Special Positions, Spring 2015



NOTE: Sites included are based on the analysis of awareness presented earlier. Omitted response categories are "disagree somewhat" and "disagree strongly."

RAND *RR2242-8.5*

special positions motivated them to improve their instruction. Agreement on this item tended to be a little lower than on the aspiration item, but, again, in each site except PPS, more than half of the teachers agreed (see the middle section of Figure 8.5). Lower agreement among teachers in PPS is again likely explained at least partly by the fact that initially some of the positions were given to less experienced teachers; other teachers might have concluded that effort was irrelevant to being awarded a CL position. In interviews, many PPS teachers noted that the effectiveness of CL positions depended on who was serving. As one teacher said, CL roles "are potentially effective, but it depends on

who is doing it." Finally, we asked teachers whether the opportunity to advance to a higher or special position increased the chances that they would remain in teaching. Responses to this item (shown in the lower section of Figure 8.5) were similar to those on the prior item.

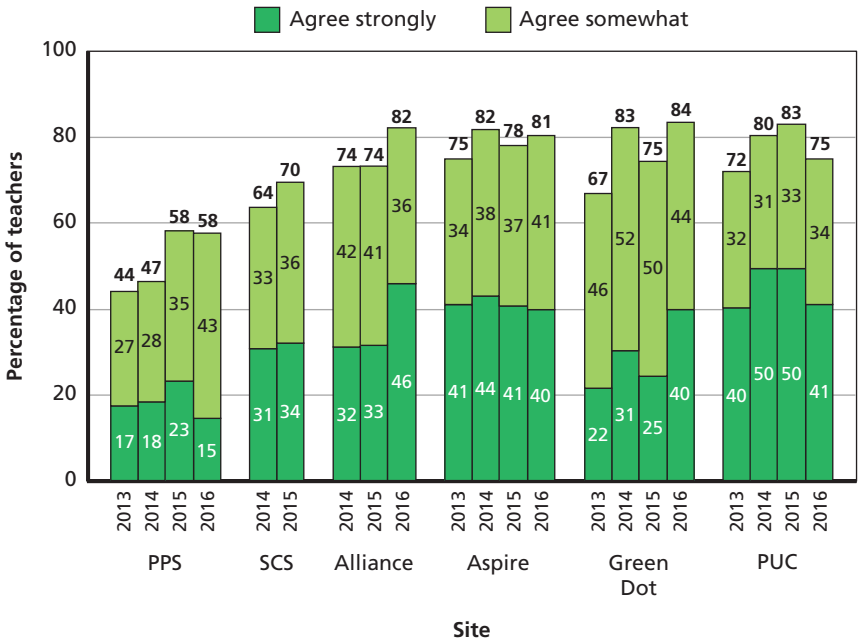
For the most part, responses to these three survey items in other years were similar (see Figure K.3 in Appendix K). Aspire placed more emphasis on CL roles than any of the other CMOs did, which might explain the higher positive responses in that site. In 2014 in PUC and 2015 in PPS, teachers in high-LIM schools (defined as schools having at least 80 percent LIM students) were more likely than teachers in other schools in those sites to agree with all three statements.

In general, novice teachers in all sites were more likely than more-experienced teachers to agree with these statements; in PPS, novice teachers were about *twice* as likely as experienced teachers to agree. For example, on the "motivated to improve my instruction" item in 2015, 79 percent of PPS novice teachers agreed, double the 38 percent of PPS experienced teachers.

Except in PPS in the early years, majorities of teachers said that the CL teachers at their school had helped them improve their instruction. Teachers who said that their school had CL teachers were also asked whether the CL teachers had helped them improve their instruction. Majorities of teachers in all sites agreed with this statement in 2015 and 2016. Agreement was lower in PPS every year, and less than half of PPS teachers agreed in 2013 and 2014. In each CMO, in at least one year, more than 80 percent of teachers agreed that CL teachers had helped them improve (see Figure 8.6). In addition, in 2014 and 2015, novice teachers in every site except PPS were more likely than experienced teachers to agree,[9] which is consistent with the focus of many CL positions on supporting the development of new teachers. In PPS, although most novice teachers valued the opportunity to become CL

[9] In every site except PPS, novice teachers were more likely to agree with the statement in both 2014 and 2015, and the difference was significant in all the sites except Green Dot. In PPS, the percentage of novice teachers agreeing was 13 points lower than the percentage of experienced teachers agreeing in 2015, but this difference was not significant. In 2014, the two PPS groups' percentages were equal.

Figure 8.6
Teachers' Agreement That the Teachers Who Held Higher-Level Positions at Their School Had Helped Them Improve Their Instruction, Selected Site-Years



NOTE: Sites and site-years included are based on the analysis of awareness presented earlier. Omitted response categories are "disagree somewhat" and "disagree strongly." (As the question appeared on the survey, there was also a "don't know" option, but, for analysis of this item, we coded responses of "don't know" as missing.)
RAND *RR2242-8.6*

teachers themselves, they were no more likely than experienced teachers to find the incumbent CL teachers helpful.

Most SLs (80 to 100 percent in most sites and years) agreed strongly or somewhat that the CL teachers at their school had helped other teachers at the school improve their instruction. Like with the teacher responses, the percentages of SLs agreeing were lower in PPS than in other sites, particularly for the percentages agreeing strongly (typically 40 percent or less in PPS, less than the 50 to 70 percent in the other sites).

Finally, it also appears that CL teachers were helpful in relieving SLs of some administrative responsibilities. We asked SLs whether the CL teachers had "taken on some responsibilities that used to be mine, freeing me to do other important things." In most years and sites, 60 to 80 percent of SLs agreed. For example, interviewees in SCS said that some PLC coaches took on administrative duties for a portion of their day.

Teacher Outcomes Associated with CL Policies

As noted earlier in this chapter, CLs were implemented to increase the retention of effective teachers and to put them into coaching and mentoring roles, in which they could improve the performance of other teachers. However, these outcomes were also influenced by other IP policies—e.g., effectiveness-based compensation was designed to improve retention, evaluation-linked PD was implemented to improve the performance of ineffective teachers—and so we cannot disentangle the simultaneous effects of these reforms.

In addition, CLs might have the secondary effects of improving LIM students' access to effective teachers if CL positions were located at high-LIM schools, and they might improve the effectiveness of new hires if CL positions targeted mentoring new hires. They might even have labor market effects if they made teaching more attractive to potential candidates. Chapter Eleven examines the retention of effective teachers and presents evidence about the likely impact of relevant policies, including CLs. Chapter Twelve presents evidence about the distribution of TE, with a focus on LIM students' access to effective teachers; it considers the effects that IP policies, including CLs, likely had on these outcomes.

Summary of the CL Lever

Research on the effectiveness of CLs is limited, and we cannot draw clear conclusions about the best structure for a set of career pathways.

For example, it is not clear whether a rigid hierarchy is preferable to a flexible lattice. On the other hand, research suggests that effective CLs benefit from good communication, meaningful tasks, fair selection, public recognition, and supplemental resources. We do not have precise measures of all these features in the IP sites, but we do have information about the extent to which the sites implemented CLs in ways that are consistent with the findings from research.

All seven sites implemented CL policies, although the positions that they created varied in number and focus. The three districts primarily used their CL positions to provide instructional support to newly hired or struggling teachers. Although the four CMOs had a few positions with a new-teacher support focus, they also created specialized positions designed to provide shared benefits, such as curriculum development, internal communication, and subject-matter leadership. In the districts, many of those tasks were assigned to central administrative departments.

For the most part, teachers were aware of their sites' CL policies, and teachers' judgments about CLs were positive. In all sites except PPS, majorities of teachers who had experience with CL teachers thought that they were effective educators who deserved the additional compensation they received. Similarly, in all sites except PPS, majorities of teachers said that the CL teachers at their school had helped them improve their instruction. Teachers generally agreed that the process for selecting CL teachers was fair, and most teachers in most sites aspired to become CL teachers. Perhaps most important is that more than half the teachers in most sites agreed somewhat or strongly that the opportunity to become a CL teacher motivated them to improve their instruction and increased the chances that they would remain in teaching.

PPS was a notable exception to this generally positive picture. For a variety of reasons—including problems that occurred during the initial implementation of the program—teachers in PPS were far less likely than teachers elsewhere to express positive sentiments about CL policies. It appears that experienced teachers did not want to change schools to take CL positions. The low interest from experienced teachers meant that some positions were given to less experienced teachers.

This decision might have depressed support for CL positions among PPS teachers.

We are unable to judge the extent to which CL policies by themselves improved teacher retention or effectiveness. In Chapters Eleven and Twelve, we examine whether the collection of policies, including CLs, had positive effects on these two intermediate outcomes.

CHAPTER NINE

Resources Invested in the IP Initiative

In Brief

How much did the sites spend, in terms of both fiscal and human resources, to implement the IP initiative?

Our analysis found that, from the inception of the IP initiative in November 2009 through June 2016, the seven sites collectively spent $575 million on the initiative, with per-pupil expenditures ranging from $868 in Green Dot to $3,541 in PPS. As a percentage of total site budget, the expenditures on the initiative ranged from 1.7 percent in Green Dot to 6.6 percent in Aspire.

The largest source of funding for the initiative was the Bill & Melinda Gates Foundation, which awarded approximately $212 million across the seven sites. In five of the seven sites, the foundation was the largest provider of funds. Across the sites, the percentage of overall expenditures on the initiative that the foundation funded ranged from 28 percent in Aspire to 53 percent in Green Dot. For HCPS, PPS, Green Dot, and PUC, the largest source of matching funds was district or CMO funds. For SCS, Alliance, and Aspire, federal funding provided the largest source of matching funds.

In five of the seven sites, activities related to compensation and CLs received the largest share of the total funding allocated to the IP initiative. The exceptions, SCS and PUC, spent the largest shares of their total IP funding—42 and 37 percent, respectively—on PD.

An additional cost of the IP initiative, beyond direct expenditures, was the time teachers and SLs spent on activities related to the teacher-evaluation component. By assigning a dollar value to the time dedicated

to these activities, we estimate that IP costs for teacher-evaluation activities totaled nearly $100 million across the seven sites in 2014–2015. We estimated the value of teacher and SL time devoted to evaluation to be about $73 million, and the direct expenditures on evaluation constituted an additional $26 million. In per-pupil terms, the overall cost for evaluation activities averaged almost $280: $201 for the value of time spent on evaluation activities and $78 for fiscal expenditures.

Introduction

Implementation of the IP initiative required substantial resources. To provide an illustrative example for other districts and education agencies considering adopting similar reforms and wondering how much they might cost and how the costs might be distributed among elements of the reform (i.e., the "levers"), we analyzed the resources that the sites invested in the initiative, in terms of both direct expenditures and time spent by school-level staff. In addition, in evaluating the Gates Foundation's reform strategy, it is useful to examine the extent to which other funds matched the foundation's investment, as intended.

In service of these goals, we used accounting records provided by the sites to determine how much each site spent on the initiative, including the various funding sources from which the sites drew and how they distributed the expenditures among the different levers. We also analyzed survey data on the time that SLs and teachers allocated to various activities, including teacher evaluation, because time school-level staff spent on activities related to the initiative is an implicit cost of the initiative. Specifically, this chapter addresses six main questions:

- Between November 2009 and June 2016, how much did each site spend on its IP initiative, and what sources of funding did each site use?
- How large was the IP budget as a percentage of each site's overall budget in 2015–2016?
- How did per-pupil expenditures on the IP initiative change in each site over the seven years of the initiative?

- How much was spent in each of the following categories related to the IP initiative: (1) teacher evaluation (including measures of effective teaching); (2) staffing (including policies related to recruitment and hiring, placement and transfer, and tenure and dismissal); (3) PD (including evaluation-linked PD); and (4) compensation and CLs?
- How much time did SLs and teachers spend on evaluation-related activities?
- What was the dollar value of the time teachers and SLs spent on evaluation-related activities in 2014–2015?

After a brief review of the literature on the costs of reforms related to TE and staff time spent on teacher evaluation, we present the results of the expenditure analyses based on accounting records, followed by the results of the time allocation analysis based on survey data. The chapter concludes with a discussion of the value of the time that SLs and teachers invested in IP activities and how the resource estimates change when we add in the value of this time.

Evidence on Resources Needed to Implement the Best Practices Described in Previous Chapters

As education-related HC reform goes, the IP initiative was unusual in its scope, attending to multiple levers as described in the previous six chapters. To our knowledge, there is no research evidence about the total financial resources needed to implement such an initiative, at least in part because such a broad, multifaceted, and long-lasting HC reform initiative has not been tried.

There is, however, some information on the costs of implementing rigorous teacher-evaluation systems based on measurement of TE, the linchpin of the IP reform (see Chapters One and Three). This information provides at least some basis for comparison for some of our findings on the sites' expenditures. In addition, research has also attended to the ways in which teachers and SLs allocate their time, and studies that have included time spent on such activities as evaluation and mea-

surement of educator effectiveness provide evidence that can be compared with our results on the ways teachers and leaders allocated their time in the IP sites.

For the cost of effectiveness-based evaluation systems, existing theory-based simulations and empirical evidence indicate a wide range of costs. With funding from Institute of Education Sciences, the Carnegie Foundation for the Advancement of Teaching developed an online "cost calculator" to help school districts calculate the costs of implementing effectiveness-based teacher-evaluation systems (Thorn, 2013; Carnegie Foundation for the Advancement of Teaching, undated); this tool illustrates all the various factors that go into determining how much an evaluation system might cost. These include contextual factors, such as teacher and observer average salaries, as well as the design of the evaluation system itself. The online calculator requires the entry of data on district size and location, average salaries, contracted services (e.g., for calculation of value added), student assessments, formal observations (including the number of observations by different types of observers and the duration of observations), training, and communication.[1] As Chapter Three demonstrated, sites can vary quite widely on any or all of these factors and thus are likely to have quite divergent costs. There is really no "typical" cost.

There is also some case-study evidence of the cost of implementing evaluation systems in particular sites. For instance, in a study of the Cincinnati Public Schools, Taylor and Tyler estimated that the cost of conducting teacher observations (three observations per year conducted by peer evaluators, similar to HCPS's policy) averaged about $7,500 per teacher (Taylor and Tyler, 2012). If each teacher taught 25 students, that would translate to $300 per pupil. Ninety percent of the total cost was associated with the salaries of the peer evaluators. More broadly, Dynarski estimated that the U.S. public education system spends $1.4 billion per year, or about $450 per teacher (or, again

[1] As a further complication threatening comparability of costs from place to place, different districts might make different decisions about the extent to which some systems or policies, such as assessment of students, should be considered as part of the evaluation system or as business-as-usual costs that, because they are incurred even in the absence of the new teacher-evaluation system, should be excluded from the cost calculation.

on the assumption of 25 students per teacher, $18 per pupil), to observe teachers (Dynarski, 2016).[2] Note that Dynarski's estimate differs from Taylor and Tyler's by more than an order of magnitude, in part because the Cincinnati example involved the use of dedicated peer evaluators, while Dynarski's estimate assumed evaluation by principals.

Many districts seek to purchase already-developed evaluation system components, such as a classroom-observation rubric, and contract out for training on the use of the rubric. Different costs are associated with different developers and training providers. Often, cost is a factor in a site's determination of which system it opts to adopt, and it can cause controversy. For example, as Eger described, the state of Oklahoma was considering three different teacher-evaluation models for statewide adoption, including one homegrown by Tulsa Public Schools (TPS) (Eger, 2011). However, the state and TPS had widely differing estimates of how much the TPS system would cost, both in absolute terms and relative to the other models being considered (one of which was the FFT, which several of the IP sites used). Mooney, 2012, meanwhile, describes New Jersey school district representatives comparison-shopping for an evaluation system model, including consideration of costs.

Toch (2008) considers the costs of four teacher-evaluation models: (1) the Teacher Advancement Program (TAP), an evaluation model based on the FFT; (2) the Toledo, Ohio, PAR program for evaluation of first-year teachers; (3) Connecticut's Beginning Educator Support and Training program, also for novice teachers; and (4) the NBPTS, which confers advanced certification on teachers demonstrating excellence. Toch's findings, shown in Table 9.1, are that the per-teacher cost can range from $2,000 to more than $10,000. He does not provide per-pupil estimates for most of the programs, but, on the assumption of 25 students per teacher, the cost estimates translate to a range of $80 to $596 per pupil.

---

[2] Dynarski's estimate is based on an assumption of 3.1 million public school teachers of kindergarten through grade 12 (from 2015 data), ten hours per year spent by a principal to observe and provide feedback to each teacher, and an average salary of $45 per hour for principals.

Table 9.1
Costs of Four Evaluation Programs

Program	Per-Teacher Cost	Note
TAP	$6,250–$14,900	40% of cost is for performance bonuses.
Toledo PAR	$5,000	This provides mentoring and evaluation for 100 first-year teachers.
Connecticut's Beginning Educator Support and Training	Just over $2,000	40% of cost is for evaluating second-year teachers' portfolios; the other 60% is for training and supporting the 1,800 participating first-year teachers.
NBPTS	$2,500	This is the cost of filing the application.

SOURCE: Toch, 2008.

For achievement-based measures, too, much depends on the specific approach taken. An analysis by the Value-Added Research Center (VARC) at the University of Wisconsin–Madison (mentioned but not cited in Sawchuk, 2013), compared three different approaches: (1) expansion of the grades and subject areas in which commercially available standardized tests annually assess students; (2) in-state development of new assessments for nontested subjects and grades; and (3) implementation of SLOs. The VARC researchers concluded that, because of high indirect costs for training teachers and others on how to craft and score SLOs, the third option would be the most expensive. According to Sawchuk, the analysis also estimated that, for "a mid-size district with 50 schools or so, evaluation costs could easily cross the $1 million mark for the first year alone."

We take the view that teacher and SL time spent on activities related to the IP reform, even if not directly financially compensated as part of the reform budget per se, is another type of cost of the reform, and one that is important to consider. If teachers and SLs are spending time on reform-related activities (e.g., observation) at the expense of time spent on other activities that serve students' educational needs, the

sites could implicitly be shouldering the time-based costs; alternatively, if the time teachers and SLs are spending on reform-related activities is coming out of their personal time (e.g., sleep or time with family), the teachers and SLs would be shouldering the time-based costs. Either way, time-based costs, as opportunity costs, are costs that *someone* shoulders, so these costs must be considered a resource invested in the reform. We therefore look at other studies and sources of data on how much time school staff spend on such activities as teacher evaluation.

Many studies and reports have examined the amount or percentage of time principals spend evaluating or observing teachers. As with the costs of evaluation systems, there is considerable variation in the findings, reflecting differences in method, definitions (e.g., what activities are included as part of evaluation), the requirements of the evaluation system, and setting (e.g., principal-to-teacher ratio, school level). Among the findings are the following:

- Sporte and her colleagues examined principals' time allocation as part of a teacher-evaluation program in Chicago (Sporte et al., 2013). They found that principals spent an average of six hours per teacher per evaluation, totaling 120 hours per year on average for elementary school principals and 168 hours per year for HS principals, translating to 7 to 9 percent of principals' total time.
- Rowan, Schilling, and their colleagues found that, in Michigan, principals spent a median of 248 hours per year on teacher-evaluation activities (including observation), translating to 16 percent of total time (Rowan, Schilling, et al., 2013).
- A report issued jointly by the National Association of Secondary School Principals and National Association of Elementary School Principals cited a 2013 survey of their members finding that "a substantive teacher evaluation requires 11–15 hours per teacher over the course of a school year" (National Association of Secondary School Principals and National Association of Elementary School Principals, undated, p. 4).
- In a longitudinal study of about 100 principals in Miami–Dade County Public Schools, Grissom, Loeb, and Master found that, in 2012 (the most recent of three studied years), elementary and

MS principals spent about 2 percent of their time evaluating teachers and HS principals spent about 1 percent (Grissom, Loeb, and Master, 2013).

- Sebastian, Camburn, and Spillane, using data from a log-based study of 52 schools in an urban district during the mid-2000s (2005–2007), found that principals allotted about 10 percent of their time to personnel issues, defined as "recruiting, hiring, supervising, evaluating, problem solving" (Sebastian, Camburn, and Spillane, 2018, p. 11).
- Larkin and Oluwole estimated the effects that the state of New Jersey's 2012 enactment of rigorous teacher-evaluation requirements would have on the amount of time school administrators spend solely on teacher observations (not including pre- and post-conferences, write-ups, or scheduling) (Larkin and Oluwole, 2014). They found that, under the previous evaluation system's requirements, administrators spent a median of 12.5 hours per teacher per year on observations, while, under the new teacher-evaluation system, they would spend 17 hours. Their estimates of median time, however, varied substantially among school districts, likely as a result of differing administrator-to-teacher ratios.

Less studied is the amount of time teachers themselves spend on evaluation-related activities. According to a survey by the Network for Public Education, 75 percent of teachers reported spending at least four hours per month on activities relating to evaluation of their teaching, with 27 percent spending eight or nine hours per month (Network for Public Education, 2016). No other studies on this topic were readily available to us. We thus believe that our own findings will be an important contribution to the field.

Total Spending on the IP Initiative

To determine each site's spending on the IP initiative, we examined the financial report that each site submitted to the Gates Foundation each year. To compare expenditures across sites, we calculated per-pupil

values by dividing each site's expenditures by the latest student enrollment figure that the site provided. We classified expenditures into four main funding sources: Gates Foundation grant dollars, federal funds (e.g., SEED, TIF, Title I), funds from state and local tax revenues, and funds from other philanthropic sources. In addition, we worked with the sites to classify each reported expenditure into the four implementation lever categories: teacher evaluation, staffing (recruitment and hiring, placement and transfer, and tenure and dismissal), PD, and compensation and CLs.[3] For further detail on methods, see Appendix L.

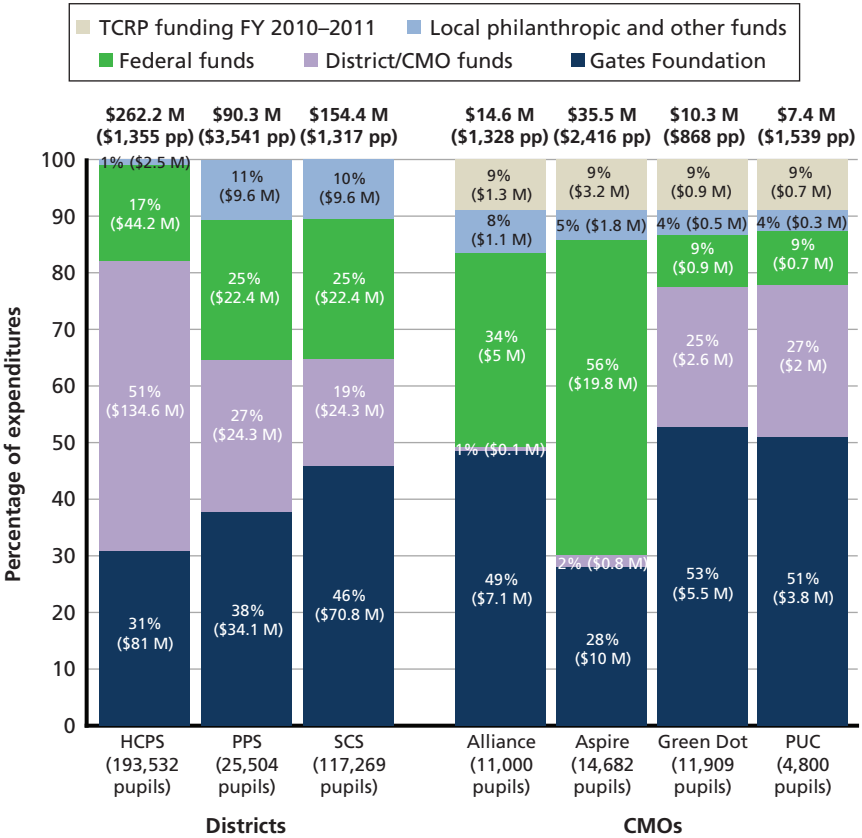
Across all seven sites, spending on IP-related activities totaled $574.7 million between November 2009 and June 2016. Figure 9.1 shows the overall expenditures for each IP site, by funding source. HCPS, the largest site, had the largest overall expenditures, $262.2 million, followed by SCS, which spent $154.4 million. Among the CMOs, Aspire, which had the highest enrollment, also had the largest expenditures with $35.5 million, followed by Alliance with $14.6 million. Because total costs might be largely a function of site size, we calculated costs on a per-pupil basis to make cross-site comparisons more meaningful and to provide a metric that others interested in adopting similar reforms might find more useful.

Overall per-pupil expenditures varied widely across the seven sites. HCPS, SCS, Alliance, and PUC spent between $1,317 and $1,539 per pupil. Aspire and PPS spent about double those amounts—$2,416 and $3,541, respectively. Green Dot had the lowest per-pupil expenditure: $868. This variation could be related to several factors, such as the specific teacher-evaluation and compensation policies that each site enacted, the number of students enrolled in each site, and size and composition of the teacher workforce (e.g., how much experience teachers tended to have). We discuss each of these factors in turn.

Although all the sites designed and implemented a multiple-measure teacher-evaluation system, some invested more on particular

[3] In cases in which expenditures supported more than one implementation lever, we worked with the sites to allocate the expenditures according to the proportion of the activities that those funds covered.

Figure 9.1
Overall Expenditures, by Funding Source, November 2009–June 2016



SOURCES: IP sites' financial reports for the fall of 2014 and springs of 2015 and 2016. See Table L.1 in Appendix L.
NOTE: FY = fiscal year. M = millions. pp = per pupil. Detailed financial reports were not available for the CMOs prior to FY 2012. For FY 2010 and FY 2011, when the CMOs were organized collectively as TCRP (see Chapter Two), we estimated each CMO's funding by prorating the total TCRP funding in those years by each CMO's share of the four CMOs' combined funding for FY 2012–2014. Enrollment figures next to site names are from 2015–2016 and come from the sites' stocktakes, data dashboards, and (PPS only) general fund budget. We base per-pupil expenditures on these 2015–2016 enrollment numbers. Because of rounding, category percentages might not sum to 100; similarly, because of rounding, category dollar amounts might not sum to the total amounts at the top.

RAND *RR2242-9.1*

aspects, such as IT systems, training of observers, and development of rubrics. For example, PPS developed both an IT system for the teacher-evaluation system and the evaluation rubric. HCPS, meanwhile, already had a teacher-evaluation rubric but had to allocate substantial resources for the peer observers. In addition, different sites made different modifications to their compensation systems, with some, for instance, offering higher effectiveness-based bonuses than others or some offering bonuses to a higher percentage of teachers (see Chapter Seven).

The size of each site also helps explain the variation in per-pupil expenditure. For example, HCPS and SCS were the two largest sites, each with more than 100,000 students, so some costs, such as developing the evaluation rubric or an IT system, were lower on a per-pupil basis, reflecting economies of scale. PPS and the CMOs were far smaller than HCPS and SCS.

Composition of the teacher force, too, might have affected expenditure levels. For example, PPS had a highly experienced teacher workforce, who, at least partly as a function of their experience, earned relatively high base salaries, and therefore proportionally higher effectiveness-based bonuses. In contrast, the CMOs had a higher proportion of teachers with less experience, so they had lower expenditures related to effectiveness-based bonuses and salary increments. On the other hand, inexperienced teachers might have been observed more times than experienced teachers and thereby incurred greater evaluation-related costs.

Grant funds from the Gates Foundation accounted for 37 percent of total IP spending (about $212 million). Depending on the site, foundation funds accounted for 28 to 53 percent of the overall expenditures. As shown in Figure 9.1, Aspire and HCPS had the lowest share of expenditures covered by foundation funds at 28 percent and 31 percent, respectively. Aspire relied more on federal funding (56 percent), while HCPS relied more on district funding (51 percent). Conversely, Green Dot and PUC had the highest shares of expenditures covered by foundation funds, with 53 percent and 51 percent, respectively. These two sites relied on CMO funds for another quarter of their expenditures.

Federal funding covered 17 to 25 percent of overall expenditures in each of the three districts and 9 to 56 percent in the CMOs. Green Dot and PUC had the lowest shares of their expenses covered by federal funding (9 percent of total expenditures for each). HCPS had the second-lowest proportion (17 percent) of its expenditures funded by federal money. The federal funds on which HCPS drew included RTT (which covered less than 10 percent of the district's total expenditures), TIF, SEED, and SIG funding. SCS, for which 25 percent of total IP-related funds came from the federal government, used RTT funds exclusively. PPS had both TIF and SIG funding. For all the CMOs except PUC, the federal funding was exclusively from TIF. PUC had a combination of TIF and Title II and IV funding. The six sites that drew on TIF funds (i.e., every site except SCS) varied in their reliance on these funds for the IP initiative. On the low end, HCPS covered about 6 percent of IP expenditures with TIF funds. On the high end, Aspire covered 56 percent of its IP expenditures with its TIF grant.

The proportion of total funding covered by the sites' own funds also varied widely across the sites, ranging from less than 5 percent in Alliance and Aspire to up to 51 percent in HCPS. The other four sites all relied on their own funds for 20 to 30 percent of their overall IP expenses.

It is not clear why there was so much variation across sites in how they funded their IP initiative. Among the possibilities are that they had differential access to other sources of funding (e.g., from government grants or local philanthropic organizations), had different degrees of flexibility to reallocate existing funding, and had different approaches to resource management and acquisition.

Proportion of Sites' Total Budgets Spent on the IP Initiative in 2015–2016

To put the cost estimates into perspective, we computed the percentage of each site's overall budget devoted to the IP initiative in 2015–2016. We chose to focus on 2015–2016 because, as the last year of the grant for most of the sites, it provides the best indicator of how much it

costs sites to sustain their reforms annually, following completion of up-front development work, which likely requires greater investment but is not a continuing expense.

In 2015–2016, the three districts each spent 2 to 3 percent of their total budget on IP-related activities. There was wider variability among the CMOs: 1.7 percent to 6.6 percent of total budget. As shown in Table 9.2, PPS spent a larger proportion of its total expenditures on the IP initiative than HCPS and SCS did. The difference, however, was not huge: close to 3 percent for PPS and closer to 2 percent for HCPS and SCS. Three of the CMOs spent a higher proportion of their total budget on the initiative: 3.4 percent for both Alliance and

Table 9.2
Proportion of Districts' Total Budgets Spent on the Initiative, 2015–2016

Spending Area	HCPS	PPS	SCS
Enrollment	193,532	25,504	117,269
Number of teachers in 2014–2015[a]	13,603	1,690	5,780
Overall budget, in dollars			
Total	2,832,043,894	539,610,000	1,347,109,782
Per pupil	14,633	21,158	11,487
Per teacher	208,193	319,296	233,064
IP initiative, in dollars			
IP expenditure	59,980,085	15,043,196	31,109,596
Per pupil	310	590	265
Per teacher	4,409	8,901	5,382
Percentage of total budget spent on the IP initiative	**2.12**	**2.79**	**2.31**

SOURCES: HCPS, PPS, and SCS financial reports for the spring of 2016; sites' budgets for 2015–2016. See Table L.1 in Appendix L. For number of teachers in 2014–2015, NCES, undated (b).

[a] The number of teachers includes only teachers in operational noncharter schools having nonmissing enrollment and serving at least one grade level from kindergarten through grade 12, in 2014–2015 (the most recent year for which data were available).

PUC and about 6.6 percent in Aspire (see Table 9.3). Green Dot had the lowest proportion of all seven sites: just 1.7 percent.

To illustrate the differences in absolute terms across IP sites, we also present the per-pupil and per-teacher IP expenditures for the 2015–2016 academic year. Of the three districts, the per-pupil IP expenditures ranged from $265 in SCS to $590 in PPS (see Table 9.2); PPS also had the highest per-teacher expenditures ($8,901), but HCPS had the lowest ($4,409). For the CMOs, the IP expenditures ranged between $212 per pupil (and $4,933 per teacher) in Green Dot and $675 per pupil (and $16,568 per teacher) in Aspire (see Table 9.3).

Table 9.3
Proportion of CMOs' Total Budgets Spent on the Initiative, 2015–2016

Spending Area	Alliance	Aspire	Green Dot	PUC
Enrollment	11,000	14,682	11,909	4,800
Number of teachers in 2014–2015	615	598	511	224
Overall budget, in dollars				
Total	161,931,516	151,178,778	152,000,000	57,514,550
Per pupil	14,721	10,297	12,763	11,982
Per teacher	263,303	252,807	297,456	256,761
IP initiative, in dollars				
IP expenditure	5,534,643	9,907,785	2,521,013	1,964,870
Per pupil	503	675	212	409
Per teacher	8,999	16,568	4,933	8,772
Percentage of total budget spent on the IP initiative	**3.42**	**6.55**	**1.66**	**3.42**

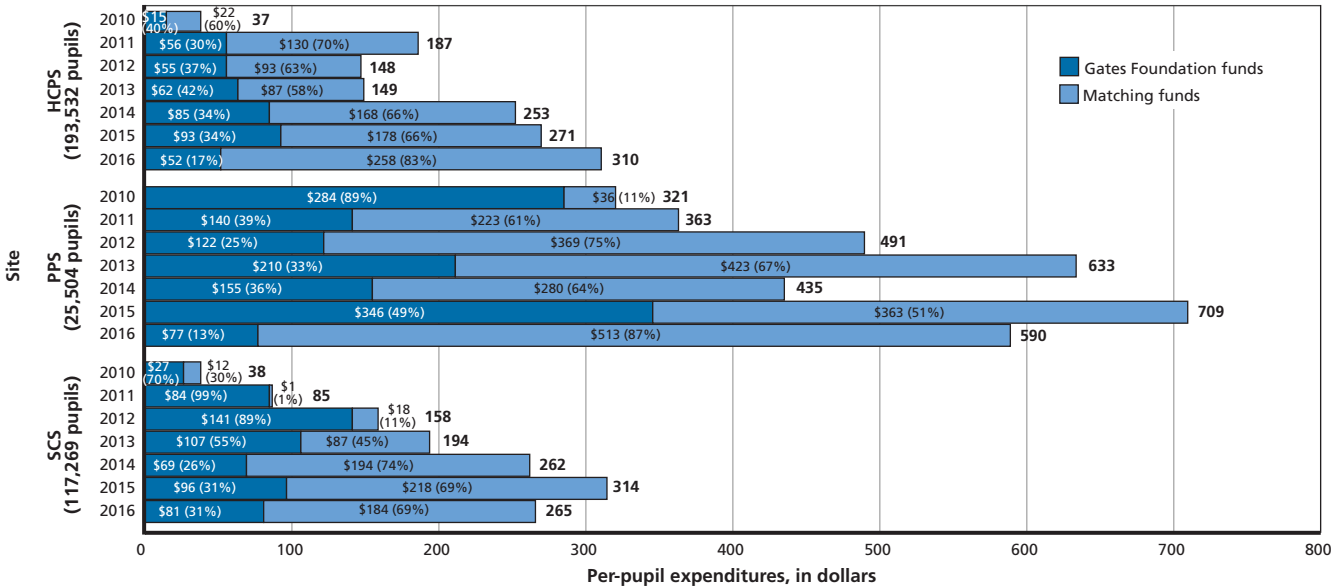
SOURCES: Alliance, Aspire, and Green Dot financial reports for the spring of 2015; PUC financial report for the fall of 2014; sites' budgets for 2015–2016. For the number of teachers in 2014–2015, NCES, undated (b).

Change over Time in Funding and Funding Source

Most sites doubled, tripled, or even quadrupled their per-pupil expenditures on the IP initiative between 2010–2011, the first full year of the initiative, and 2015–2016. As shown in Figures 9.2 (for the three districts) and 9.3 (for the four CMOs), HCPS and PPS almost doubled their per-pupil IP expenditures over this time period; SCS and Green Dot about tripled theirs; and the other three CMOs about quadrupled their per-pupil spending on IP efforts. Aspire had the largest increase in its IP expenditures, rising from $159 per pupil in 2011–2012 to $675 in 2015–2016. In PPS, although the per-pupil expenditures grew over the course of the initiative, the growth was not as large as in most of the other sites; this is because, in the initial year, PPS's per-pupil expenditures were quite high compared with those in the other sites. PPS's first-year expenditures were high because of investment in IT systems, linking together various databases that had previously been unconnected, used in the new teacher-evaluation system.

As planned, the share of expenditures covered by matching funds grew over the years, and the share covered by Gates Foundation funds declined. As the initiative matured, the sites generally drew less and less on Gates Foundation funding; this was according to plan. In HCPS, the foundation provided 40 percent of the district's (relatively low) initiative funding in the initial year (2009–2010), and the Gates Foundation–funded proportion remained at 30 to 40 percent over the next five years before dreopping to 17 percent in the final year (see Figure 9.2). PPS began the initiative in 2010 with a high proportion of its expenditures (89 percent) covered by Gates Foundation funds. Over the subsequent years, PPS drew less on Gates Foundation funds than on other funding sources, until 2014–2015, when the proportion of funding covered by the foundation rose again to nearly half. Then in 2015–2016, the foundation's share of PPS's expenditures fell dramatically, to 13 percent—the lowest percentage for any of the three districts over these seven years. In SCS, the portion of expenses covered by the Gates Foundation was high in the first three years (peaking at

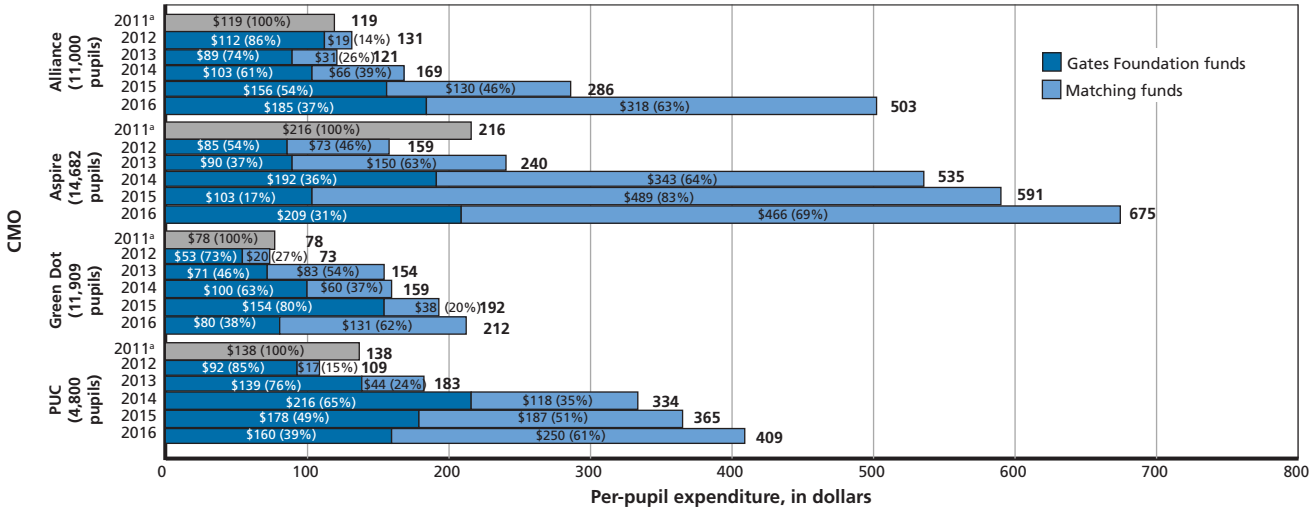
Figure 9.2
Districts' Per-Pupil Expenditures on the Initiative: Gates Foundation Funding Versus Matching Funds, November 2009–June 2016



SOURCES: IP site financial reports for the fall of 2014 and the springs of 2015 and 2016.
NOTE: Years shown indicate the spring of the school year (for example, *2010* indicates the 2009–2010 school year). Enrollment figures next to site names are from 2015–2016 and come from the sites' stocktakes, data dashboards, and (PPS only) general fund budget. We base per-pupil expenditures on these 2015–2016 enrollment numbers.

RAND *RR2242-9.2*

Figure 9.3
CMOs' Per-Pupil Expenditures on the Initiative: Gates Foundation Funding Versus Matching Funds, November 2009–June 2016



SOURCES: IP site financial reports for the fall of 2014 and the springs of 2015 and 2016.
NOTE: Years shown indicate the spring of the school year (for example, 2010 indicates the 2009–2010 school year). Enrollment figures next to site names are from 2015–2016 and come from the sites' stocktakes and data dashboards. We base per-pupil expenditures on these 2015–2016 enrollment numbers.
[a] Combined funding for FY 2010 and FY 2011. Detailed financial reports were not available for the CMOs prior to FY 2012. For FY 2010 and FY 2011, when the CMOs were organized collectively as TCRP, we estimated each CMO's funding by prorating the total TCRP funding in those years by each CMO's share of the four CMOs' combined funding in FYs 2012–2014.

RAND *RR2242-9.3*

nearly 100 percent in 2010–2011), before falling to about half in 2012–2013 and then further to a quarter or third in the final three years.

From 2011–2012 through 2015–2016,[4] all the CMOs increased the proportion of expenditures covered by matching funds. In both Alliance and PUC, the percentage of per-pupil expenditures covered by Gates Foundation funding dropped steadily over the five years, beginning in 2011–2012 at about 85 percent and ending in 2015–2016 at slightly less than 40 percent. The pattern for Aspire was similar, although the decrease in the Gates Foundation's share (from 54 percent in the first year to 31 percent in the final year) was not as large, and the pattern was interrupted in 2014–2015 when the foundation's proportion hit a low of 17 percent. In Green Dot, the Gates Foundation's share fluctuated over the five years, falling, rising, and then falling again—with the final change, from 80 percent in 2014–2015 to 38 percent the following year, being the largest year-to-year change.

Spending by Implementation Lever

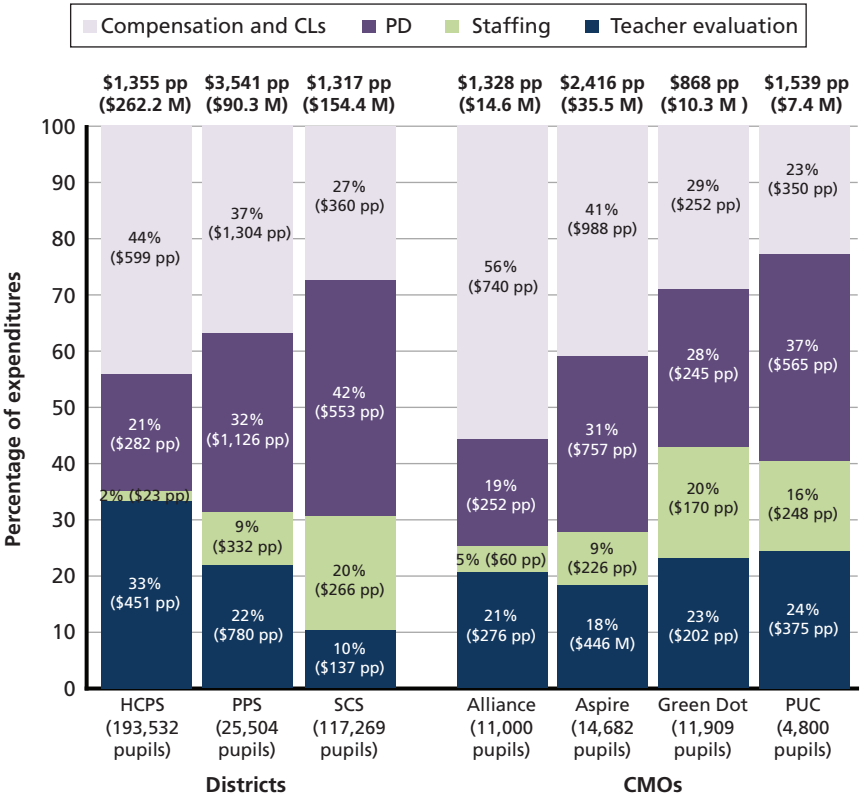
Five of the seven sites spent the largest share of their IP expenditures on compensation and CL activities. The staffing lever typically received the smallest share of spending. Figure 9.4 shows the allocation of per-pupil expenditures across the main implementation levers during the whole grant period (November 2009 through June 2016). The levers are defined as follows: (1) teacher evaluation, (2) staffing, (3) PD, and (4) compensation and CLs.[5] As Figure 9.4 shows, sites differed both in how much they spent on each lever and, perhaps relatedly, in how they allocated funds across the levers.

---

[4] Because the CMOs did not track their expenditures by source the first two years of the initiative, we could determine their expenditures by funding source starting only in 2011–2012.

[5] *Staffing* refers to activities related to recruitment and hiring, placement and transfer, and tenure and dismissal. *Compensation* refers to additional amounts of compensation staff received as a result of the initiative (e.g., effectiveness-based bonuses and salary increments); it does not include base salaries.

Figure 9.4
Overall Proportion and Total Per-Pupil Expenditures, by Implementation Lever, November 2009–June 2016



SOURCES: IP site financial reports for the fall of 2014 and the springs of 2015 and 2016. See Table L.1 in Appendix L.
NOTE: M = millions. pp = per pupil. Detailed financial reports were not available for the CMOs prior to FY 2012. For FY 2010 and FY 2011, when the CMOs were organized collectively as TCRP, we estimated each CMO's funding by prorating the total TCRP funding in those years by each CMO's share of all CMO funding for FYs 2012–2014 and apportioned it across the different levers. Enrollment figures next to site names are from 2015–2016 and come from the sites' stocktakes, data dashboards, and (PPS only) general fund budget. We base per-pupil expenditures on these 2015–2016 enrollment numbers. Because of rounding, category percentages might not sum to 100; similarly, because of rounding, category dollar amounts might not sum to the total amounts at the top.

RAND *RR2242-9.4*

For the teacher-evaluation lever (the bottom part of each bar), per-pupil expenditures ranged from $137 in SCS to $780 in PPS, with a cross-site average of $381. As a proportion of spending across all four levers, expenditures for evaluation ranged from 10 percent in SCS to 33 percent in HCPS. HCPS's use of peer evaluators was a major expense, accounting for nearly one-fifth of the site's total IP expenditures. SCS spent proportionally less on evaluation than the rest of the sites; one reason could be that it used the state's measure of value added, as well as the state's weighting of TE components and performance levels, so it did not have to pay for the development and calculation of these elements of the evaluation system.

In all the sites except SCS, the staffing lever received less funding than the other levers. Expenditures ranged from $23 per pupil in HCPS to $332 per pupil in PPS (averaging $189 across the seven sites), and percentage of total expenditures ranged from 2 percent in HCPS to 20 percent in SCS and Green Dot. SCS spent more per pupil on the staffing lever ($266) than it did on the evaluation lever ($137). SCS's partnership with TNTP helps explain the site's relatively high proportional spending on the staffing lever.

For the PD lever, per-pupil expenditures ranged from $245 in Green Dot to $1,126 per pupil in PPS. As a percentage of total spending, PD accounted for as low as 19 percent in Alliance and as high as 42 percent in SCS, which put a lot of resources into coaching, particularly after the merger between legacy MCS and legacy SCS (see Appendix F for details about PD and coaching in SCS). In two sites—PUC (37 percent) and SCS (42 percent)—the PD lever had the largest share of total IP expenditures. Like SCS, PUC also put considerable resources into coaching.

Finally, per-pupil expenditures on the compensation and CL lever (top part of each bar) ranged from $252 in Green Dot to $1,304 in PPS. This lever accounted for 23 percent (in PUC) to 56 percent (in Alliance) of sites' overall IP expenditures. In five of the seven sites (HCPS, PPS, Alliance, Aspire, and Green Dot), expenditures related to compensation and CLs received the largest share of the total IP expenditures. Alliance spent more than half of its total expenditures on this lever; no other site spent even as much as half on a single lever. As

noted in Chapter Seven, Alliance was one of only two sites (the other being Aspire) that implemented an effectiveness-based salary schedule for all teachers.

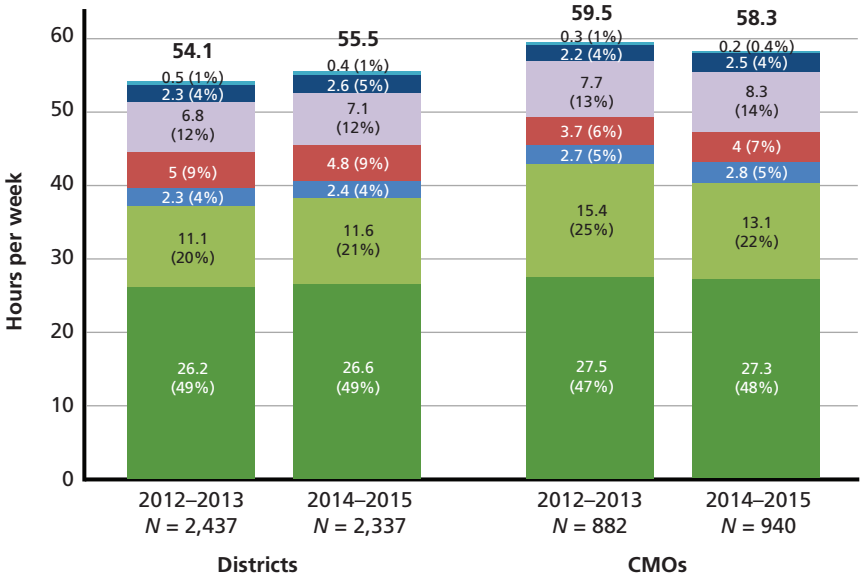
Teachers' and SLs' Allocation of Time

The expenditures we reported based on financial data are only one part of the story of the resources needed to implement the IP initiative. To provide a more complete picture, it is important to take into account the amount of time that teachers and SLs spent in IP-related activities, such as planning, training, observing, and guiding staff. Even though most teachers and SLs were salaried and not paid on an hourly basis—so the sites did not have to pay them for additional time they spent on initiative activities—time spent on such activities might have been at the expense of other activities on which personnel could have spent their time. Thus, we must consider time spent on initiative activities in terms of opportunity costs; we can also think of staff time as an additional resource invested in the initiative. This section explores how much time teachers and SLs reported spending on different activities, including those related to the IP initiative. These data were reported on the teacher and SL surveys discussed elsewhere in this report. In particular, we used data from the teacher and SL surveys administered in the springs of 2013 and 2015.[6] We report overall time and time spent on individual activities, such as administration, instruction, evaluation, PD, and reform. (See Appendix L for more information about the time allocation methods.)

Figure 9.5 shows how teachers said they allocated their time (in hours) across different activities in a typical week during the 2012–

[6] The teacher surveys administered in 2014 and 2016 did not include the detailed questions on time allocation used for our analyses, and the surveys administered in 2011 used a different structure for the time allocation section, as well as different questions, that made comparisons with later years challenging. See Appendix L for further detail on the selection of years of data for the time allocation analysis.

Figure 9.5
Teachers' Overall Time Allocation, 2012–2013 and 2014–2015



Activity	Districts	CMOs
Total weekly hours	+	–
Classroom instruction		
Instructional planning	+	–
Administration		
Contact with students and families		+
PD		+
Evaluation	+	+
Reform		–

SOURCES: SL surveys conducted in the springs of 2013 and 2015.
NOTE: + denotes that weekly hours in 2014–2015 were statistically significantly more than weekly hours in 2012–2013 ($p < 0.05$). – denotes that weekly hours in 2014–2015 were statistically significantly less than weekly hours in 2012–2013 ($p < 0.05$).

RAND *RR2242-9.5*

2013 and 2014–2015 school years.[7] It also shows whether differences between the two years were statistically significant.[8] Figure 9.6 presents the corresponding results for SL time,[9] and Figures 9.7 and 9.8 disaggregate the SL results by leadership position (principals versus APs), first for the districts (Figure 9.7) and then for the CMOs (Figure 9.8). For ease of presentation and interpretation, we average results (1) across the three districts and (2) across the four CMOs.[10] In Appendix M, we provide time allocations by individual site.

Teachers' Time

On average, teachers in the three districts reported working 55.5 hours per week in 2014–2015, a slight increase (1.4 hours) from 2012–

---

[7] Teacher activities fell into seven main categories: (1) classroom instruction (directly teaching and assessing student progress, including classroom teaching during and outside the regular school day); (2) instructional planning (planning and reviewing student work and data); (3) administration (attending meetings, supervising other staff, and similar activities); (4) contact with students and families (dealing with disciplinary issues, monitoring detention or study hall, sponsoring or coaching after-school activities, and meeting with parents); (5) PD (activities related to participating in workshops or training as part of district- or school-wide PD, as well as participating in learning communities or other collaborative activities); (6) evaluation (preparing for one's own evaluation and, for those who reported serving as formal evaluators or mentors, formally evaluating or mentoring other teachers and receiving training to be a mentor or a coach); and (7) reform (participating in activities related to TE reforms, such as participating in meetings and responding to surveys, as well as other district and CMO reform activities not related to TE).

[8] We conducted t-tests to calculate whether differences were statistically significant ($p < 0.05$).

[9] SL activities consisted of seven categories: (1) administration: general administration activities (e.g., management, meetings); (2) classroom instruction: teaching classes, only for SLs who reported having official teaching responsibilities; (3) evaluation: activities related to the formal evaluation of teachers; (4) receiving PD: participating in PD; (5) providing PD: leading PD for teachers and nonteaching staff; (6) recruitment: hiring of teachers and support staff; and (7) reform: other activities related to TE reforms (e.g., meetings, responding to surveys, and communicating or discussing policies related to TE reforms), as well as other district and CMO reform activities not related to TE.

[10] Each average is the average of the component sites (i.e., the three districts for the district averages and the four CMOs for the CMO averages); we did *not* calculate the average by pooling all respondents together and ignoring site. In other words, each site contributes equally to the average for its set; sites with larger numbers of respondents do not overinfluence the average.

2013 (see Figure 9.5). The corresponding figure for the CMOs was 58.3 hours, a slight decrease (1.2 hours) from 2012–2013. Both of these changes—the increase for the districts and the decrease for the CMOs—were statistically significant. In both years, and in both the districts and the CMOs, the largest portion of teachers' time—about one-half of their total time—was spent on classroom instruction. The second-largest portion was instructional planning, and the third largest was PD.[11]

From the 2012–2013 school year to the 2014–2015 school year, teachers spent an increasing amount of time on evaluation-related activities. In 2012–2013, teachers in the districts spent an average of 2.3 hours on evaluation-related activities, and teachers in the CMOs spent an average of 2.2 hours.[12] In 2014–2015, the average amount of time spent on evaluation-related activities was 2.6 hours in the districts and 2.5 hours in the CMOs. Thus, the amount of time spent on evaluation increased by about 0.3 hours both in the districts and in the CMOs; for both groups of sites, the increase was statistically significant. In the districts, the increase in time spent on evaluation-related activities appears to have come from the overall increase in weekly hours, which was statistically significant; none of the other individual categories had a statistically significant decrease. In the CMOs, the increase in time spent on evaluation was accompanied by a significant reduction in time spent on instructional planning and on activities related to reform. In both years, the average amount of time teachers allocated to evaluation was very similar across the seven sites.

Our estimates, particularly for 2015, are higher than those that the Network for Public Education found (Network for Public Educa-

---

[11] See Table M.1 in Appendix M for teacher time allocation by site.

[12] We defined *evaluation-related activities* as activities related to evaluations of teachers' own performance, attending training to conduct observations of other teachers, preparing for and conducting observations of other teachers, and working with other staff as a mentor. Less than 10 percent of teachers in either year reported any time spent on activities related to evaluating other teachers, which is consistent with our understanding that very few teachers with active teaching responsibilities evaluated other teachers, in any of the seven sites. Slightly more than half of teachers reported time spent mentoring or coaching other teachers. Almost all teachers reported spending time on the evaluation of their own teaching.

tion, 2016). It found that most teachers spent four to nine hours per month on activities related to evaluation of their teaching. We, on the other hand, found that, in 2015, teachers in the IP sites spent about 2.5 hours per week on evaluation-related activities, which translates to 11 hours per month. However, our definition of *evaluation-related activities* might have been somewhat broader than the definition used in the Network for Public Education study.

SLs' Time

On average, SLs in the three districts worked about 58 hours per week in both years (see Figure 9.6). The SLs in the four CMOs, meanwhile, worked 61.5 hours per week in 2014–2015, a small but significant increase (1.4 hours) from 2012–2013. In both years, and in both the districts and the CMOs, the largest portion of SLs' time was spent on administrative tasks, followed by time spent on evaluation. The third-largest portion of time was spent receiving PD (see Figure 9.6).[13]

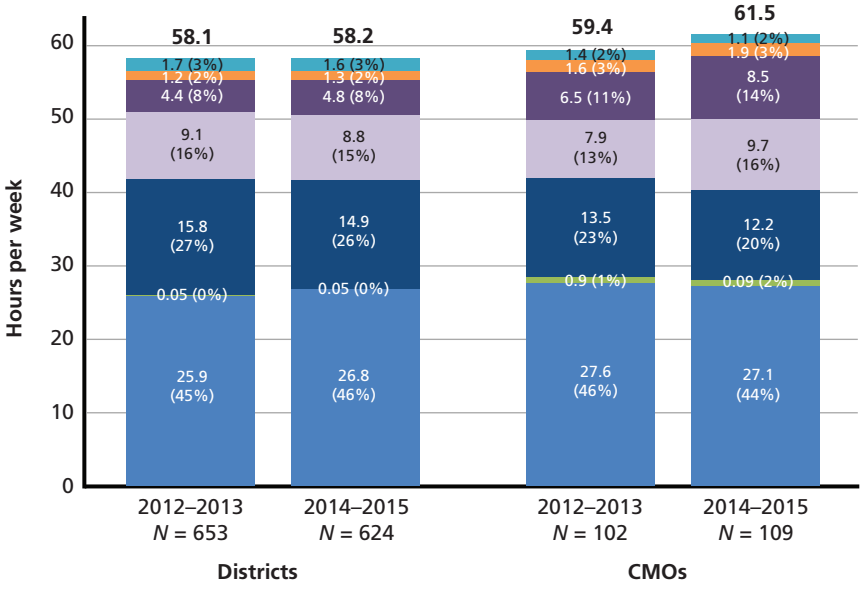
On average, across the seven sites, SLs spent slightly less than half of their time on administrative tasks. District SLs' time allocated to administrative-related tasks increased from 25.9 hours (45 percent of their time) in 2012–2013 to 26.8 hours (46 percent of their time) in 2014–2015. This increase was statistically significant. On the other hand, CMO SLs' time allocated to administration decreased slightly, from 27.6 hours (46 percent of time) to 27.1 hours (44 percent of time)—also a statistically significant change.

The time that district and CMO SLs reported spending on evaluation-related activities declined by about one hour from 2012–2013 to 2014–2015. Evaluation-related activities constituted SLs second-largest share of time, at about one-fourth or one-fifth of their total hours. In both the districts and the CMOs, the amount of time spent on evaluation declined slightly from 2012–2013 to 2014–2015, from about 16 to 15 hours in the districts and from 13.5 to 12.2 hours in the CMOs. This decrease is consistent with an overall move toward fewer evaluation hours starting in the middle years of the IP initiative.

---

[13] See Table M.2 in Appendix M for SL time allocation by site.

Figure 9.6
SLs' Overall Time Allocation, 2012–2013 and 2014–2015



Activity	Districts	CMOs
Total weekly hours		+
Administration	+	–
Classroom instruction		
Evaluation	–	–
PD received		+
PD provided		+
Recruitment		
Reform		–

SOURCES: SL surveys conducted in the springs of 2013 and 2015.
NOTE: + denotes that weekly hours in 2014–2015 were statistically significantly more than weekly hours in 2012–2013 ($p < 0.05$). – denotes that weekly hours in 2014–2015 were statistically significantly less than weekly hours in 2012–2013 ($p < 0.05$).
RAND *RR2242-9.6*

Principals Versus APs

As part of the IP initiative, the sites required much more classroom observation, including the provision of feedback to teachers, than they had previously. We were interested in examining whether APs started to assume some of the tasks previously done by principals to allow

principals to devote more time to teacher evaluation, as well as whether APs took on some of the evaluation responsibilities themselves. Thus, we disaggregated the SL time allocation data by leadership position: principals versus APs.[14]

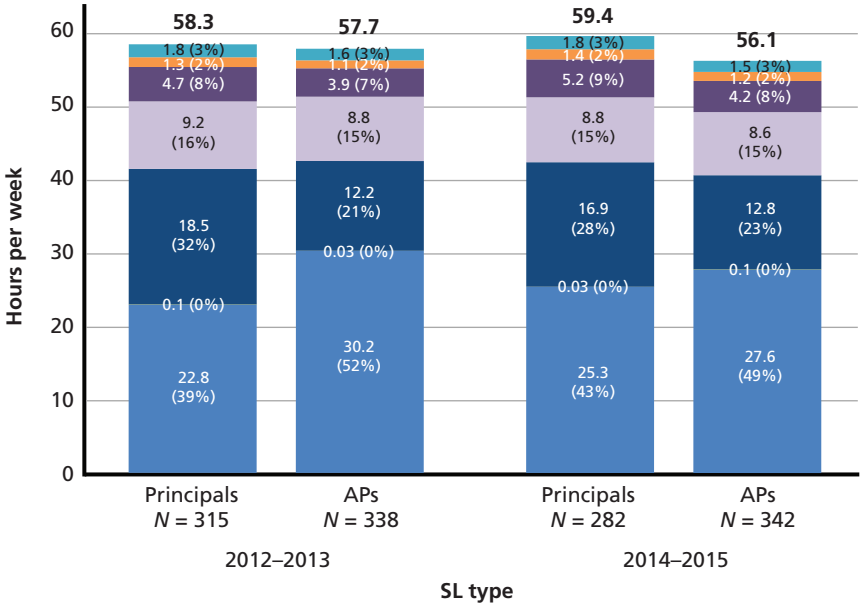
On average across the three districts, principals allocated more time than APs to evaluation and PD-related activities and less time to administrative tasks. This was the case in both 2012–2013 and 2014–2015. As shown in Figure 9.7, the differences between the two groups in the administration and evaluation categories were substantial and statistically significant, although larger in the earlier year than in the later year. In 2014–2015, principals allocated four more hours per week to evaluation (16.9 hours) than APs did (12.8 hours); in 2012–2013, principals allocated 18.5 hours to evaluation, while APs allocated just 12.2. For time spent on administration, APs allocated eight more hours per week than principals in 2012–2013 but only about two hours more in 2014–2015.

In the CMOs, principals allocated significantly more time to evaluation than APs did in both years, but the difference was larger in 2014–2015. On average, CMO principals spent 13 hours per week on evaluation in 2012–2013 and 14.2 in 2014–2015; APs spent 12.5 hours per week in the earlier year and 10.6 in the later year (see Figure 9.8). For time spent on administration, principals significantly decreased their hours per week from 29.6 in 2012–2013 to 27.2 in 2014–2015; APs' time on administrative activities remained about the same (27.6 hours per week in 2012–2013 and 27.4 in 2014–2015). In neither year was the difference between principals and APs in administrative time statistically significant.

That principals spent more time than APs on evaluation activities, in both the districts and the CMOs, is not surprising. As a consequence of the IP initiative, principals' focus shifted toward evaluating teachers, providing PD, and providing support to improve the quality of instruction. To help principals make time for these new activities, APs took on some of the administrative tasks that had previously been principals' responsibilities. Moreover, in some sites, central-office staff

[14] See Table M.3 in Appendix M for the principal and AP time allocations by site.

Figure 9.7
District Principals' and APs' Time Allocation, 2012–2013 and 2014–2015

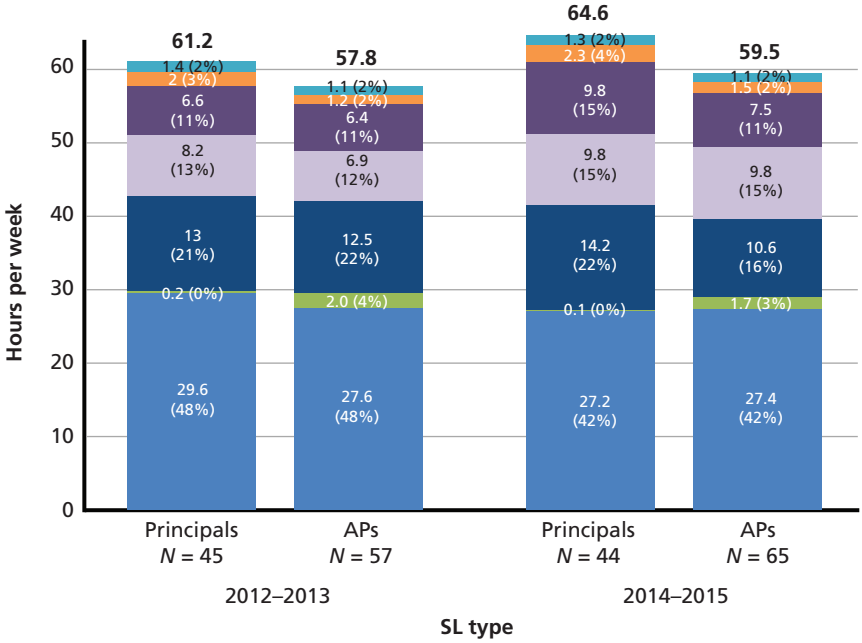


Activity	Differences Between Principals and APs		Differences Between Years	
			Principals	APs
	2012–2013	2014–2015	2012–2013 to 2014–2015	2012–2013 to 2014–2015
Total weekly hours	+	–	+	
Administration	–	–	+	–
Classroom instruction	+			
Evaluation	+	+	–	
PD received			–	
PD provided	+	+		
Recruitment	+			
Reform	+			

SOURCES: SL surveys conducted in the springs of 2013 and 2015.
NOTE: In the first two columns, + denotes that weekly hours for principals were statistically significantly higher than weekly hours for APs ($p < 0.05$), and – denotes that weekly hours for principals were statistically significantly lower than weekly hours for APs ($p < 0.05$). In the second two columns, + denotes that weekly hours in 2014–2015 were statistically significantly higher than weekly hours in 2012–2013 ($p < 0.05$), and – denotes that weekly hours in 2014– 2015 were statistically significantly lower than weekly hours in 2012–2013 ($p < 0.05$).

RAND *RR2242-9.7*

Figure 9.8
CMO Principals' and APs' Time Allocation, 2012–2013 and 2014–2015



Activity	Differences Between Principals and APs		Differences Between Years	
			Principals	APs
	2012–2013	2014–2015	2012–2013 to 2014–2015	2012–2013 to 2014–2015
Total weekly hours	+	+	+	
Administration			–	
Classroom instruction	–	–		
Evaluation	+	+		–
PD received	–			+
PD provided			+	
Recruitment	+	+		
Reform	+		–	

SOURCES: SL surveys conducted in the springs of 2013 and 2015.
NOTE: In the first two columns, + denotes that weekly hours for principals were statistically significantly higher than weekly hours for APs ($p < 0.05$), and – denotes that weekly hours for principals were statistically significantly lower than weekly hours for APs ($p < 0.05$). In the second two columns, + denotes that weekly hours in 2014–2015 were statistically significantly higher than weekly hours in 2012–2013 ($p < 0.05$), and – denotes that weekly hours in 2014–2015 were statistically significantly lower than weekly hours in 2012–2013 ($p < 0.05$).

RAND *RR2242-9.8*

we interviewed reported that the central office had removed managerial tasks to help principals focus on evaluation activities. PPS was one such site.

The CMOs, meanwhile, created new SL positions to support their increasing student enrollments and their expansion. Alliance, Aspire, and PUC almost doubled their numbers of APs during the first two years of the IP initiative. In the earlier years of the initiative, principals were responsible for teacher observations, but, in the later years, APs took on some of the evaluation tasks.

Our estimates of the amount of time principals spent on evaluation are considerably higher than what other studies have found. Our analyses of principals' time allocation in 2014–2015 indicate that, on average, principals spent 13 hours per week (23 percent of their time) on evaluation-related activities. None of the studies we reviewed, as noted earlier in this chapter, obtained nearly as high an estimate. The IP sites might have required more observations than the places studied in the extant literature. However, some of the difference between our results and previous results might also be that our definition of *time spent on evaluation* included time spent attending training to conduct teacher evaluations, as well as actually conducting the evaluations. In addition, some of the other studies estimate principals' time spent on evaluation in terms of hours per evaluated teacher, so their results are not directly comparable to ours because we use a different metric.

Comprehensive Cost of Operating the Teacher-Evaluation System in 2014–2015

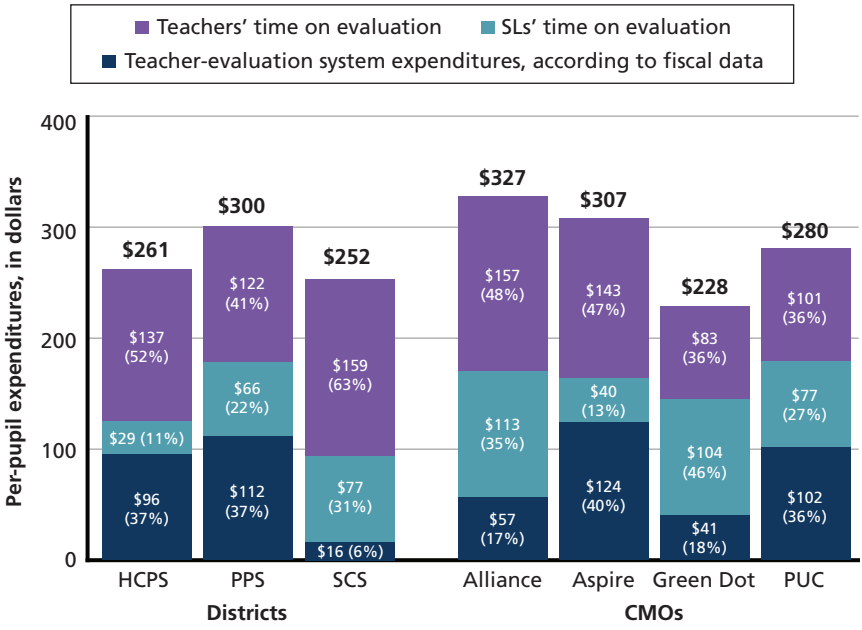
The value of the time that teachers and SLs spent on implementing evaluation activities exceeded the direct expenditures on teacher evaluation. In an effort to account for the total resources that the sites invested in the IP initiative, we estimated the monetary value of the time that teachers and SLs reported allocating to activities related to the teacher-evaluation system, as described in the previous section. We focused on 2014–2015 because, by then, most of the sites were implementing the IP initiative at the level they were likely to continue into

future years. To estimate a financial value for time, we used compensation data that the sites provided. These data included information on teachers' and SLs' base salaries, overtime pay, bonuses, and benefits. (See Appendix L for further detail.)

Figure 9.9 shows each site's per-pupil expenditures for the teacher-evaluation implementation lever (bottom section of the bar), as well as the estimated value of SLs' time allocated to evaluation (middle part of the bar) and the estimated value of teachers' time allocated to evaluation (top part of the bar).

Accounting for teachers' and SLs' time increases the cost of teacher evaluation by about $200 per pupil. In 2014–2015, the eval-

Figure 9.9
Comprehensive Cost of the Teacher-Evaluation Lever in 2014–2015



SOURCES: IP site financial reports for the springs of 2015 and 2016, teacher and SL surveys conducted in the spring of 2015, and compensation data for FY 2015. See Table L.1 in Appendix L for details about the financial reports. Because of rounding, category percentages might not sum to 100; similarly, because of rounding, category dollar amounts might not sum to the total amounts at the top.

RAND *RR2242-9.9*

uation system expenditures based on the fiscal data averaged about $78 per pupil across the seven sites. The SL time spent on evaluation activities (12 to 15 hours per week) added $72. The teacher time spent on evaluation activities (about 2.5 hours per week) added, on average across sites, another $129 to the per-pupil expenditures. In other words, the inclusion of the time-based per-pupil costs more than triples the expenditures based solely on the fiscal data, from $78 to $280 per pupil.

This estimate of $280 per pupil is within the range of per-pupil costs that the studies cited earlier found. It is slightly lower than the $300 per pupil suggested in the Cincinnati study (Taylor and Tyler, 2012) but within the range of the per-pupil costs of TAP (Toch, 2008).

The estimated per-pupil value of staff time varied widely across the seven sites, from $166 in HCPS to $270 in Alliance. The addition of the time-based costs had the greatest implications for SCS, in which the time-based costs added $236 per pupil to the very low $16 per pupil based solely on fiscal costs. (Recall from earlier in this chapter that SCS spent less per pupil than the other sites, in terms of direct expenditures, on the teacher-evaluation lever.) Alliance and Green Dot also saw a substantial—almost six-fold—increase in their estimated costs.

The comprehensive costs for teacher-evaluation activities totaled almost $100 million across the seven sites in 2014–2015. The estimated expenditures based on the fiscal data totaled $26 million, and the estimated value of teachers' and SLs' time added $73 million. As previously discussed, by including the value of staff time, the comprehensive estimate reflects a complete picture of the resources invested in the IP initiative.

Summary

Not including the cost of staff time, the seven IP sites spent, on average, $1,766 per pupil on the IP initiative between its inception in 2009 and the end of the 2015–2016 school year. Overall IP expenditures varied substantially from site to site, as did per-pupil expenditures, per-

haps depending on such factors as the starting point of their teacher-evaluation systems, the design of each lever (e.g., number of observations required, size of effectiveness-based bonuses), the use of outside consultants (e.g., to calculate VAM scores or to help recruit teachers), and investment in technology. The previous chapters of this report have detailed the many ways in which the sites differed along these and other dimensions, as do several of the appendixes.

The IP evaluation was funded mainly by the Gates Foundation, but the share of costs that the foundation funded decreased over time, as planned, as sites increased the funding based on local or federal sources. We examined direct expenditures on four implementation levers: teacher evaluation, staffing, PD, and compensation and CLs. Over the seven years of the initiative, the largest expenditures were related to the compensation and CL lever, with a cross-site average of $656 per pupil spent on this lever. The PD lever constituted the second-largest expenditure, with an average of $540 per pupil. Per-pupil expenditures for teacher evaluation averaged $381, and the staffing lever had the lowest per-pupil expenditures, at $189 averaged across the seven sites.

In addition to direct expenditures, we estimated the value of the time that teachers and SLs reported spending on evaluation activities during the 2014–2015 school year. According to our estimates across the seven sites, school principals spent, on average, about 25 percent of their time on evaluation-related activities in 2014–2015. This estimate is considerably higher than what other researchers have reported. Rowan, Schilling, and their colleagues found that, in Michigan, principals spent about 16 percent of their time on teacher evaluation (Rowan, Schilling, et al., 2013); Sporte and her colleagues, looking at school principals in Chicago, reported 7 to 9 percent of time spent on teacher evaluation (Sporte et al., 2013). Other estimates are even smaller (Grissom, Loeb, and Master, 2013; Sebastian, Camburn, and Spillane, 2018). Our estimates might be higher because they include time that principals spent attending training on conducting teacher evaluations, but, most likely, that alone does not account for the differences between our findings and those of other studies.

Including the value of time in cost estimates, which we did for 2014–2015, puts the cost 3.5 times larger than an estimate based only on fiscal data. Across the seven sites, the overall (i.e., taking time into account) cost related to the teacher-evaluation implementation lever averaged $280 per pupil. The value of SLs' time spent on evaluation activities represented 26 percent of this overall cost, and the value of teachers' time spent on these activities accounted for 46 percent of the total cost.

CHAPTER TEN

The Effectiveness of Newly Hired Teachers

In Brief

Did the effectiveness of newly hired teachers improve over time?

Our analysis found little evidence that new policies related to recruitment, hiring, new-teacher support, and other staffing matters in the IP sites led to hiring more-effective teachers. Most SLs were satisfied with the quality of new hires at the beginning of the initiative, and their levels of satisfaction did not increase over the next six years. Although the distribution of the site-calculated composite TE levels of new teachers grew more positive over time in some sites, the increases appear to be the result of inflation in effectiveness scores; we did not find similar increases in study-calculated VAM effectiveness levels, which are not subject to such drift, and we saw similar increases in the effectiveness of experienced teachers. Small samples and other data limitations prevent us from assessing the effects of specific IP levers on new-teacher effectiveness.

Introduction

The preceding chapters—Chapter Four especially—identified policies implemented during the IP initiative that were designed, in whole or in part, to attract and hire new teachers who would prove to be highly effective. Sites designed policies to meet their unique needs, expand the recruitment pool, improve the identification of effective teachers, incorporate new technologies, and be responsive to state and local contexts. Changes to recruitment and hiring policies were generally uncontro-

versial. However, all the sites experienced challenges in implementing the new policies. This chapter describes efforts to improve the effectiveness of new hires, presents survey evidence on SLs' perceptions about the success of the efforts, and shares empirical analyses of changes in measured effectiveness over time.

IP Policies Designed to Influence New-Hire Effectiveness

Although all the sites had hiring policies prior to the IP initiative, they made many modifications to attract more and stronger candidates and to recruit such candidates earlier. Large-scale changes included SCS contracting with TNTP to expand its recruitment efforts, identify vacancies, and screen candidates. Large-scale changes that the CMOs implemented included Aspire's teacher residency program, which started in 2010–2011, and the introduction of an HC department in Green Dot in 2011–2012 focused on teacher recruitment, retention, PD, and performance management.

Small-scale changes included advancing hiring deadlines, which HCPS did in 2013–2014. Other policy changes included mentoring programs and diversity efforts initiated in the districts and an increase in early offers of employment in some CMOs. HCPS introduced a mentor program beginning in 2010–2011 and highlighted it in recruitment materials to increase the district's attractiveness. Each of the three districts adopted some type of online application process during the IP initiative, while the CMOs already had such systems. These application processes included gathering information about candidates' effectiveness as measured by the district's rubric. Additional technologies to automate and speed up elements of the hiring process were adopted during the IP initiative. For instance, to facilitate candidate screening and application tracking, HCPS piloted the AppliTrack system in 2013–2014 and fully implemented it in 2015–2016. Some CMOs also changed in their applicant screening processes and provided occasional signing bonuses.

Hiring policies in each of the IP sites were driven partly by local and state contexts. For example, SCS faced unique hiring challenges

because of the merger of legacy MCS and legacy SCS and the creation of the ASD. According to our interviews with central-office staff, many effective teachers left SCS to take positions in ASD schools or because of the uncertainty that accompanied the merger. Additionally, Tennessee state laws eliminated seniority and internal candidate status as factors in hiring. PPS faced budget shortfalls and declining enrollment, which meant that few new teachers were hired. Most of the CMOs' schools are in California, where local teacher labor markets were highly competitive throughout the period of the initiative, and each CMO made improvements to recruitment and hiring to try to attract the best teachers. All the IP sites faced difficulties filling positions in schools with low-income, low-performing, and high-minority populations.

Unfortunately, these changes were not staged in a way that allowed us to test their separate impact, and we do not have adequate comparison or control groups to estimate causal effects of specific policy changes. Thus, our analysis was limited to examining changes over time in the effectiveness of newly hired teachers and to associating these changes with hiring and recruitment policies that the sites implemented. If the policies worked as intended, we would expect the quality of new hires to have increased over time. However, the effects of the policies might not be seen immediately following their implementation. Additionally, because each site implemented multiple policies simultaneously, the independent effects of any one policy cannot be determined. To the extent possible, we examine the impact of all policy changes collectively in terms of the hiring trends seen in each IP site.

Our findings related to the effectiveness of newly hired teachers come from two sources of evidence: (1) self-reports from SLs in response to annual surveys and (2) trends over time in comparisons between the effectiveness of newly hired teachers and that of experienced teachers, as measured by the sites' measures of TE and the study-calculated measures of teacher value added.

SL Perceptions

In all the sites, most SLs indicated that they were satisfied with the performance of newly hired teachers. Across sites, there was no consistent increase over time in SLs' satisfaction with new teachers. In every year from 2011 through 2016, we asked SLs how satisfied they were with the performance of new teachers (see Figure 10.1). Averaging across the years, we found that SLs were most likely to express satisfaction in HCPS (90 percent), followed closely by Alliance (88 percent); however, the percentages varied more from year to year in Alliance than in HCPS. The average percentage expressing satisfaction was lowest in Aspire (78 percent) and PPS (77 percent). There appears to be no trend over time that is consistent across sites. As Figure 10.1 shows, the percentages expressing satisfaction seem stable in HCPS and SCS; Aspire and Green Dot might have trended downward, while PUC might have trended upward.[1] In the three districts, SLs at lower-LIM schools (defined as schools in the bottom half of their site's school LIM distribution) were more likely than teachers at higher-LIM schools to say that they were satisfied with the performance of newly hired teachers.[2] For example, in SCS in 2015, 87 percent of SLs at lower-LIM schools reported satisfaction with new hires, more than the 71 percent at higher-LIM schools. We observed a similar difference in some of the CMOs in some years, although, in PUC in 2016, the difference was in the opposite direction.

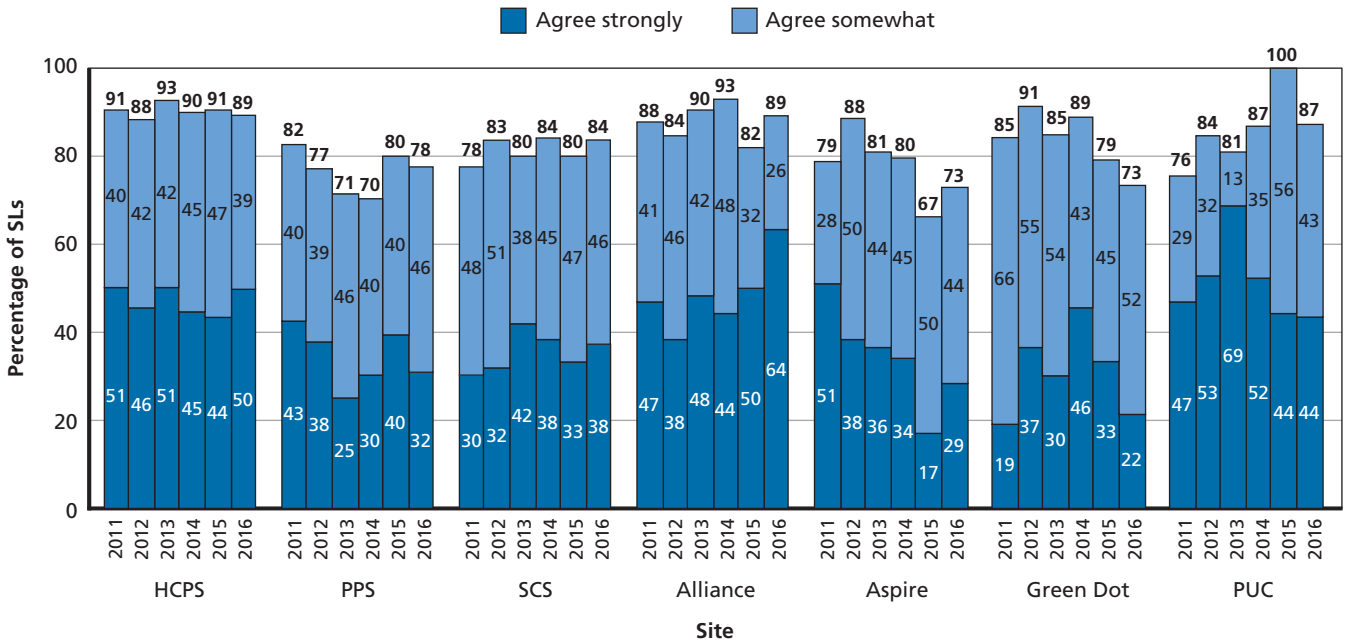
Most SLs were satisfied with the performance of both traditionally and alternatively prepared teachers, although satisfaction with traditionally prepared teachers was slightly higher than with

---

[1] In addition to the variability across years, the smallness of samples in the CMOs make trends difficult to discern. In Aspire, only in 2012 was the percentage agreeing (88 percent) statistically significantly different from the percentage in 2016 (73 percent) ($p < 0.05$). In Green Dot, the percentages in 2012 (91 percent) and 2014 (89 percent) were statistically significantly different from the percentage in 2016 (73 percent). In PUC, the increase in the percentage from each prior year to 2015 was statistically significant but so too was the decline in the percentage from 2015 to 2016.

[2] The difference was significant ($p < 0.05$ or lower) in HCPS and SCS in 2014, 2015, and 2016 and in PPS in 2014 and 2015. (We did not make the comparison prior to 2014.)

Figure 10.1
SLs' Agreement That, Overall, They Were Satisfied with the Performance of Teachers Who Were Newly Hired to Their School That Year, Springs 2011–2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly." The question as it appeared in the survey also had a "not applicable" option, but we coded responses of "not applicable" as missing in the calculation of the percentages shown.

RAND *RR2242-10.1*

teachers hired through alternative programs. Starting in 2013, we asked SLs how satisfied they were with "new/beginning" teachers from each of two sources: (1) traditional teacher-preparation programs and (2) alternative teacher-preparation programs. (See Chapter Four for a discussion of the various programs from which the sites recruited teachers during the initiative.)

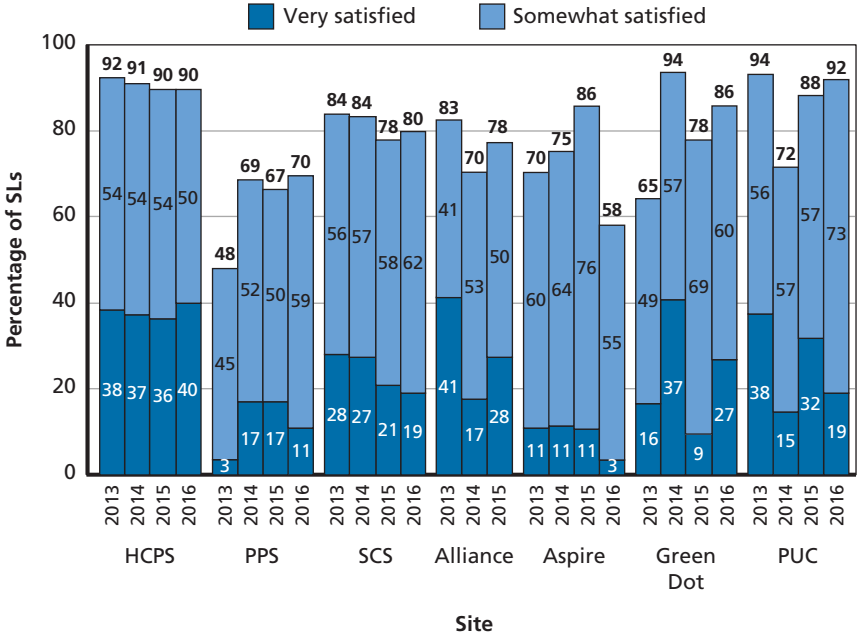
With the exception of PPS in 2013, at least 75 percent of SLs in each site each year reported that their school had hired teachers from traditional preparation programs that year. Of those who reported having hired such teachers, about 90 percent of HCPS SLs and 80 percent of SCS SLs reported being satisfied with the performance of these teachers. The percentages reporting satisfaction were lower in PPS but rose over time, starting at 48 percent in 2013 (when only about half the PPS SLs said that their school even *had* new teachers from traditional programs) and then increasing to about 70 percent in the three subsequent years.[3] In the CMOs, the percentages of SLs reporting satisfaction with traditionally prepared new teachers varied (see Figure 10.2).

According to SLs, new hires from alternative teacher-preparation programs were less common than new hires from traditional programs. In PPS, less than 40 percent of SLs in any year reported that their school had hired new teachers from alternative preparation programs, which is consistent with the district's recruitment procedures as discussed in Chapter Four. In the other six sites, typically 55 to 80 percent of SLs reported having new teachers from alternative preparation programs, with the exact percentage varying by year.

Among the SLs who reported that their school had hired alternatively prepared teachers, on average across years, about 75 percent of SLs in each site except PPS were satisfied with these new hires (see Figure 10.3). However, the percentages varied from year to year more in some sites than others (likely because of smaller samples—i.e., fewer SLs—in the more-varying sites). In Aspire, Green Dot, and PUC, the

[3] The increase from 2013 to each of the three later years was statistically significant ($p < 0.05$).

Figure 10.2
SLs' Satisfaction with Recently Hired Beginning Teachers from Traditional Teacher-Preparation Programs, Springs 2013–2016



NOTE: We asked this question only of SLs who reported that their schools had at least one new teacher in that school year. We inadvertently did not ask the question of Alliance SLs in 2016. Omitted response categories are "somewhat dissatisfied" and "very dissatisfied." The question as it appeared in the survey also had a "not applicable—no teachers from this source" option, but we coded responses in this category as missing in the calculation of the percentages shown.

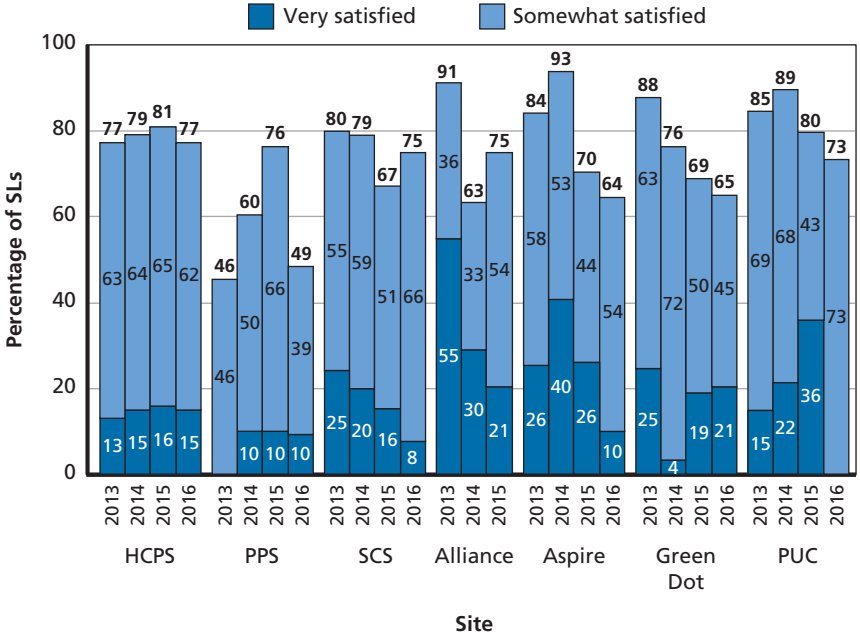
RAND *RR2242-10.2*

percentages expressing satisfaction were lower in 2016 than in any prior year,[4] and there appears to be a downward trend.

Figures 10.2 and 10.3 together appear to suggest that SLs were slightly more likely to express satisfaction with traditionally prepared teachers than with alternatively prepared teachers. In both 2015 and 2016, SLs in six of the seven sites were more likely to be satisfied with

[4] The decrease from 2013 to 2016 was statistically significant ($p < 0.05$) in Aspire and Green Dot. The decrease from 2014 to 2016 was significant only in Aspire. The decrease from 2015 to 2016 was not significant in any of the three sites.

Figure 10.3
SLs' Satisfaction with Recently Hired Beginning Teachers from Alternative Teacher-Preparation Programs, Springs 2013–2016



NOTE: We asked this question only of SLs who reported that their schools had at least one new teacher in that school year. We inadvertently did not ask the question of Alliance SLs in 2016. Omitted response categories are "somewhat dissatisfied" and "very dissatisfied." The question as it appeared in the survey also had a "not applicable—no teachers from this source" option, but we coded responses in this category as missing in the calculation of the percentages shown.

RAND *RR2242-10.3*

traditionally prepared teachers than with alternatively prepared teachers. In 2014, the difference in satisfaction between traditionally prepared teachers and alternatively prepared teachers was significant in five of the sites; in 2013, it was significant in four sites.

Changes in the Measured Effectiveness of New Hires

Because all the sites created composite measures of TE, we can directly examine the performance of newly hired teachers and how that performance changes over time (once teachers have taught long enough to receive effectiveness scores). We use two approaches to analyzing changes in the effectiveness of newly hired teachers. First, we assess changes in the effectiveness of newly hired teachers by looking at trends in their study-calculated VAM scores. Although we can estimate VAM scores on only a subset of teachers, these scores are more reliable than the composite TE levels as a fixed measure of changes in effectiveness because they have a similar meaning over time. VAM scores are adjusted based on state National Assessment of Educational Progress (NAEP) performance trends to make them equivalent across states and over time. VAM scores for some newly hired teachers were available in HCPS, PPS, SCS, and Aspire.[5] Appendix A describes the methodology we employ for estimating value added; Appendix N contains additional analyses of trends in the effectiveness of teachers with three to five years of experience and six or more years of experience.

Second, we examine site-specific trends in the composite TE scores of newly hired and experienced teachers.[6] These data are harder to interpret because the distribution of composite TE scores in a site can shift from year to year even if the underlying effectiveness of the teachers remains constant. In fact, evidence presented in Chapter Four shows that these ratings have been increasing over time, but the source of these trends is unclear and could be due to SLs inflating the observation scores they give to teachers. To avoid being misled by a drift in the composite TE measures that is not a true change in TE, we compare composite TE score trends with VAM score trends, and we compare TE trends in newly hired teachers with TE trends in experienced teachers. The first comparison is important because, if we see changes

---

[5] In Chapter Eleven, we describe our methods for calculating VAM scores.

[6] We have sufficient data to examine these trends in all sites except PUC. We could not analyze these trends in PUC because PUC stopped giving teachers composite effectiveness ratings after 2012–2013.

in composite TE trends but not in VAM score trends, that could be evidence of an upward drift that is not due to true changes in effectiveness. There are two possible explanations for an increase in composite TE when VAM score trends are constant: (1) a drift in the composite TE measure that does not reflect true improvement or (2) there is true improvement in effectiveness for a subset of teachers—namely, teachers who received composite TE ratings without VAM scores are improving, whereas effectiveness ratings for teachers with VAM scores are not changing. Here, the second explanation is possible but not likely. We conduct the second comparison as an additional test of a drift in composite TE scores. If the changes in the composite TE scores of new hires parallel changes for more-experienced teachers, there are again two possible explanations: (1) a drift in the composite TE score that does not reflect true improvement or (2) an increase in TE for all existing teachers *and* an improvement in teacher-preparation programs such that new teachers are also more effective over time. Again, the second explanation seems less plausible. Taken together, these two comparisons provide the most-rigorous evidence we can produce with the available data about whether the changes in new-hire effectiveness are true improvements.

For the analysis, we grouped teachers into three categories based on experience: one to two years of experience, three to five years of experience, and six or more years of experience. We measure experience as the number of years teaching in the district, and new hires are those with one to two years of experience.[7] For the analysis, we consider a teacher to be any employee in the district personnel file who is associ-

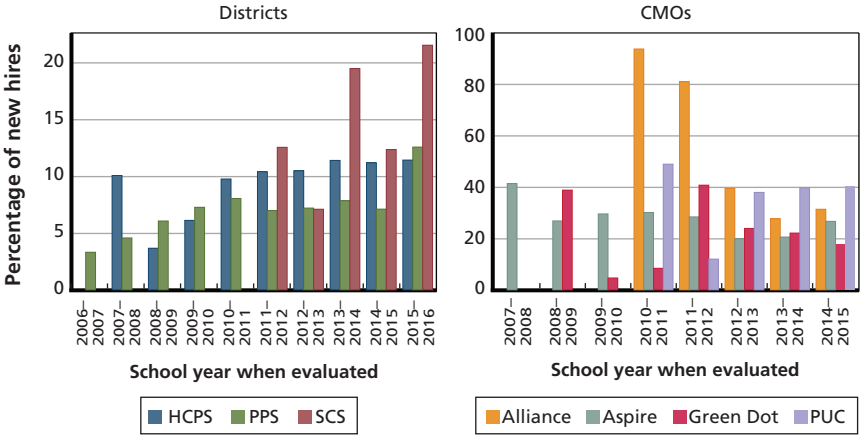
[7] We could not isolate experience at the school level and consequently cannot assess the change in the composition of internal candidates (i.e., those who previously taught within the district).

To increase the size of the new-hire sample, we grouped hires with one or two years of experience together as new hires. Furthermore, the measures of effectiveness for teachers with one or two years of experience are likely to reflect largely their training prior to hiring rather than on-the-job training.

ated with a course with a nonzero number of students.[8] We calculated a VAM score for every teacher who taught either mathematics or ELA to at least ten students in grades 4 through 8.

Figure 10.4 shows the percentage of all teachers who were new hires, by site and year (where data are available). SCS stands out among the districts, both because the proportion of new hires was as great as 20 percent and because the proportion varied considerably over time. This variability is likely associated with the merger of legacy MCS with legacy SCS, the subsequent split-off of some municipal schools, and the creation of the ASD, among other factors. The percentage of new hires in each CMO is generally larger than it is in the districts (as reflected in the vertical scales of the two charts); this is due to a few factors, including the growth in the number of schools managed by the CMOs and the amount of annual teacher turnover they tended to experience. The level and variability in hiring in the CMOs generally declined during the period we studied. It is important to remember when interpret-

Figure 10.4
New Hires in Each Site



RAND *RR2242-10.4*

[8] The number of employees defined in this manner is greater than the number of employees who had the job title of teacher and received composite TE scores, which is the sample used for some of the other analyses.

How to Read the Effectiveness Figures

This chapter includes several figures depicting change over time in the effectiveness of newly hired teachers (those with one year or two years of experience teaching in the district). The following is a guide to how to read them:

- *The panels* describe our analysis by VAM scores (left) and composite TE measures (right).
- *The horizontal axis* indicates the school year in which teachers' effectiveness was evaluated.
- *The vertical axis* indicates the percentage of newly hired teachers. In each year, the percentages for the three effectiveness levels sum to 100.
- *The blue bars* depict the percentage of newly hired teachers in the bottom 20 percent of the VAM score distribution or in the lowest level of composite TE. For VAM scores, the expectation is that this bar is around 20 percent. Because composite TE levels are not equally distributed, there is no a priori expectation for composite TE.
- *The red bars* describe the percentage of newly hired teachers who scored in the middle 60 percent of the VAM score distribution or who received middle composite TE ratings. For VAM scores, the expectation is that this bar is around 60 percent. Because composite TE levels are not equally distributed, there is no a priori expectation for composite TE.
- *The green bars* show the percentage of newly hired teachers in the top 20 percent of the VAM score distribution or in the highest level of composite TE. For VAM scores, the expectation is that this bar is around 20 percent. Because composite TE levels are not equally distributed, there is no a priori expectation for composite TE.

ing the findings that follow that we have composite measures of TE for majorities of teachers in each site, but the study-calculated VAM score is available for only a small proportion of teachers. Additionally, the number of teachers is not evenly distributed across composite TE levels; typically, the lowest level has the smallest number of teachers, so statistics describing this group are less precise (i.e., they have a greater margin of error).

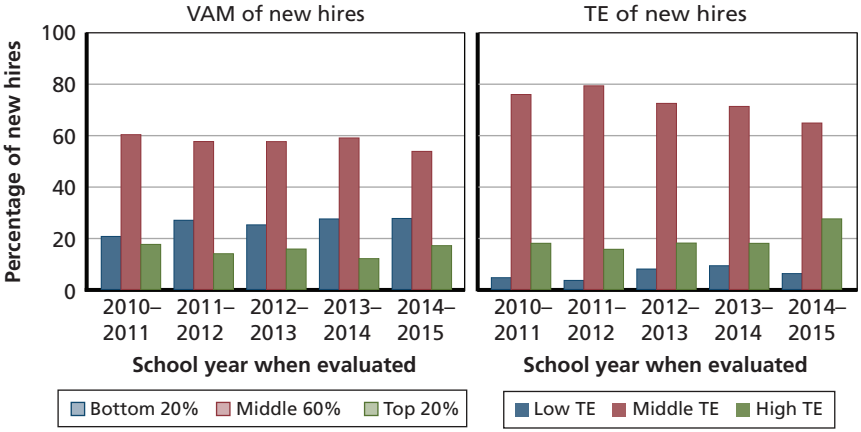
Figures 10.5 through 10.10 illustrate changes over time in the relationship between the composite TE measure and teaching experience for each site. We have two different measures of effectiveness: the study-calculated VAM score and each site's composite TE score. We separate VAM into three levels: bottom 20 percent, middle 60 percent, and top 20 percent. If the policies had no effect on newly hired teachers, we would expect 20 percent of newly hired teachers to be in the bottom 20 percent on the VAM score, 60 percent to be in the middle 60 percent, and 20 percent to be in the top 20 percent. Deviations from these levels would indicate that newly hired teachers differ from the rest of the teachers in the site in terms of effectiveness. The expec-

tation differs when it comes to the sites' composite TE levels. The sites originally divided teachers into four or five levels (depending on the site) based on the composite TE score; we collapsed each site's TE levels into three: low, middle, and high. Because teachers are not distributed across these composite TE levels predictably (as is the case with the three VAM levels), we do not have an expectation for the distribution of newly hired teachers across levels.

HCPS

Over time, effectiveness of newly hired teachers in HCPS did not increase. Figure 10.5 shows the effectiveness of new hires in HCPS over time. The left-hand panel shows the trend in effectiveness categories based on VAM scores. In 2010–2011, the distribution of the effectiveness of newly hired teachers is very close to what we would expect: Twenty percent of newly hired teachers were in the bottom 20 percent of the VAM score distribution, 60 percent of newly hired teachers were in the middle 60 percent, and 20 percent of newly hired teachers were in the top 20 percent. However, the percentage of newly hired teachers who were in the bottom 20 percent of the VAM score distribution

Figure 10.5
HCPS's New-Hire Effectiveness, by VAM Score and Composite TE Level



NOTE: For TE, low TE = U or NI, middle TE = E, high TE = HE and most E.

RAND *RR2242-10.5*

significantly increased in 2011–2012. This increase was countered by a decrease in the percentage of newly hired teachers in the top 20 percent (or, in 2014–2015, in the middle 60 percent). Thus, the VAM measure of effectiveness shows a small decline in the effectiveness of newly hired teachers during this period.

The right-hand panel of Figure 10.5 shows the trends in composite TE levels for newly hired teachers during the same time frame. In 2010–2011, approximately 5 percent of newly hired teachers had low composite TE, 75 percent had middle composite TE, and 20 percent had high composite TE. The percentage of newly hired teachers in the low-TE level increased in 2012–2013 and 2013–2014, although it declined slightly in 2014–2015. During the same period, the percentage of newly hired teachers in the high-TE category remained roughly the same before significantly increasing in 2014–2015.

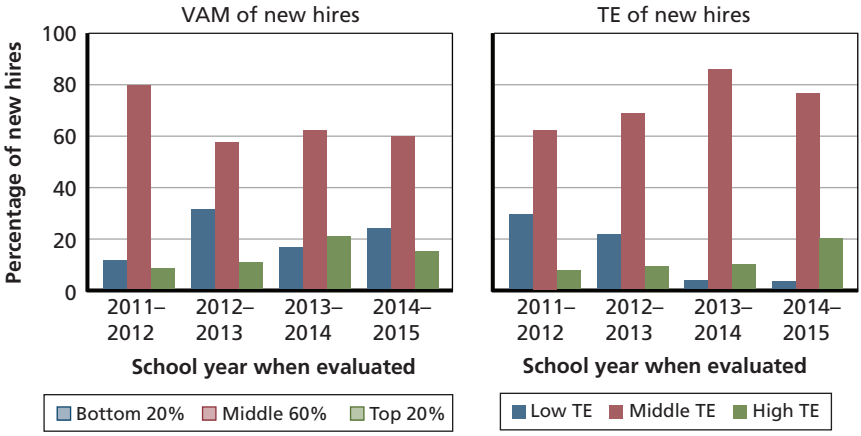
How should one interpret these changes? There were more newly hired teachers in the low-VAM and low-TE categories each year from 2010–2011 onward; this does not suggest general improvement over time. However, composite TE levels improved in 2014–2015. To interpret this result, it helps to look at changes in the composite TE scores of more-experienced teachers. Figures N.1 and N.2 in Appendix N show that the percentage of teachers with three to five and six or more years of experience who had high composite TE levels also increased in 2014–2015. Taken together, these comparisons suggest that the increase in the percentage of newly hired teachers who were high-TE teachers in 2014–2015 was likely the result of measurement drift. Overall, the evidence in HCPS suggests that, over time, the district did not manage to hire new teachers who were more effective early in their careers than it had in the past. As discussed earlier, it is impossible for us to associate this trend with specific policies implemented by the district.

PPS

There is no clear pattern in the effectiveness of newly hired teachers in PPS during the IP initiative. We were able to calculate VAM scores for only a small number of teachers in PPS (fewer than 400 each year); when divided into nine groups (three VAM levels and three experience levels), the total number of teachers in each group is smaller still.

Consequently, there is substantial variation in the calculated proportions over time. This limits the conclusions that we can draw about the impact of hiring policies in PPS based on VAM levels. In the left-hand panel of Figure 10.6, we see the trends in the distribution of newly hired teachers' effectiveness by VAM score. In 2011–2012, approximately 10 percent of newly hired teachers were in the bottom 20 percent of the VAM score distribution, 80 percent in the middle 60 percent, and 10 percent in the top 20 percent. The percentage of newly hired teachers in the bottom 20 percent dramatically increased in 2012–2013 (although the change was not statistically significant), mirrored by a significant decrease in middle-VAM teachers. In 2013–2014, the percentage of newly hired teachers in the top 20 percent of the VAM score distribution increased to slightly above 20 percent, while the percentage of low-VAM new hires decreased. However, the improvement was short-lived; in 2014–2015, the percentage of low-VAM new hires increased to slightly more than 20 percent, and the percentage of high-VAM teachers decreased to approximately 15 percent.

Figure 10.6
PPS's New-Hire Effectiveness, by VAM Score and Composite TE Level



NOTE: For TE, low TE = F or NI, middle TE = P, and high TE = D. We base 2011–2012 TE results on a pilot version of PPS's composite measure that was never shared with teachers or SLs; the composite TE measure became fully operational, with stakes attached, beginning in 2013–2014.

RAND *RR2242-10.6*

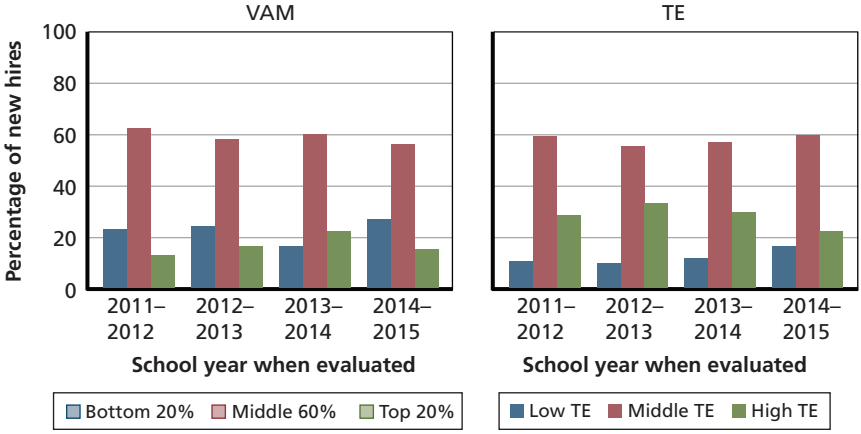
The right-hand panel of Figure 10.6 presents the new-hire trends by composite TE levels.[9] In 2011–2012, 30 percent of new hires had low composite TE, more than 60 percent had middle composite TE, and just under 10 percent had high composite TE. Over the next three years, the percentage of new hires who had high composite TE levels increased, particularly in 2014–2015, while the percentage of new hires who had low composite TE levels dropped dramatically, to less than 5 percent. (However, the year-by-year changes were not statistically significant.) If trustworthy, these numbers indicate very positive changes. However, we do not observe a similar increase in the effectiveness of new hires as measured by VAM scores (the left-hand panel), although these two trends might not be directly comparable, given that the composite TE levels of new teachers in PPS did not include a site-calculated VAM score component. Interestingly, the percentage of experienced teachers classified with high composite TE levels increased during this period, particularly in 2014–2015 (see Figures N.3 and N.4 in Appendix N). These findings in combination suggest that these changes might have been related to changes in the composite TE measure. Prior to 2013–2014, RISE was the sole composite measure of TE used for personnel decisions in PPS. Then, in 2013–2014, PPS implemented a new composite measure, of which RISE was one component.

SCS

There were no significant changes in the effectiveness of newly hired teachers in SCS over the period of the IP initiative. In 2011–2012, more than 20 percent of newly hired SCS teachers were in the bottom 20 percent of the VAM score distribution, more than 60 percent were in the middle 60 percent of the distribution, and 15 percent were in the top 20 percent (the left-hand panel of Figure 10.7). This suggests that new hires were slightly less effective than expected. Furthermore, except for 2013–2014, the effectiveness of new hires remained relatively constant. There was some improvement in 2013–2014, the first

[9] In PPS, the composite TE scores for teachers in their first or second year do not include a site-calculated VAM score component because PPS uses a three-year average in its site-calculated VAM scores.

Figure 10.7
SCS's New-Hire Effectiveness, by VAM Score and Composite TE Level



NOTE: For TE, low TE = below or significantly below expectations, middle TE = meeting expectations, and high TE = above and significantly above expectations.
RAND *RR2242-10.7*

year following the merger of legacy SCS with legacy MCS, when the percentage of low-VAM newly hired teachers dropped slightly and the percentage of high-VAM newly hired teachers increased somewhat, but that trend reversed the following year.

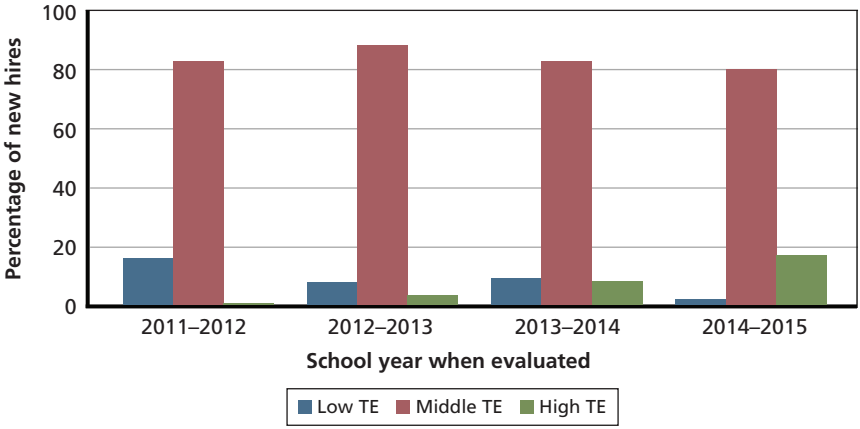
The right-hand panel of Figure 10.7 shows trends in the distribution of new-hire effectiveness as measured in terms of composite TE levels. In 2011–2012, approximately 10 percent of new hires had low composite TE, 60 percent had middle composite TE, and 30 percent had high composite TE. As previously mentioned, there is no a priori expectation about the distribution of newly hired teachers across the three composite TE levels. However, we observe relative stability in the distribution of new-hire effectiveness measured by composite TE level between 2011–2012 and 2013–2014, followed by a nonsignificant decline in new-hire effectiveness in 2014–2015. We found a similar change in the distribution of newly hired teachers' effectiveness measured by VAM scores, suggesting that the change is unlikely to reflect drift in the composite TE levels but reflects a real decline in the effectiveness of new hires.

Although it is difficult to associate these trends with specific hiring policies, it is worth mentioning that the decline in the effectiveness of newly hired teachers noted in 2014–2015 corresponds to a change in state policies related to hiring. In 2014–2015, Tennessee eliminated preferences for internal candidates in hiring, requiring that all new-teacher candidates be entered into the same hiring pool as internal transfer candidates. Theoretically, this larger pool should have led to more-effective hires, but our results suggest that this was not the case in SCS in 2014–2015.

Alliance

We could not determine with certainty whether there were changes in the effectiveness of newly hired teachers in Alliance. Because of the relatively small number of teachers in Alliance (and the even smaller number teaching tested subjects and grade levels), our study did not calculate VAM scores for any teacher in this CMO. Thus, for Alliance, we present new-hire effectiveness only in terms of composite TE levels in Figure 10.8. In 2011–2012, approximately 17 percent of newly hired Alliance teachers had low composite TE, more than 80 percent

Figure 10.8
Alliance's New-Hire Effectiveness, by Composite TE Level



NOTE: Low TE = entering or achieving, middle TE = E, and high TE = HE or master.
RAND *RR2242-10.8*

had middle composite TE, and only 2 percent had high composite TE. Over time, the percentage of newly hired teachers who received high composite TE scores increased. By 2014–2015 nearly 20 percent of newly hired teachers were high-TE. This increase was mirrored by a decrease in the percentage of newly hired teachers who were identified as having low TE. However, this trend could reflect inflation in TE scores. Without VAM scores for comparison, we examined similar distribution trends for teachers with more experience. For teachers with three to five and six or more years of experience, the percentage of high-TE teachers increased even more dramatically than among new hires (see Figures N.7 and N.8 in Appendix N).

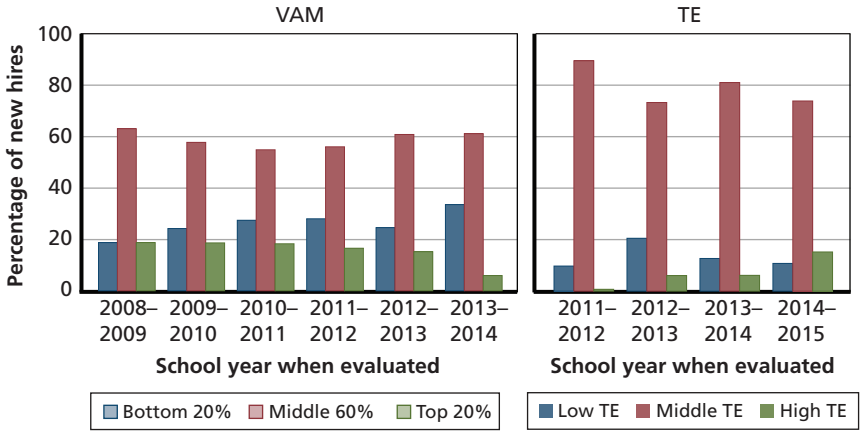
At face value, our results suggest that, over time, more high-TE teachers and fewer low-TE teachers were hired. However, because we observe an even greater increase in TE for teachers with three to five and six or more years of experience, a more likely explanation is that there was inflation in Alliance's composite TE measure over time.

Aspire

There was a decline in the effectiveness of newly hired teachers in Aspire as measured by VAM scores. Figure 10.9 shows trends in effectiveness of newly hired teachers in Aspire. In 2008–2009, just less than 20 percent of newly hired Aspire teachers were in the bottom 20 percent of the VAM score distribution, just more than 60 percent were in the middle 60 percent of the distribution, and slightly less than 20 percent were in the top 20 percent. This is essentially the expected distribution of new hires by VAM score. Unfortunately, over the next few years, new hires were increasingly in the bottom 20 percent of the VAM score distribution and fewer were in the top 20 percent. This suggests that the effectiveness of new hires decreased over time and that policies aimed at hiring more-effective teachers did not work as hoped.

On the other hand, the right panel of Figure 10.9 shows a decrease in the percentage of new hires who had low composite TE and an increase in the percentage who had high composite TE. Because this does not match the trend shown using the VAM score categories, we think that it represents upward drift in composite TE scores. This interpretation is substantiated by an increase in the percentage of teachers

Figure 10.9
Aspire's New-Hire Effectiveness, by VAM Score and Composite TE Level



NOTE: For TE, low TE = entering or achieving, middle TE = E, and high TE = HE or master.

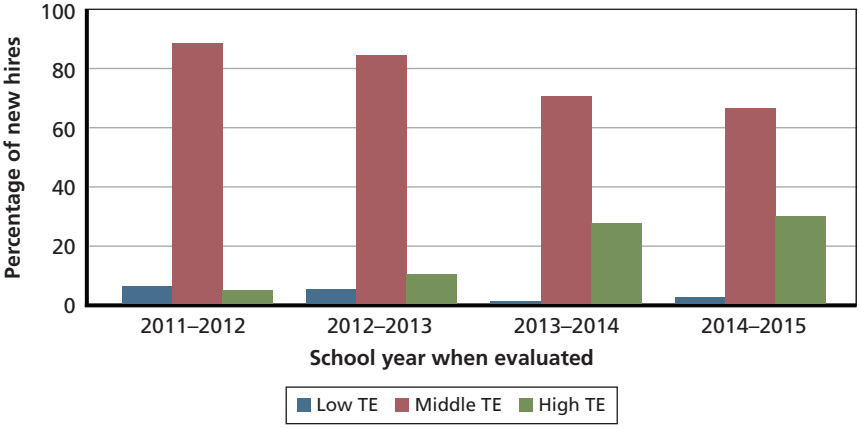
RAND *RR2242-10.9*

with three to five years of experience and six or more years of experience in the high-TE category between 2011–2012 and 2014–2015 (see Figures N.9 and N.10 in Appendix N).

Green Dot

We could not determine with certainty whether there were changes in the effectiveness of newly hired teachers in Green Dot. Figure 10.10 shows hiring trends of Green Dot. VAM scores were not calculated in Green Dot because of the small number of teachers (and the even smaller number teaching tested subjects and grade levels), so Figure 10.10 shows results for composite TE levels only. In 2011–2012, approximately 5 percent of newly hired Green Dot teachers had low TE, nearly 90 percent had middle TE, and 5 percent had high TE. By 2014–2015, nearly 30 percent of newly hired teachers had high TE. This increase in the percentage of newly hired high-TE teachers was mirrored by a large decrease in the percentage of newly hired middle-TE teachers and a slight decrease in low-TE new teachers. However, composite TE is subject to inflation over time, so we examine similar

Figure 10.10
Green Dot's New-Hire Effectiveness, by Composite TE Level



NOTE: Low TE = entering or achieving, middle TE = E, and high TE = HE or master.
RAND *RR2242-10.10*

distribution trends for teachers with more experience to see whether there is a similar pattern. For teachers with three to five and six or more years of experience, the percentage of high-TE teachers increased even more dramatically than among new hires (see Figures N.11 and N.12 in Appendix N). This suggests that there might have been inflation in composite TE levels in Green Dot, so we cannot determine with any confidence whether new policies, such as changes to the HR department, increases in the number of student teachers, establishment of partnerships with local universities, expansion of recruitment efforts, and early offers for hard-to-staff positions, led to more-effective newly hired teachers. We know that Green Dot removed the student achievement measure from the composite TE score in 2012–2013, a change that could have caused the composite TE level inflation that we observe.

Summary

Overall, our analysis shows little evidence that recruitment and hiring policies in the sites led to hiring more-effective teachers. Although the composite TE levels of newly hired teachers increased over time in some sites, these changes appear to be the result of awarding higher composite TE levels to teachers overall and not specifically to increased effectiveness among new hires. This conclusion is supported by analysis of study-calculated VAM levels that are not subject to rater drift over time and by comparisons with the effectiveness of experienced teachers. SL survey responses told a similar story; most SLs were satisfied with the quality of new hires at the beginning of the initiative, and this satisfaction did not increase further over the next six years.

CHAPTER ELEVEN

Retention of Effective Teachers

> Teachers are already motivated because we are here and want to make a change, but I think [that effectiveness-based salary] helps sustain the teachers [who] are in for the long haul.
>
> —Alliance teacher

In Brief

Did the new policies that the sites implemented—including those related to recruitment and hiring, dismissal and tenure, PD, compensation and CLs—collectively increase retention of effective teachers?

Our analysis found that, in the districts and CMOs, highly effective teachers were generally more likely than less effective teachers to remain teaching; however, the likelihood that they would remain in teaching did not increase as a result of the IP initiative. That is, we found little evidence that the policies designed, in whole or in part, to increase retention of effective teachers achieved their goal. On the other hand, there is some evidence that, over the course of the IP initiative, less effective teachers were more likely than more-effective ones to exit from teaching, a positive effect that could be due to some of the same policies. In some sites, we found decreasing retention of less effective teachers when we used the site composite TE scores as the measure of effectiveness; in some sites, we found decreasing retention when we used study-calculated VAM scores as the measure of effectiveness; and, in HCPS, the largest site, we found it for both measures.

Teachers' career plans, admittedly an imperfect measure of future actions, were somewhat consistent with these findings. In some sites, more-

effective teachers and more-experienced teachers were more likely than less effective and less experienced teachers to expect to continue teaching, but, in other sites, the opposite was true. In most sites, less than half the teachers thought that they would still be teaching in their current sites in five years' time, and this was true from 2011 through 2015. Thus, future intentions were not consistently related to the sites' implementation of retention-related policies.

Introduction

In this chapter, we examine whether the policies put in place as a result of the IP initiative affected teacher retention. Following a brief discussion of the policies designed to influence retention and our ability to assess their impact, we present findings from our analysis of teacher retention. We bring two types of evidence to bear on this issue. First, we present findings from surveys of teachers, whom we asked about their future employment intentions, and SLs, whom we asked about their experiences with teacher retention. Second, we report analyses of teacher retention based on employment records obtained from each of the sites.

IP Policies Designed to Influence Retention and Issues Assessing Their Impact

The preceding chapters identified several policies that IP sites implemented that were designed, in whole or in part, to retain effective teachers and dismiss ineffective ones. First, increased attention to teacher support and PD might have improved teacher job satisfaction and commitment. Second, effectiveness-based compensation, whether bonuses or effectiveness-based salary schedules, provided cash incentives to make teaching more financially attractive to effective teachers. Third, CL positions offered some effective teachers additional responsibility and leadership opportunities (as well as additional compensation). Fourth, as we noted in Chapter Four, retention is also influenced

by school-level factors that are within the direct control of SLs. For example, research shows that teacher retention is higher in schools that give teachers more autonomy and more discretion over discipline and in schools with more-responsive grievance policies This research highlights the fact that retention—a teacher remaining in the same site the following year—is jointly determined by the site and the teacher. A teacher might leave for personal reasons (e.g., retirement, relocation out of the area, change of careers) or for administrative reasons (e.g., dismissal for poor performance, furlough because of declining enrollment). We use the term *separation* to include cases in which a teacher does not remain teaching in his or her site. Dismissal (discussed in Chapter Five) is just one type of separation. It is worth noting that the sites also introduced policies aimed at less effective teachers, with the goal of either improving their performance or dismissing them. Our retention analysis focuses on efforts to keep effective teachers teaching, but the evidence is also relevant to analysis of the dismissal of ineffective teachers.

Both individually and collectively, sites hoped that these policy changes would increase the likelihood that effective teachers would remain in the sites working as teachers or mentoring other teachers to improve their practice. Each individual policy could, in theory, have contributed to effective-teacher retention. If the policies improved retention independently of each other, we would expect to see increases in year-to-year teacher retention following the implementation of each policy.[1] Under this ideal assumption, the effects would also be cumulative—as more policies were implemented, there would be greater increases in retention. Then again, if the policies were differentially effective—that is, if some were more potent than others—the retention trends would mirror the implementation pattern of the more-potent policies. And, if the policies interacted in some manner—for example, if implementation of a particular bonus policy reduced the

[1] During the IP initiative, there were broad changes to state laws that could have counteracted any positive effects of retention policies at IP sites. For instance, changes in state tenure rules in Florida and Tennessee that reduced job security for teachers could have reduced teacher retention at the same time as the sites were enacting policies to increase retention. Changes in pension plans might also have had an effect on retention that we cannot estimate.

independent effect of a CL policy—the pattern of responses could be quite varied.

Evidence presented in Chapters Four to Eight showed that the sites implemented retention-related policies at different points in time and that, once the policies were implemented, they were generally not discontinued. Therefore, if the policies were working as imagined, the overall effect should have increased retention of HE teachers (and decreased retention of ineffective teachers) over time. The effects might also be observable as an increase in the retention of HE teachers after IP implementation compared with that before the IP initiative. An exception to the "generally not discontinued" trend was the effectiveness-based bonuses in SCS and the CMOs, most of which were discontinued; if these compensation policies worked as imagined, we would expect to see retention improvements with their introduction, followed by declines when discontinued.[2] Because the sites often implemented multiple policies simultaneously, there are very few cases in which it is possible to isolate the impact of a single policy. For example, PD and CL policy changes were implemented simultaneously in Alliance in 2013–2014, so it is impossible to determine whether one policy had an effect and the other did not or whether they had similar effects. As another example, effectiveness-based bonuses might have led to retention improvements in SCS following their introduction in 2010–2011, but this effect might have been obscured because some HE teachers exited because Tennessee eliminated tenure in the same year. Generally, it is difficult to predict what the policies' impact might be in any given year within any given site because the sites did not stage their reforms in a manner allowing us to test the separate effects of each reform. The best we can do is to examine trends in retention and see whether, collectively, the reforms were associated with greater retention of effective teachers over time.

First, we present survey results reflecting teachers' opinions about their likelihood of remaining in teaching five years in the future. Then

[2] If they were discontinued because they were perceived not to have worked, we might observe no change at their onset and no change at their elimination.

we present empirical evidence about year-to-year retention of teachers, by effectiveness level, in each of the sites.

Teacher Intentions and SL Perceptions

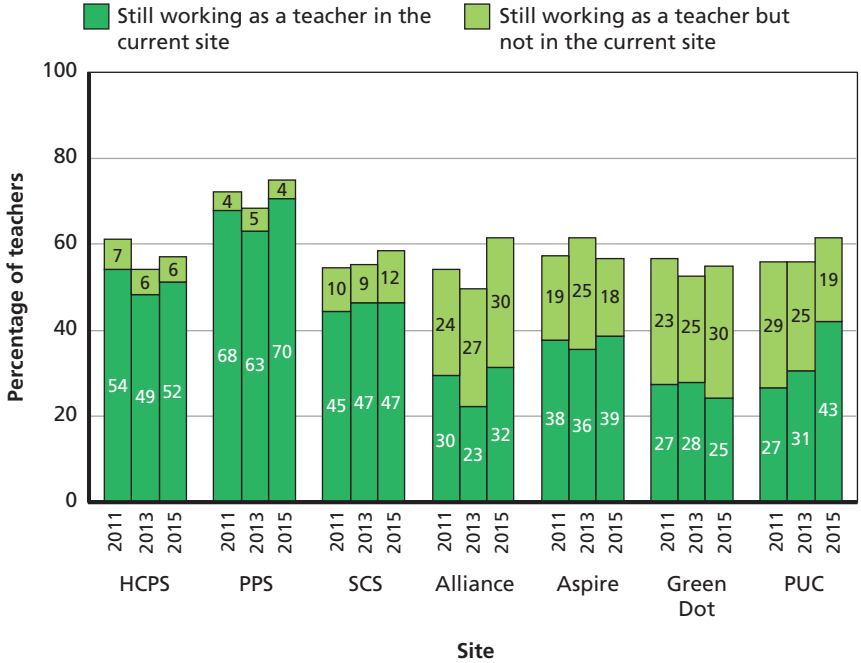
The first evidence we consider related to the effectiveness of policies to increase teacher retention comes from the analysis of survey data on teachers' future employment intentions and SLs' experiences with teacher retention.

In most sites, from 2011 through 2015, less than half the teachers thought that they would be "still working as a teacher" in their current site in five years. In the springs of 2011, 2013, and 2015, we asked teachers what they expected to be doing in five years' time. Figure 11.1 (lower parts of the bars) shows that only in PPS did a large majority of teachers say that they thought that they would still be working as teachers in their current site. In the other two districts, about half the teachers had this expectation, while, in the CMOs, typically 20 to 40 percent of teachers thought that they would still be teaching in their current CMO in five years. As shown by the upper parts of the bars in Figure 11.1, CMO teachers were more likely than teachers in the districts to say that they expected to be working as teachers in five years somewhere other than in their current site.

Within each site, over time, there was relatively little change in the percentage of teachers who expected to still be teaching in the site in five years. A possible exception is PUC, in which the percentage increased from 27 percent in 2011 to 43 percent in 2015, a statistically significant increase.[3]

[3] In no other site was the change from 2011 to 2015 statistically significant ($p < 0.05$), but a few sites did have significant changes from 2011 to 2013 (HCPS and Alliance) or from 2013 to 2015 (PPS, Alliance, and PUC).

Figure 11.1
Teachers Planning to Still Be Teaching in Five Years, Springs 2011, 2013, and 2015



NOTE: Omitted response categories are "working in a non-teaching instructional role," "working as a school or district/CMO administrator," "working in another type of education-related position," "working in a non-education job or career," "attending a full-time advanced degree program," "no longer in paid labor force," and "other."

RAND *RR2242-11.1*

Of course, numerous personal and school factors affect teachers' long-term plans. Our analysis of the 2015 data found the following significant differences in intentions to remain in teaching:

- In HCPS, SCS, and PUC, elementary school teachers were more likely than MS or HS teachers to say that they would likely still be working as teachers in their current site in five years. In Alliance and Green Dot, HS teachers were more likely than MS teachers to expect that they would still be teaching in the site in five years.

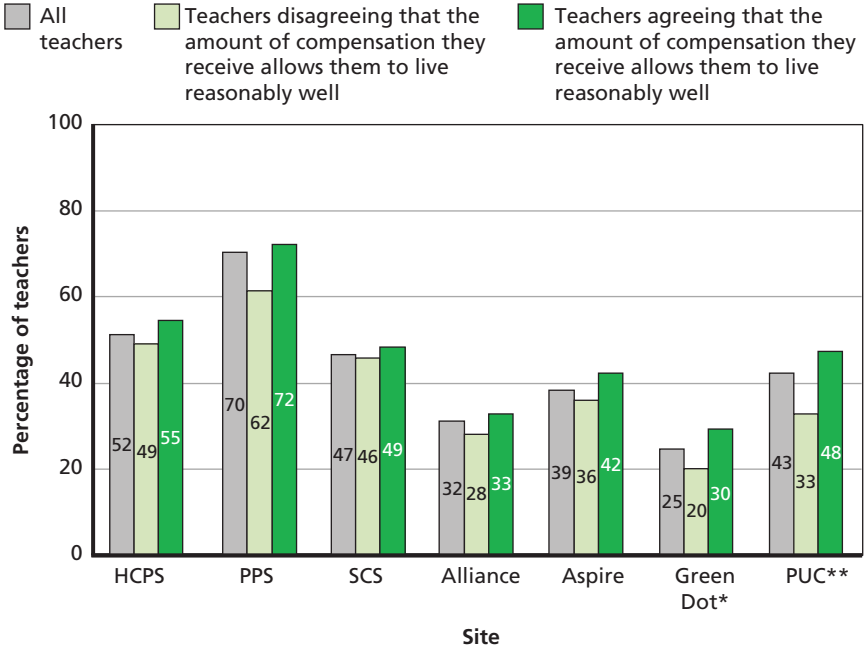
- In PPS and Green Dot, male teachers were more likely than female teachers to expect that they would still be teaching in the site in five years.
- In PPS and Aspire, teachers in their first two years of teaching were more likely than more-experienced teachers to say that they would likely still be teaching in their current site in five years. The opposite was true in PUC. In HCPS, teachers with more than 20 years of experience were less likely than teachers with two to 20 years of experience to have this five-year expectation.
- In HCPS and Aspire, teachers in the bottom third of the within-site LIM distribution (i.e., the lowest-LIM schools) were more likely than teachers in middle-third or top-third LIM schools to say they would likely still be teaching in their current site in five years. The opposite was true in Alliance and Green Dot. And in PPS and Alliance, teachers in middle-third LIM schools were more likely than teachers in top-third LIM schools to have this five-year expectation.

In addition to personal factors, site policies can also influence teachers' plans and decisions to remain. For instance, teachers who agreed that the amount of compensation they received allowed them "to live reasonably well" were more likely than teachers who disagreed to expect to still be teaching in the site in five years, although the difference was significant only for Green Dot and PUC.[4] (See Figure 11.2.) This finding is consistent with prior research evidence (e.g., Podolski et al., 2016).

In only two sites—HCPS and PPS—were teachers who received high effectiveness ratings more likely than low-rated teachers to say that they were planning to still be working as a teacher in their current site in five years. Given the policies put in

[4] As noted in Chapter Seven, teachers in PPS were more likely than teachers in the other sites to agree that the amount of compensation they received allowed them to live reasonably well. In the years in which we asked this question (2014, 2015, and 2016), at least three-quarters of PPS teachers agreed, far more than the 40 to 65 percent in the CMOs and 35 to 50 percent in the other two districts. This could help explain why PPS teachers were more likely than teachers elsewhere to expect to continue teaching in their site.

Figure 11.2
Teachers Planning to Still Be Teaching in Their Current Site in Five Years, by Perception of Salary Adequacy, Spring 2015

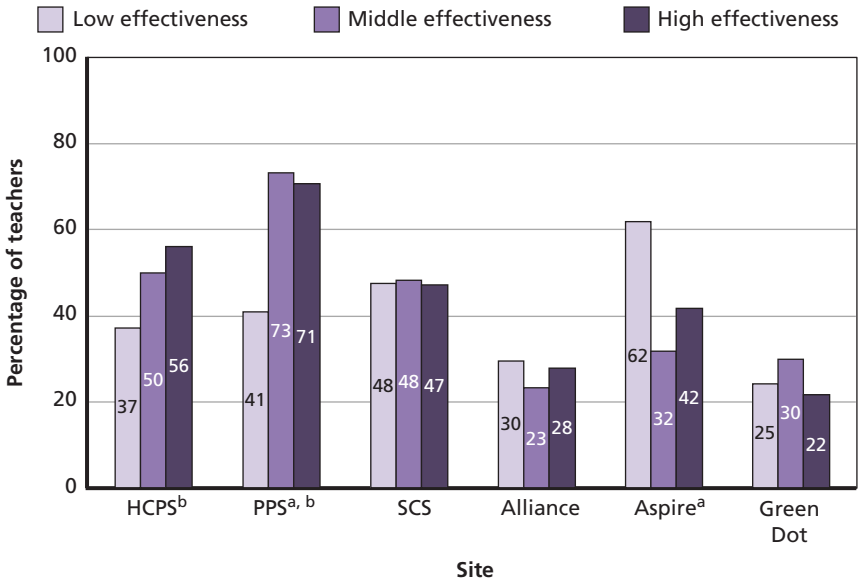


NOTE: * denotes that the difference between agreeing and disagreeing percentages is significant at $p < 0.05$. ** denotes that the difference between agreeing and disagreeing percentages is significant at $p < 0.01$.

RAND *RR2242-11.2*

place by the IP initiative, sites were looking for policies to help them retain *effective* teachers. Figure 11.3 shows how teachers' plans to still be teaching in their current site in five years are related to their TE ratings. As Figure 11.3 shows, in HCPS and PPS, teachers who received the lowest effectiveness ratings for their teaching during the 2013–2014 school year were significantly less likely than the highest-rated teachers to say in spring 2015 (after they had seen their ratings) that they were planning to still be teaching in their current district in five years. However, there were no differences in SCS, Alliance, or Green Dot. In Aspire, low-rated teachers were *more* likely than higher-rated teachers

Figure 11.3
Teachers Planning to Still Be Teaching in Their Current Site in Five Years, by Prior-Year TE Rating, Spring 2015



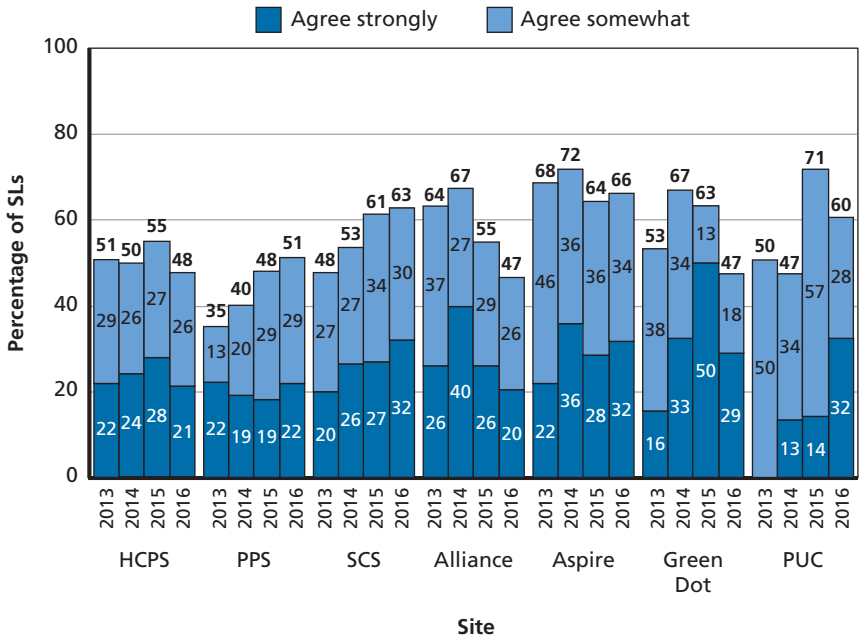
NOTE: We based definitions of *low*, *middle*, and *high* on sites' own categorizations; for details, see Appendix A. Effectiveness ratings were not available for PUC.
[a] Low–middle difference is significant at $p < 0.05$.
[b] Low–high difference is significant at $p < 0.05$.
RAND *RR2242-11.3*

to say that they were still planning to be teachers in the CMO in five years.[5]

A final look at intentions to remain in teaching comes from SLs' responses to the statement, "More often than is good for my school, good teachers leave my staff because they perceive better opportunities elsewhere." (See Figure 11.4.) During the period from 2013 through 2016, Aspire SLs were consistently more likely than SLs in other sites

[5] In interpreting this counterintuitive finding in Aspire, it is worth noting that the low-rated teachers were significantly less likely than high-rated teachers to say that they were planning to be teachers elsewhere. In other words, lower-rated teachers, perhaps as a function of their low ratings, might have perceived themselves to have fewer options for "shopping themselves around" than high-rated teachers had.

Figure 11.4
SLs' Agreement That, "More Often Than Is Good for My School, Good Teachers Leave My Staff Because They Perceive Better Opportunities Elsewhere," Springs 2013–2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly."
RAND *RR2242-11.4*

to agree with this statement. The percentage agreeing rose in PPS and SCS and declined in Alliance.[6] In PUC, a high percentage of SLs agreed with the statement in 2015, and a PUC central-office staff member whom we interviewed in the fall of 2016 provided a possible explanation, referencing local labor market conditions: "I'd say we did have higher-than-normal turnover this year, and that's because people had options. If anybody was commuting, the school down the street hired them. They had a ton of options. One candidate would have, like, five

[6] The SCS increases from 2013 to 2014, 2014 to 2015, and 2013 to 2016 were statistically significant ($p < 0.05$). The PPS increase from 2013 to 2016 was statistically significant ($p < 0.05$). The Alliance decrease from 2014 to 2016 was statistically significant ($p < 0.05$).

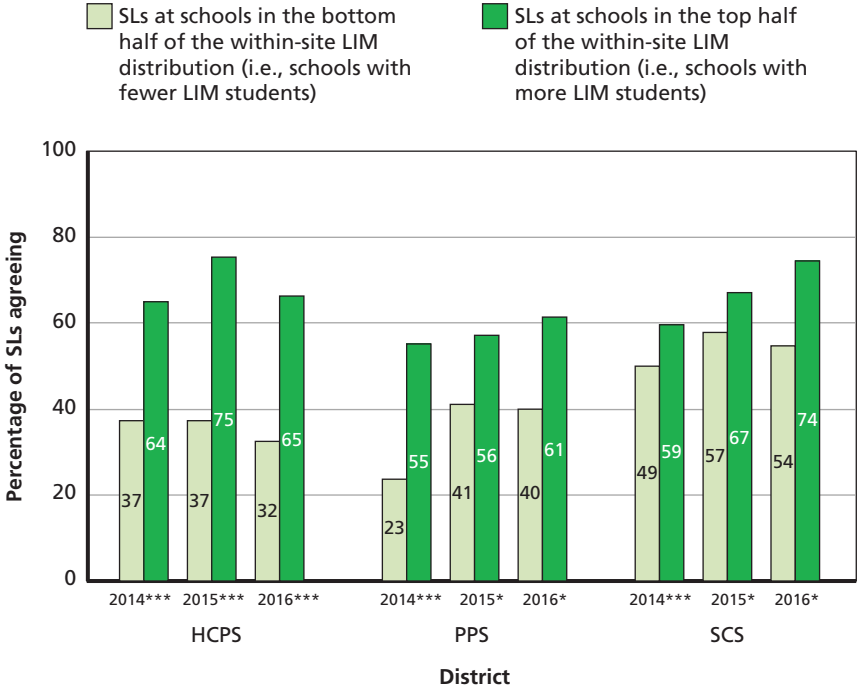
offers." And in our interviews with Aspire teachers, those in the notoriously expensive San Francisco Bay area frequently complained about the high cost of living and the difficulty of living on their salaries.

In the three districts, SLs at schools with higher percentages of LIM students (the top half of the within-site school-LIM distribution) were much more likely to agree than leaders in schools with lower percentages of LIM students that their school frequently lost good teachers to better opportunities. (Figure 11.5 presents results for 2014, 2015, and 2016.) There is considerably less LIM variation in the percentage of LIM students across schools within the CMOs than there is in the districts, so the top-half/bottom-half comparisons are less stable. As a result, we exclude the CMOs from Figure 11.5.

Changes in Retention

We also measured retention directly by examining whether teachers in one year continued to work as teachers in the same site the subsequent year and whether this varied by TE. We have two ways of characterizing effectiveness. For the majority of teachers in each district and CMO, we have the composite measure of TE that the sites provided after the beginning of the IP initiative. The study-calculated VAM measure is available only for mathematics and ELA teachers in grades 4 through 8 in the districts and Aspire but is available in years prior to the IP initiative. For PUC, we received composite TE level information for only two years, and we did not have VAM scores, so we could not compare trends in retention by TE. To make the analyses comparable across sites, we standardized the number of levels based on composite TE scores. Composite TE was categorized into four or five levels depending on the site, and we collapsed each of these into three levels: low, middle, and high. In reviewing our analyses, it is also important to remember that the number of teachers is not evenly distributed across composite TE levels; typically, the lowest level contains the fewest teachers. For our analysis of retention by VAM scores, we separated VAM scores into three levels: bottom 20 percent, middle 60 percent, and top 20 percent.

Figure 11.5
District SLs' Agreement That, "More Often Than Is Good for My School, Good Teachers Leave My Staff Because They Perceive Better Opportunities Elsewhere," by School's Percentage of Students with LIM Status, Springs 2014–2016



NOTE: * denotes that the difference between bottom-half and top-half LIM percentages is significant at $p < 0.05$. *** denotes that the difference is significant at $p < 0.001$. We have combined the original response categories of "agree strongly" and "agree somewhat." Omitted response categories are "disagree somewhat" and "disagree strongly."

RAND *RR1295-11.5*

The estimates we present in Figures 11.6 through 11.27 result from modeling teacher retention (whether the teacher teaches the following year at any school in the site) as a function of TE (measured in terms of the site's composite TE score or the study-calculated VAM score) in each time period, controlling for the teacher's age, teaching experi-

ence, educational attainment, gender, and race.[7] (In Appendix O, we describe these methods.) The time periods that we analyze here break the study period into three periods: pre-IP up through 2009–2010, early IP between 2010–2011 and 2012–2013, and late IP from 2013–2014 onward. Our main analysis focuses on these period-by-period changes in order to assess the effects of various retention policies implemented during the IP initiative. We also analyzed the relationships by year (described in Appendix P) to assess whether retention changed on an annual basis. Although the two models (by initiative period and by year) are not identical, the results of the initiative period model presented here generally reflect an aggregation of the results of the year model presented in Appendix P. Additionally, Appendix P shows the results of a sensitivity analysis in which we model teacher retention in terms of two consecutive low effectiveness ratings. Generally, sites allowed teachers to receive two low annual effectiveness ratings and to undertake an improvement process before formally dismissing them, although a site might have taken steps to counsel low-performing teachers out of the profession before taking formal dismissal actions. The analyses of retention rates following two consecutive years of effectiveness ratings more closely mimics the policies in the sites.

For each site except PUC, we present information about the number of teachers included in the analyses, followed by information about the average likelihood that teachers in each composite TE or VAM score category would be retained the following year, controlling for the aforementioned characteristics. Because of differences in the number of teachers in each site and the number of teachers at each composite TE level within a site, the confidence intervals and the strength of the comparisons differ across sites and levels.

---

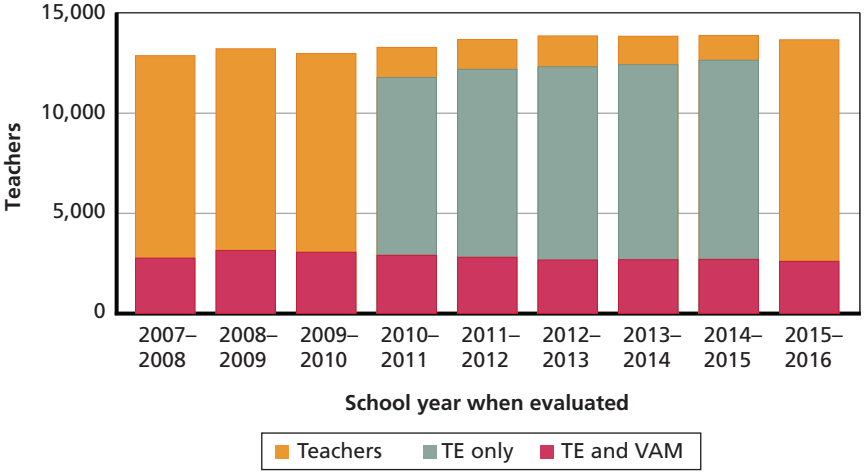
[7] We use a linear probability model, which avoids bias introduced by model misspecification (i.e., arbitrarily assigning a distribution to the error terms). Although previous studies have documented the minimal differences between the marginal effects estimated by linear probability models and other, more-parametric models (e.g., logistic), the downside of the linear probability model is that it can generate predicted values outside the boundaries of the limited dependent variable. See Angrist and Pischke, 2008, for a discussion of the advantages and disadvantages of the different approaches.

HCPS's Changes in Teacher Retention

Number of HCPS Teachers and Distribution of Teachers by TE Level

HCPS had the largest number of teachers among the IP sites between 2007–2008 and 2016–2017, leading to the most-precise estimates. Figure 11.6 shows the total number of teachers, the number with *both* a district composite TE score and a study-calculated VAM score, and those with only a study-calculated TE score, by year.[8] In Figure 11.7, we show the distribution of composite TE levels. Composite TE scores in HCPS range from 1 to 5; we grouped together the bottom two score categories and the top two score categories to obtain three levels. We also categorized the VAM score into three levels: bottom 20 percent, middle 60 percent, and top 20 percent. Although the distribution of the VAM categories is fixed by definition and does not vary from year

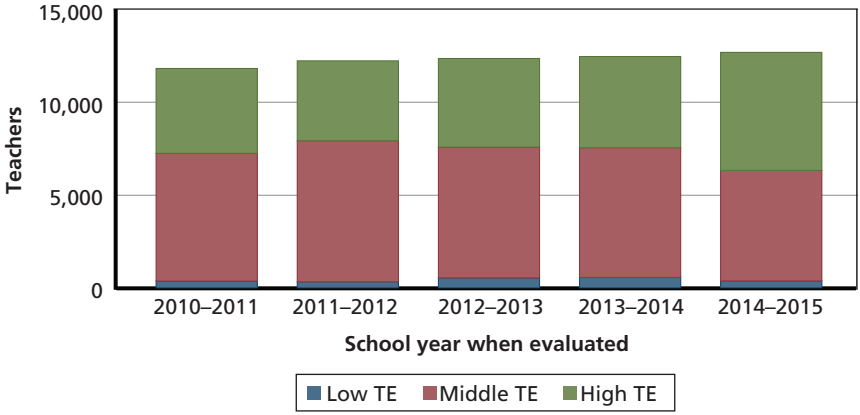
Figure 11.6
Numbers of HCPS Teachers and of Those with District Composite TE Levels or Study-Calculated VAM Scores



RAND *RR2242-11.6*

[8] For these analyses, we define the total number of teachers by the number of district employees associated with any course with a nonzero number of students. This number is greater than the number of district employees who had the job title of "classroom teacher," the distinction we used when developing the survey sample.

Figure 11.7
HCPS Teachers with District Composite TE Levels, by Composite TE Level



NOTE: Low TE = U or NI, middle TE = E, and high TE = HE level 4 or HE level 5.
RAND *RR2242-11.7*

to year, the number of teachers in each composite TE level changed over time, depending on the ratings the teachers actually received. In HCPS, the number of teachers in each level remained relatively stable from 2010–2011 through 2013–2014, but, in 2014–2015, the number of high-TE teachers increased, while the number of middle-TE teachers decreased.

Retention of Effective HCPS Teachers

Using either measure, HE HCPS teachers were more likely to remain in teaching than less effective teachers; however, the retention of HE teachers did not increase over time. The retention of HCPS teachers who scored low on either the composite TE measure or the VAM measure declined over time. We compared the likelihood of retention of HCPS teachers by year based on their effectiveness levels (both site composite TE levels and study-calculated VAM scores), after controlling for relevant characteristics. Figure 11.8 shows

How to Read the Retention Figures

This chapter includes several figures depicting change over time in the rate of teacher retention by composite TE or VAM level. The following is a guide to how to read the figure:

- *Retention* is based on the teacher working the following year (e.g., teachers evaluated during the 2012–2013 school year would be tracked at the start of the 2013–2014 school year).
- *The horizontal axis* indicates the period when teachers were evaluated (pre-IP up through 2009–2010, early IP between 2010–2011 and 2012–2013, and late IP from 2013–2014 onward).
- *The vertical axis* indicates the percentage of teachers who are likely to remain teaching in the site from one year to the next, conditional on site average demographic and educational information included in the model.
- *The blue, red, and green bars* depict the low, middle, and high composite TE levels (or, for VAM levels, bottom, middle, and top), respectively. Error bars show the 95-percent confidence intervals.
- *The values of the estimates* and their standard errors are available in Tables O.1 and O.2 in Appendix O.

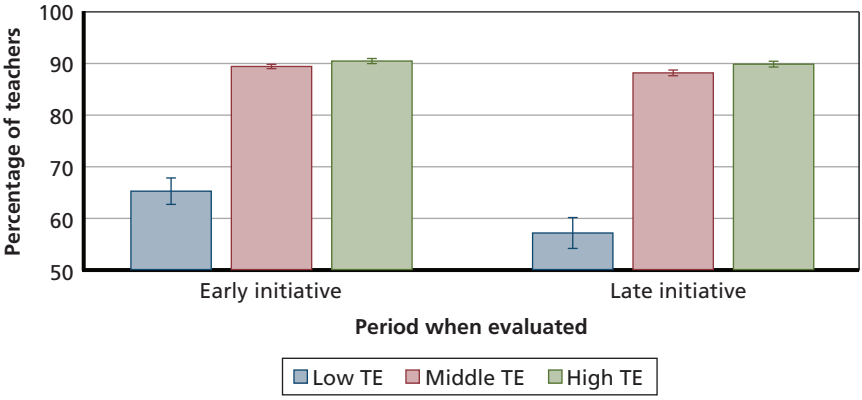
NOTE: The text accompanying the figures indicates whether there was a significant change in retention by TE or VAM level during the IP initiative. Differences can be significant, even if confidence intervals overlap slightly.

retention in the early and late IP periods by composite TE level.[9] We observe that low-TE teachers were significantly less likely to remain teaching than middle- and high-TE teachers in each period. Additionally, the results show that, compared with low-TE teachers in the early IP period, low-TE teachers were significantly less likely to remain teaching in the late IP period. However, the likelihood that high-TE teachers would remain teaching did not significantly differ from the early to the late IP period.

Figure 11.9 shows retention in the pre-, early, and late IP periods by VAM level. By VAM level, we also observe that low-VAM teachers were less likely to remain teaching than middle- and high-VAM teachers (the difference is significant in the early and late IP periods). Also, the results show that, compared with those in the pre-IP period,

[9] We emphasize the analyses using composite TE scores by presenting the results for these analyses ahead of the figures for the VAM scores because more teachers had composite TE scores than VAM scores, so we base these results on a larger sample. In contrast, in Chapter Ten, we emphasized the VAM scores because VAM scores are not subject to the sort of drift or inflation we observed in composite TE scores, and drift or inflation reduces the interpretability of trends. Consequently, in Chapter Ten, the VAM scores come first, and, in this chapter, the composite TE results come first.

Figure 11.8
Teachers Remaining in HCPS, by Period and by Composite TE Level



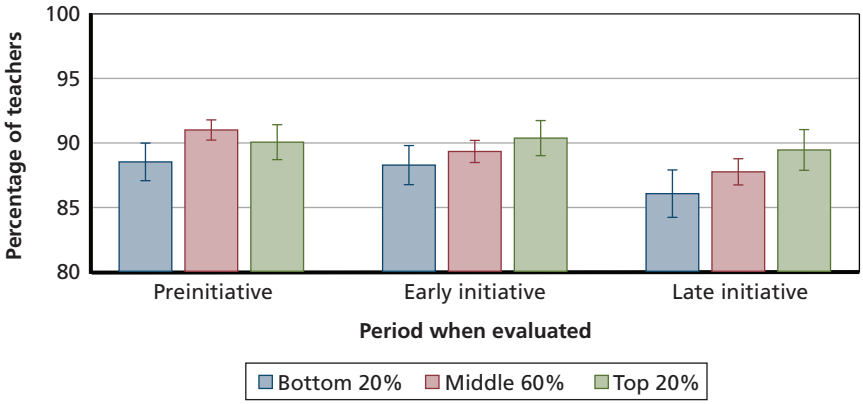
NOTE: Low TE = U or NI, middle TE = E, and high TE = HE 4 or HE 5. For any given year, we have classified every teacher as either remaining as a teacher in the district the following year or not. Error bars show 95-percent confidence intervals; estimates control for teacher characteristics. The TE measure is available beginning in 2010–2011. We have adjusted percentages for teacher characteristics, including gender, experience level, and education level, using multiple regression. Therefore, differences between percentages can be associated with year or effectiveness level rather than these characteristics. The values of the estimates and their standard errors are available in Table O.1 in Appendix O.

RAND *RR2242-11.8*

low- and middle-VAM teachers were significantly less likely to remain teaching in the late IP period. However, for high-VAM teachers, the differences in retention rates by period are not significantly different.

In general, the sites did not dismiss teachers based on low performance in a single year; a more common policy was to subject teachers to greater scrutiny and perhaps dismissal if they had two consecutive years with low composite TE ratings. We conducted a second analysis that more closely mimicked this policy by classifying teachers into levels based on two consecutive years of composite TE scores or two consecutive years of VAM scores. In Appendix P, we describe the methods. Requiring two years of composite TE or VAM scores restricts the sample, so the estimates are not as precise and we are less likely to detect changes over time. In fact, we could not perform these analyses on the CMOs because the samples were too small. The results of the two-

Figure 11.9
Teachers Remaining in HCPS, by Period and VAM Score



NOTE: For any given year, we have classified every teacher as either remaining as a teacher in the district the following year or not. Error bars show 95-percent confidence intervals; estimates control for teacher characteristics. We have adjusted percentages for teacher characteristics, including gender, experience level, and education level, using multiple regression. Therefore, differences between percentages can be associated with year or VAM score rather than these characteristics. The values of these estimates and their standard errors are available in Table O.2 in Appendix O.

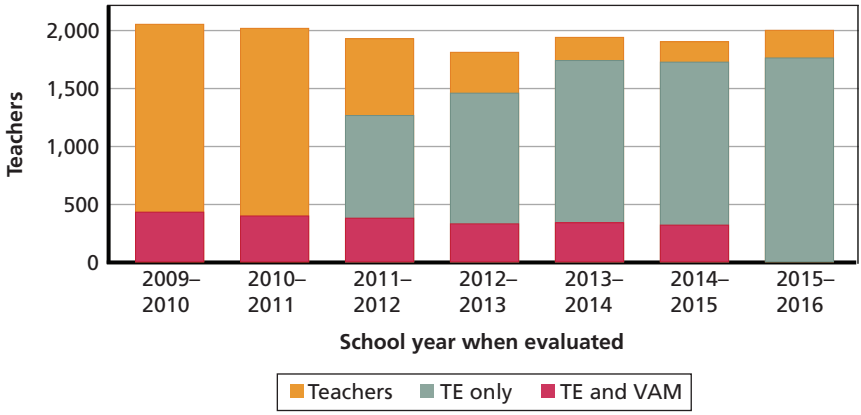
RAND *RR2242-11.9*

year analyses in HCPS were similar to the results presented above. The most-effective teachers were significantly more likely to remain teaching than the least effective teachers, but their likelihood of remaining in teaching did not increase over time (see Figure P.7 in Appendix P). As expected and similar to the results above, when we classified teachers based on two years of effectiveness data, the likelihood that the least effective teachers would remain in teaching gradually decreased over time. In Chapter Five, we more fully describe the use of effectiveness scores for dismissal decisions.

PPS Changes in Teacher Retention

In this section, we discuss the results of our analysis of retention in PPS. PPS had the fewest teachers of the IP districts, so the estimates are the least precise. Figures 11.10 and 11.11 show the introduction of composite TE scores in 2011–2012 and their rapid expansion there-

Figure 11.10
Numbers of PPS Teachers and of Those with District Composite TE Levels or Study-Calculated VAM Scores

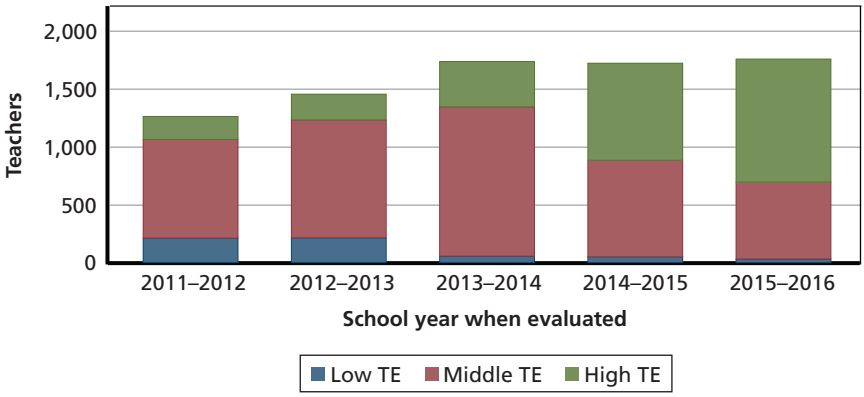


NOTE: We based the distribution of ratings for 2011–2012 on a pilot version of PPS's composite measure that was never shared with teachers or SLs.
RAND *RR2242-11.10*

after; they also show that we were able to calculate VAM scores on only a small fraction of teachers in PPS.[10] PPS uses four performance levels, and we grouped together the bottom two levels to obtain three levels; Figure 11.11 shows the change over time in the distribution of composite TE levels. As described in Chapter Four, the distribution of composite TE dramatically changed in PPS beginning in the 2013–2014 school year. The initial composite TE measure used in PPS, called RISE, was piloted in 2009–2010, and the results were available to teachers, principals, and central-office staff in 2010–2011. In the 2011–2012 and 2012–2013 school years, the RISE observation measure was the sole measure used for personnel decisions. PPS began

[10] We include as a teacher anyone who was the staff person of record at the end of the year for a course in which students were registered. Because of changes in personnel during the year or because of other district requirements, not everyone that we define as a teacher is included in the district's definition of *teacher* for the purposes of teacher evaluation; thus, not every teacher has a composite TE score. We use a definition based on course records rather than composite TE score because we can apply it consistently throughout this extended period of time.

Figure 11.11
PPS Teachers with District Composite TE Levels, by TE Level



NOTE: Low TE = F or NI, middle TE = P, and high TE = D. We based the distribution of ratings for 2011–2012 on a pilot version of PPS's composite measure that was never shared with teachers or SLs.

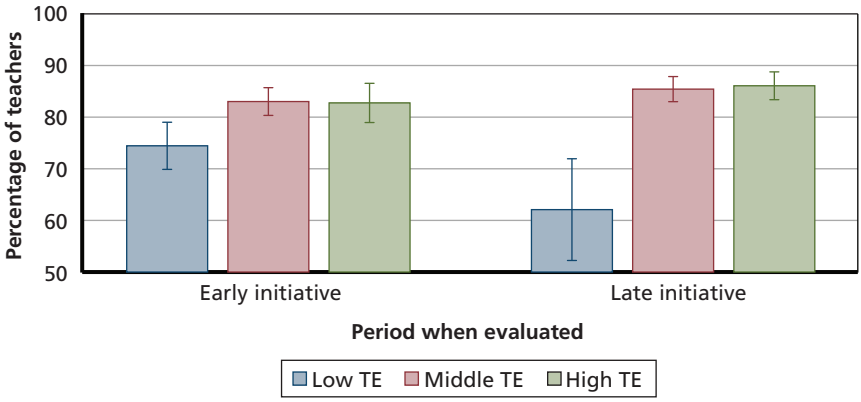
RAND *RR2242-11.11*

pilot-testing a composite measure, of which RISE was one component, in 2010–2011, and it became the measure used for personnel decisions beginning in the 2013–2014 school year. With the introduction of the composite measure in the 2013–2014 school year, the number of teachers in the high-TE categories rose, while the number of low- and middle-TE teachers fell.

Each year, using either measure, we found that HE PPS teachers were more likely than less effective teachers to remain teaching in the site; when measured by composite TE, the retention of less effective teachers decreased over time and the retention of HE teachers increased.[11] Figure 11.12 shows retention in the early and

[11] As mentioned in the "How to Read the Retention Figures" box above, the difference between two estimates can be significant even if the confidence intervals of the two estimates overlap. If the correlation of the two estimates is equal to –1, the confidence intervals will not overlap only if there is a significant difference in the two estimates. For any other correlation between the two estimates, there will be some values of the estimates for which the confidence intervals overlap but there is still a significant difference between the estimates. Therefore, although we present confidence intervals in Figures 11.8, 11.9, 11.12, 11.13, 11.16,

Figure 11.12
Teachers Remaining in PPS, by Period and Composite TE Level



NOTE: Low TE = U or NI, middle TE = E, and high TE = HE 4 or HE 5. For any given year, we have classified every teacher as either remaining as a teacher in the district the following year or not. Error bars show 95-percent confidence intervals; estimates control for teacher characteristics. The TE measure is available beginning in 2010–2011. We have adjusted percentages for teacher characteristics, including gender, experience level, and education level, using multiple regression. Therefore, differences between percentages can be associated with year or effectiveness level rather than these characteristics. The values of the estimates and their standard errors are available in Table O.1 in Appendix O.

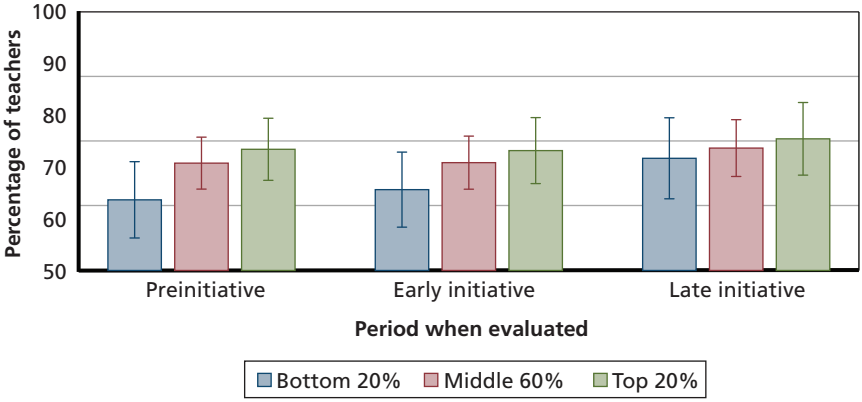
RAND *RR2242-11.12*

late IP periods by composite TE level. We observe that low-TE teachers were significantly less likely than middle- or high-TE teachers in each period to remain teaching. The results also show that the likelihood that low-TE teachers would remain teaching significantly decreased. In contrast, the likelihood that middle- and high-TE teachers would remain teaching was significantly higher in the late IP period.

Figure 11.13 shows retention in the pre-, early, and late IP periods by VAM level. First, we observe that the middle- and high-VAM teachers were significantly more likely to remain teaching than low-VAM teachers in the pre- and early IP periods. However, in the late

11.17, 11.20, 11.23, 11.24, and 11.27 to provide information about the precision of the estimates, we include explicit tests of the change in retention over time in Tables O.1 and O.2 in Appendix O.

Figure 11.13
Teachers Remaining in PPS, by Period and VAM Score



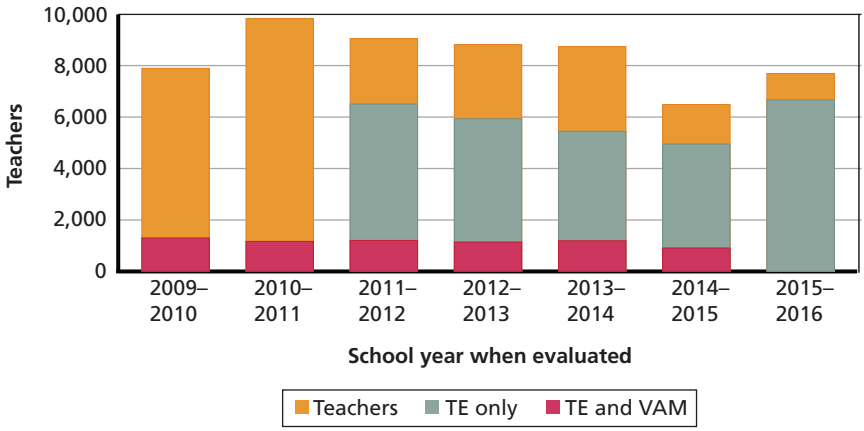
NOTE: For any given year, we have classified every teacher as either remaining as a teacher in the district the following year or not. Error bars show 95-percent confidence intervals; estimates control for teacher characteristics. We have adjusted percentages for teacher characteristics, including gender, experience level, and education level, using multiple regression. Therefore, differences between percentages can be associated with year or VAM score rather than these characteristics. The values of these estimates and their standard errors are available in Table O.2 in Appendix O.

RAND *RR2242-11.13*

IP period, there is no statistically significant difference between low-, middle-, and high-VAM teachers in the likelihood of retention. Additionally, the likelihood of retention among low-VAM teachers increased in the early and late IP periods. However, compared with that of the pre-IP period, the change in the likelihood of retention in the early IP period is not statistically significant, and the change from the pre-IP to the late IP period is significant only at $p < 0.10$. Finally, the results show that the likelihood of retention for middle- and high-VAM teachers did not significantly change across the three periods.

We did not see major differences when we classified teachers on the basis of two years of composite TE scores or two years of VAM estimates. (Figure P.8 in Appendix P presents these results.)

Figure 11.14
Numbers of SCS Teachers and of Those with District Composite TE or Study-Calculated VAM Scores



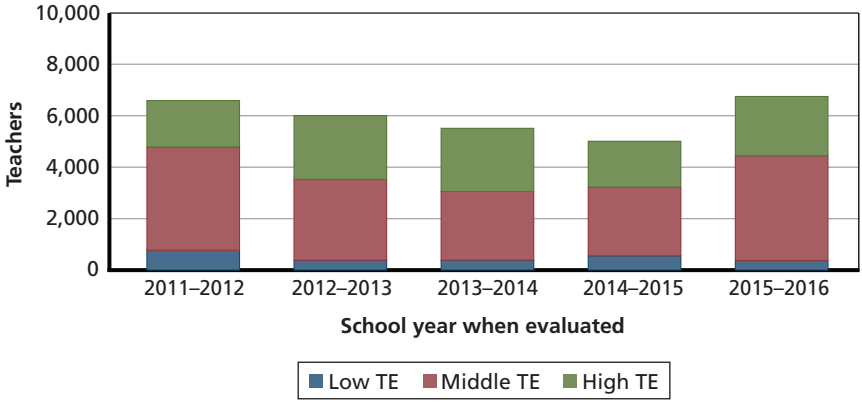
RAND *RR2242-11.14*

SCS Changes in Teacher Retention

Figure 11.14 shows, by year, the total number of SCS teachers, the number of teachers with composite TE and VAM scores, and the number with only TE scores; Figure 11.15 shows the distribution of composite TE levels by year. The number of teachers in our analysis shows a large increase in 2010–2011 and a large drop in 2014–2015. We tried to insulate our analyses from these large changes in the teacher workforce by including only teachers and schools that had been part of legacy MCS. However, even this restriction did not completely negate changes arising from the merger because legacy-MCS teachers who transferred to newly added SCS schools would appear in our analysis as separations, not retentions, as would legacy-MCS teachers who remained in schools that were transferred to the ASD.

Composite TE scores in SCS range from 1 to 5, and we grouped together both the bottom two score categories and the top two score categories to obtain three levels. The number of high-TE teachers grew during the first three years shown and then shrunk a bit; the opposite was roughly true for low-TE teachers. The greatest change was in the

Figure 11.15
SCS Teachers with District Composite TE Levels, by TE Level



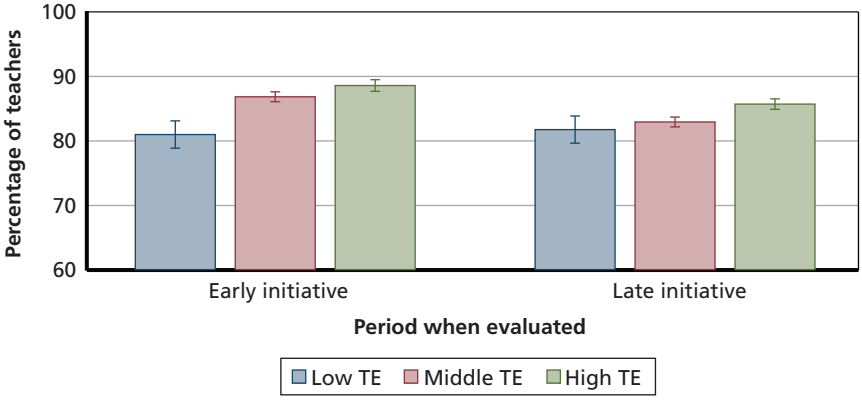
NOTE: Low TE = below or significantly below expectations, middle TE = meeting expectations, and high TE = above or significantly above expectations.
RAND *RR2242-11.15*

number of middle-TE teachers, which decreased between 2011–2012 and 2014–2015 and then increased in 2015–2016.

By either measure, HE teachers were more likely than less effective teachers to remain teaching in SCS after the start of the IP initiative; when we categorized them by VAM score, retention decreased for all SCS teachers during the initiative, but it decreased the most for the lowest-VAM teachers. Figure 11.16 shows retention in the early and late IP periods by composite TE level. Like in HCPS and in PPS, high-TE teachers were significantly more likely than low-TE teachers in both the early and late IP periods to remain teaching. Retention decreased for middle-TE and high-TE teachers in SCS during the initiative; it also decreased for the lowest-VAM teachers.

Figure 11.17 shows retention in the pre-, early, and late IP periods by VAM level. By VAM level, we do not observe any significant difference in the likelihood of retention between low–, middle–, and high–VAM level teachers in the pre-IP period. However, in the early and late IP periods, high-VAM teachers were significantly more likely than low-VAM teachers to remain teaching. Over time, we observe a general decline in the likelihood of remaining in teaching at SCS for

Figure 11.16
Teachers Remaining in SCS, by Period and by District Composite TE Level



NOTE: Low TE = U or NI, middle TE = E, and high TE = HE 4 or HE 5. For any given year, we have classified every teacher as either remaining as a teacher in the district the following year or not. Error bars show 95-percent confidence intervals; estimates control for teacher characteristics. The TE measure is available beginning in 2010–2011. We have adjusted percentages for teacher characteristics, including gender, experience level, and education level, using multiple regression. Therefore, differences between percentages can be associated with year or effectiveness level rather than these characteristics. The values of the estimates and their standard errors are available in Table O.1 in Appendix O.

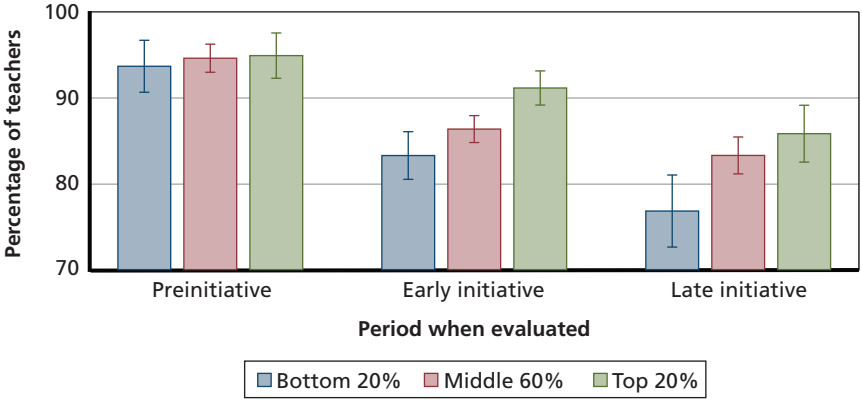
RAND *RR2242-11.16*

each VAM level, with the decline being the largest for the lower-VAM teachers. Compared with those in the pre-IP period, low-, middle-, and high-VAM teachers were all significantly less likely to remain teaching in the early IP period. And compared with those in the early IP period, low-, middle-, and high-VAM teachers were significantly less likely to remain teaching in the late IP period.

Alliance Changes in Teacher Retention

Figure 11.18 shows the total number of teachers and the number of teachers with composite TE scores by year. Because of the relatively small number of teachers in the site (and the even smaller number who taught in tested subjects and grade levels), we did not calculate VAM scores for any teacher in Alliance. The total number of teachers in Alliance grew rapidly following the 2010–2011 school year as

Figure 11.17
Teachers Remaining in SCS, by Period and VAM Score



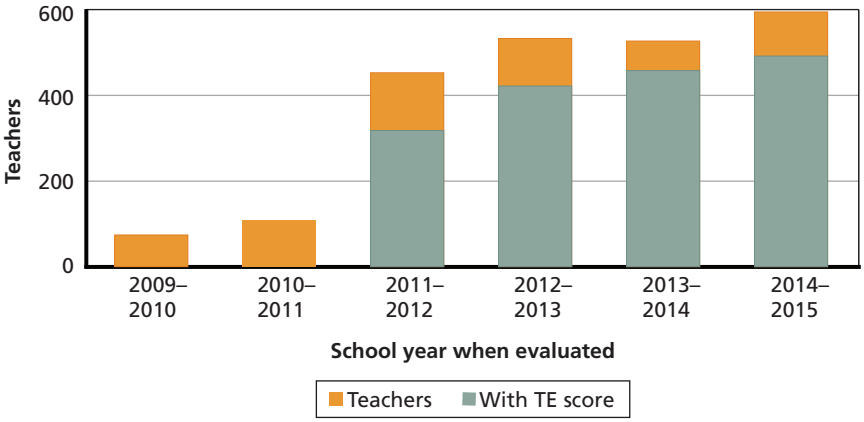
NOTE: For any given year, we have classified every teacher as either remaining as a teacher in the district the following year or not. Error bars show 95-percent confidence intervals; estimates control for teacher characteristics. We have adjusted percentages for teacher characteristics, including gender, experience level, and education level, using multiple regression. Therefore, differences between percentages can be associated with year or VAM score rather than these characteristics. The values of these estimates and their standard errors are available in Table O.2 in Appendix O.

RAND *RR2242-11.17*

Alliance opened new schools, and, by 2014–2015, there were approximately 600 teachers in the CMO. In Figure 11.19, we also show the distribution of composite TE levels. Composite TE scores in Alliance ranged from 1 to 5, and we grouped both the bottom two score categories and the top two score categories in order to obtain three levels. For Alliance, the majority of teachers fell into the middle-TE category; over time, the number of low-TE teachers decreased and the number of high-TE teachers increased.

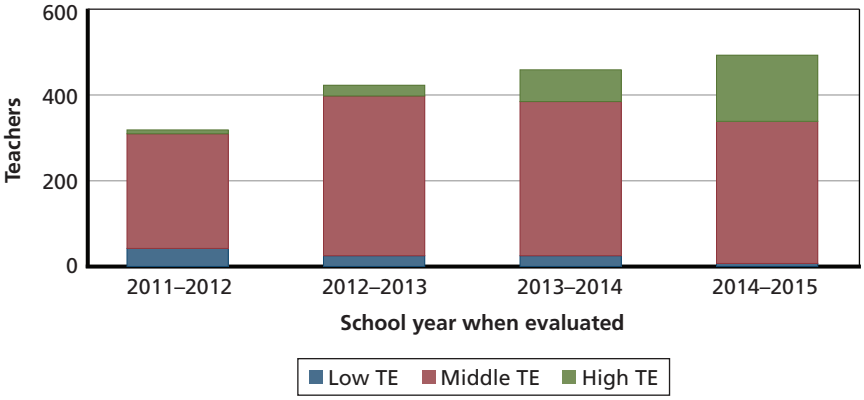
Each year, high-TE Alliance teachers were more likely than middle- or low-TE teachers to remain teaching; however, over time, the likelihood that low-TE and high-TE teachers would remain did not significantly change. Figure 11.20 shows retention in the early and late IP periods by composite TE level. We observe that high-TE teachers were significantly more likely than low- and middle-TE teachers to remain teaching in both the early and late IP periods.

Figure 11.18
Numbers of Alliance Teachers and of Those with CMO Composite TE Levels



RAND *RR2242-11.18*

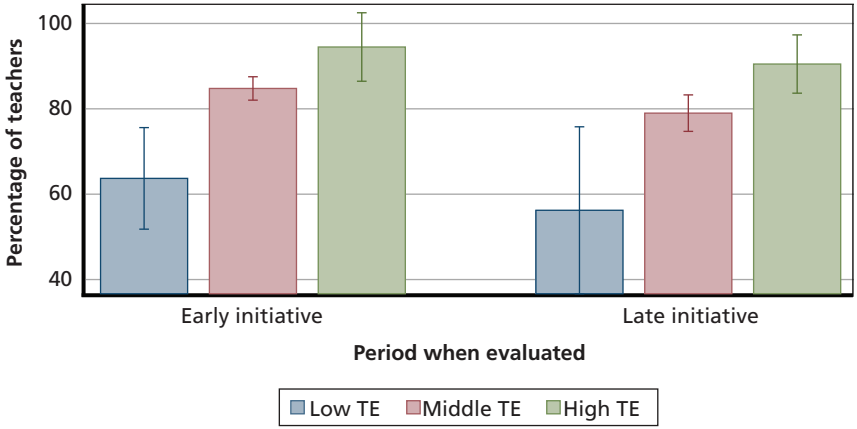
Figure 11.19
Alliance Teachers with CMO Composite TE Levels, by TE Level



NOTE: Low TE = entering or achieving, middle TE = E, and top TE = HE or master.
RAND *RR2242-11.19*

Also, the likelihood that middle-TE teachers would remain teaching significantly decreased in the late IP period. In contrast, the likelihood

Figure 11.20
Teachers Remaining in Alliance, by Period and by CMO Composite TE Level



NOTE: Low TE = entering or achieving, middle TE = E, and high TE = HE or master. For any given year, we have classified every teacher as either remaining as a teacher in the district the following year or not. Error bars show 95-percent confidence intervals; estimates control for teacher characteristics. The TE measure is available beginning in 2011–2012. We have adjusted percentages for teacher characteristics, including gender, experience level, and education level, using multiple regression. Therefore, differences between percentages can be associated with year, effectiveness level, or level of value added rather than these characteristics. The values of the estimates and their standard errors are available in Tables O.1 and O.2 in Appendix O.

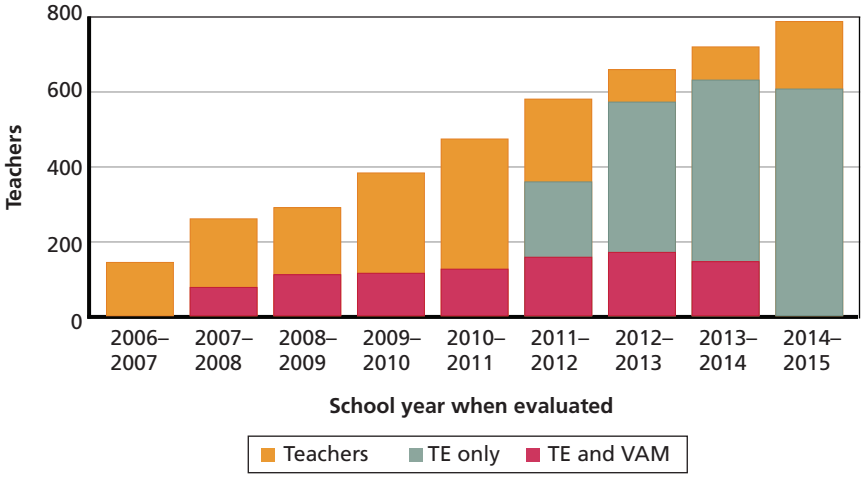
RAND *RR2242-11.20*

that low- and high-TE teachers would remain teaching does not significantly change across periods.

Aspire Changes in Teacher Retention

Figure 11.21 shows the total number of Aspire teachers, the number of teachers with composite TE scores, and the number of teachers with both TE and VAM scores, by year. Composite TE scores are available from 2011–2012 through 2014–2015. Aspire is the only CMO for which we could calculate VAM scores, and we could do so for each year from 2007–2008 through 2013–2014. Like in Alliance, the total number of teachers in Aspire rapidly grew over time as the CMO expanded its number of schools. Figure 11.22 shows the distribution of

Figure 11.21
Numbers of Aspire Teachers and of Those with CMO Composite TE Levels or Study-Calculated VAM Scores

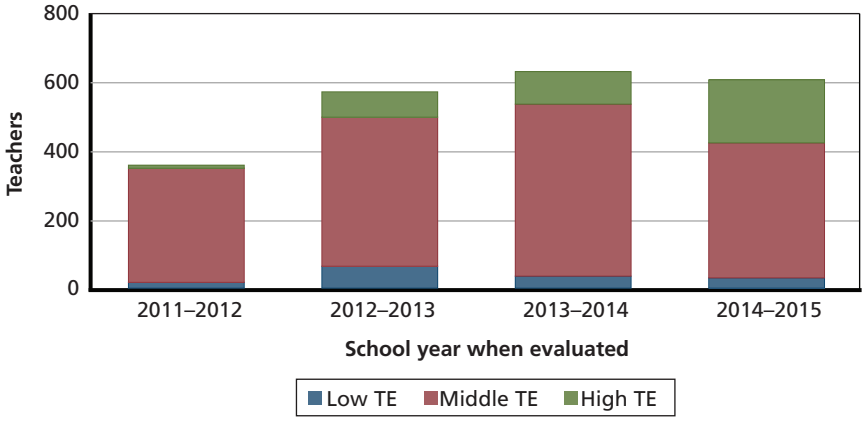


RAND *RR2242-11.21*

composite TE levels. Composite TE scores in Aspire range from 1 to 5, and we grouped both the bottom two score categories and the top two score categories in order to obtain three levels. The majority of teachers fell into the middle–composite TE category, and, over time, the number of high-TE teachers increased.

Middle-TE and high-TE teachers in Aspire were more likely than low-TE teachers to remain teaching, although the differences were not always statistically significant. There were no statistically significant changes over time in the likelihood that high-TE teachers would remain teaching in Aspire, but low- and high-VAM teachers were significantly less likely to remain teaching over time. Figure 11.23 shows retention in the early and late IP periods by composite TE level. First, we note that the likelihood that low- and high-TE teachers would remain teaching in the late IP period did not significantly change from that in the early IP period. Additionally, the likelihood that low- and high-TE teachers would remain teaching in the late IP period did not significantly change from that in the early IP

Figure 11.22
Aspire Teachers with CMO Composite TE Levels, by TE Level



NOTE: Low TE = entering or achieving, middle TE = E, and top TE = HE or master.
RAND *RR2242-11.22*

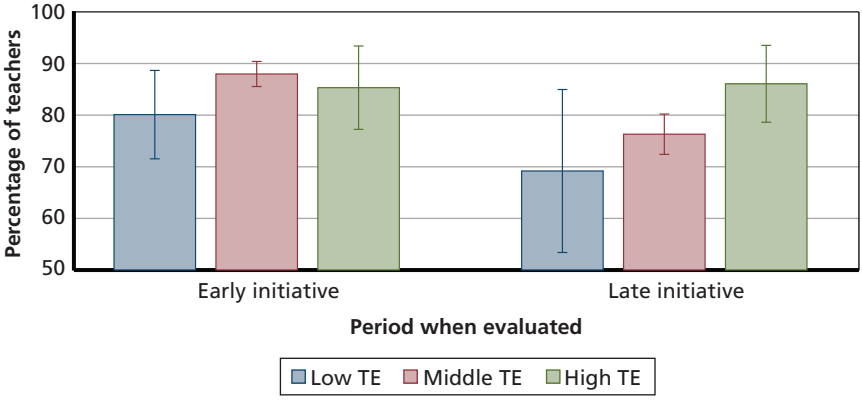
period. However, the likelihood that middle-TE teachers would remain teaching significantly decreased in the late IP period.

Figure 11.24 shows retention in the pre-, early, and late IP periods by VAM level. Only in the early IP period were high-VAM teachers significantly more likely than low-VAM teachers to remain teaching. Also, we observe a general decline in retention likelihood: Compared with those in the pre-IP period, low-, middle-, and high-VAM teachers were all significantly less likely to remain teaching in the late IP period.

Green Dot Changes in Teacher Retention

Figure 11.25 shows the total number of Green Dot teachers and the number of teachers with composite TE scores, by year. Composite TE scores are available from 2011–2012 through 2014–2015. Because of the relatively small number of teachers in the site (and the even smaller number teaching tested subjects and grade levels), we did not calculate VAM scores for any teachers in Green Dot. Like in the other CMOs, the total number of teachers in Green Dot rapidly grew over time as the CMO opened new schools. Figure 11.26 shows the distribution of composite TE levels. Composite TE scores in Green Dot range from 1

Figure 11.23
Teachers Remaining in Aspire, by Period and by Composite TE Level



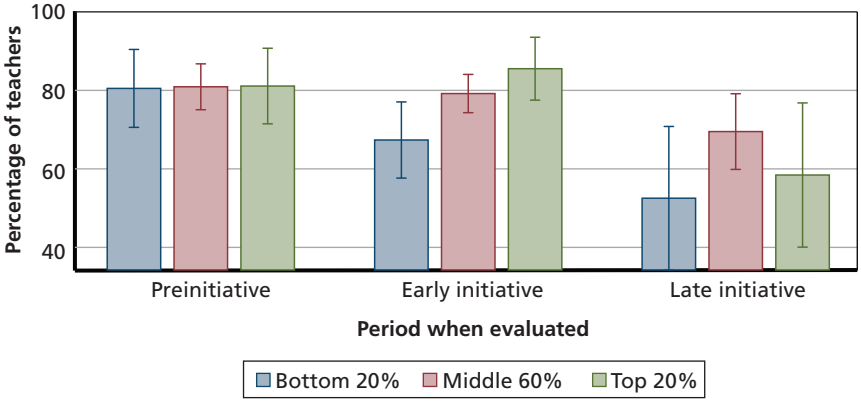
NOTE: Low TE = U or NI, middle TE = E, and high TE = HE 4 or HE 5. For any given year, we have classified every teacher as either remaining as a teacher in the district the following year or not. Error bars show 95-percent confidence intervals; estimates control for teacher characteristics. The TE measure is available beginning in 2010–2011. We have adjusted percentages for teacher characteristics, including gender, experience level, and education level, using multiple regression. Therefore, differences between percentages can be associated with year or effectiveness level rather than these characteristics. The values of the estimates and their standard errors are available in Table O.1 in Appendix O.

RAND *RR2242-11.23*

to 5, and we grouped both the bottom two score categories and the top two score categories in order to obtain three levels. In 2011–2012, we categorized the majority of teachers as middle composite TE, but, over time, the number of high-TE teachers increased, and, by 2014–2015, more teachers fell into the high-TE category than the middle-TE category. The number of low-TE teachers was small and grew smaller over time. These changes are likely related to a change in the calculation of composite TE: Beginning in 2013–2014, student achievement was not included in the composite TE measure.

More-effective teachers in Green Dot were more likely than less effective teachers to remain teaching. There were no statistically significant changes over time in the likelihood that low- and high-TE teachers would remain teaching in Green Dot. Figure 11.27 shows retention in the early and late IP periods by com-

Figure 11.24
Teachers Remaining in Aspire, by Period and VAM Score

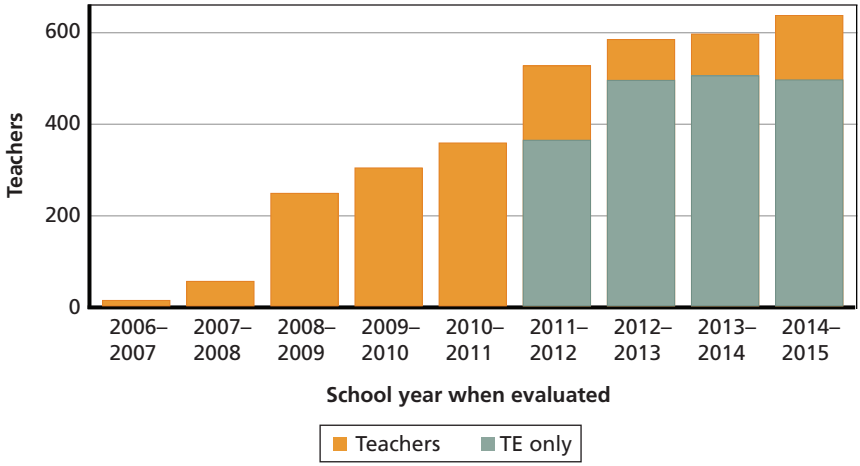


NOTE: For any given year, we have classified every teacher as either remaining as a teacher in the district the following year or not. Error bars show 95-percent confidence intervals; estimates control for teacher characteristics. We have adjusted percentages for teacher characteristics, including gender, experience level, and education level, using multiple regression. Therefore, differences between percentages can be associated with year or VAM score rather than these characteristics. The values of these estimates and their standard errors are available in Table O.2 in Appendix O.

RAND *RR2242-11.24*

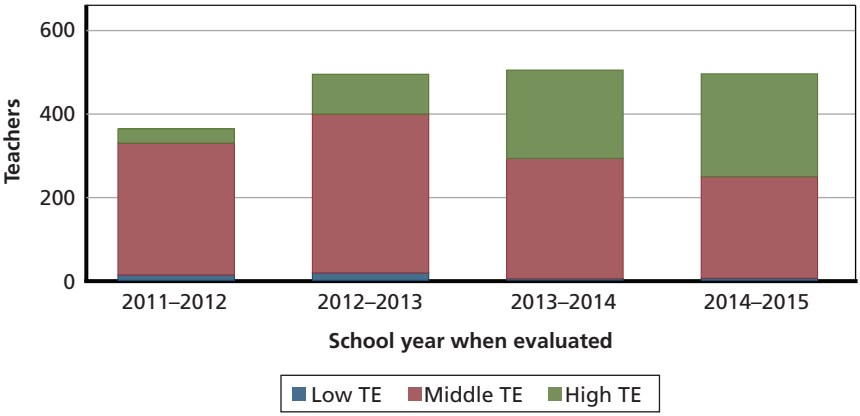
posite TE level. In the early IP period, middle- and high-TE teachers were significantly more likely than low-TE teachers to remain teaching. In the late IP period, both middle- and high-TE teachers were more likely than low-TE teachers to remain teaching, but, because of the small number of teachers in the low-TE category and resulting wide confidence intervals, this difference is not significant. Although the likelihood that middle-TE teachers would remain in teaching significantly decreased in the late IP period, the likelihood of retention of low- and high-TE teachers did not significantly differ.

Figure 11.25
Numbers of Green Dot Teachers and of Those with CMO Composite TE Levels



RAND *RR2242-11.25*

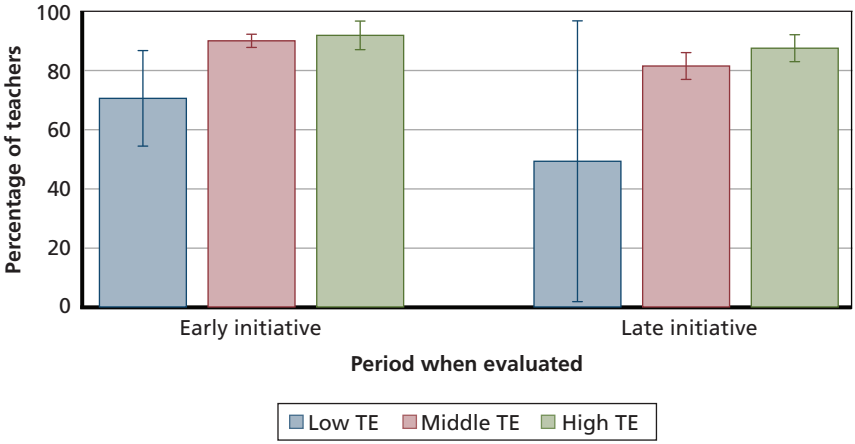
Figure 11.26
Green Dot Teachers with CMO Composite TE Levels, by TE Level



NOTE: Low TE = entering or achieving, middle TE = E, and high TE = HE or master.

RAND *RR2242-11.26*

Figure 11.27
Teachers Remaining in Green Dot, by Period and by Composite TE Level



NOTE: Low TE = entering or achieving, middle TE = E, and high TE = HE or master. For any given year, we have classified every teacher as either remaining as a teacher in the district the following year or not. Error bars show 95-percent confidence intervals; estimates control for teacher characteristics. The TE measure is available beginning in 2011–2012. We have adjusted percentages for teacher characteristics, including gender, experience level, and education level, using multiple regression. Therefore, differences between percentages can be associated with year, effectiveness level, or level of value added rather than these characteristics. The values of the estimates and their standard errors are available in Tables O.1 and O.2 in Appendix O.

RAND *RR2242-11.27*

Summary

We drew on two sources of evidence related to teacher retention: survey results and empirical analyses of employment records. In most sites, less than half the teachers reported on our surveys that they would still be working as a teacher in their current site in five years' time, and this was true from 2011 through 2015. Many possible reasons could underlie teachers' plans to depart, but, whatever the reasons, such plans present a significant challenge for schools. In some sites, including HCPS and PPS, more-effective teachers and more-experienced teach-

ers reported that they were more likely to expect to continue teaching, but, in other sites, the opposite was true.

Analyses of teacher employment data in the three districts and three largest CMOs show that HE teachers were generally more likely than less effective teachers to remain teaching. However, as Table 11.1 summarizes, over time, the likelihood that HE teachers would remain in teaching did not increase in most sites despite changes in HR policies and practices. In fact, we were more likely to find statistically significant estimates of the opposite trend showing *decreases* in the retention of HE teachers. In HCPS, the retention of HE teachers changed minimally over time. In PPS, the retention of HE teachers improved over time when measured by the site's composite effectiveness measure but not when measured by the project-calculated VAM measure. In SCS, retention of high-VAM teachers decreased significantly over time. Both sample size and the availability of effectiveness measures limited our analysis of retention patterns in the CMOs. Considering these limitations, we found no clear evidence that the retention of HE teachers increased over time in Alliance. In Aspire, we observed a discrepancy between the changes observed by composite TE score and by VAM score. In Green Dot, we found no evidence that HE teachers were increasingly likely to remain teaching, but there is some (insignificant) evidence that less effective teachers became less likely to remain teaching over time.

On the other hand, as shown in Table 11.1, in HCPS and PPS, there was a decrease over time in the likelihood that low-TE teachers would remain teaching in the site. In addition, when classifying effectiveness in terms of VAM score rather than composite TE level, there was a decrease in the retention of low-VAM teachers in HCPS, SCS, and Aspire. Therefore, five of the six sites showed a decline in the retention of less effective teachers by one measure or the other. Notably, every statistically significant trend in the retention of less effective teachers was in the desired direction. During the IP initiative, sites enacted changes in policies related to dismissal, which could account for some of these changes. The increase in the number of policies designed to reward teachers on merit might also have induced less effective teachers to exit voluntarily.

Table 11.1
Change over Time in Retention of High-Performing and Low-Performing Teachers

	High-Performing Teachers		Low-Performing Teachers	
Site	**TE (Change from Early to Late Initiative)**	**VAM (Change from Pre- to Late Initiative)**	**TE (Change from Early to Late Initiative)**	**VAM (Change from Pre- to Late Initiative)**
HCPS	N/S	N/S	–	–
PPS	+	N/S	–	N/S
SCS	–	–	N/S	–
Alliance	N/S	N/A	N/S	N/A
Aspire	N/S	–	N/S	–
Green Dot	N/S	N/A	N/S	N/A

NOTE: + denotes positive and statistically significant at $p < 0.05$. – denotes negative and statistically significant at $p < 0.05$. N/S = not significant. N/A = not applicable because of data limitations.

Although policies designed in whole or in part to increase the retention of HE teachers did not appear to have achieved that goal during this period, the sites generally experienced a decrease in the retention of ineffective teachers.

CHAPTER TWELVE

The Initiative's Effects on TE and on LIM Students' Access to Effective Teaching

In Brief

Did the overall IP initiative improve the effectiveness of the teacher workforce, and did LIM students have greater access to effective teachers?

Overall, we found that TE (measured in terms of the study-calculated VAM) did not improve meaningfully in the three IP districts, except for reading teachers in SCS, who were much more effective at the end of the initiative. Mathematics teachers made small but significant improvements in HCPS and SCS, and there were moderate to small but significant declines in reading for HCPS and PPS. Nevertheless, most SLs reported that they had taken steps to assign effective teachers to students with the greatest needs, and, in HCPS and SCS, most reported that teacher evaluations were used to at least a moderate extent to assign teachers to students and classes. However, our analysis of student access to effective teaching found that, in 2015, LIM students had less access than non-LIM students to effective teachers overall, although that negative difference was statistically significant only for HCPS mathematics and reading and SCS reading. When we decomposed access into two components—within schools and between schools—we found that LIM students tended to have less access to effective teachers within their schools than they did between schools. Comparing access before the IP initiative with that after, there was no improvement in LIM students' access associated with the IP initiative, and, in some cases, it worsened. Finally, we looked for evidence of each of three possible mechanisms by which sites could, in theory, affect LIM students' access to effective teaching: (1) improving the effectiveness of teachers already working in high-LIM classrooms, (2) moving effective teachers to high-LIM

classrooms, and (3) increasing the number of LIM students in the classrooms of effective teachers. We found no evidence that any of the sites consistently employed any of the three mechanisms, which is consistent with the overall finding that LIM students did not have better access than non-LIM students to effective teaching.

Introduction

Two main objectives of the IP initiative were to improve TE overall and to promote LIM students' access to the most-effective teachers. The IP sites tried to accomplish these changes in a variety of ways, as described in previous chapters, including targeted PD to improve the effectiveness of current teachers, hiring and placing more-effective teachers in schools with many disadvantaged students, and encouraging effective teachers to continue teaching in these schools.

Other researchers have examined LIM students' relative access to effective teachers in other districts. Most notably, Isenberg et al. (2016) analyzed 26 school districts in the United States and found only small differences in the access that disadvantaged students had, compared with that of more-advantaged students, to effective teachers. Of course, achievement can improve for some or all students for reasons other than an increase in TE. Chapter Thirteen examines the impact of the initiative on achievement for all students and select subgroups.

This chapter presents the following information about the distribution of TE as we measured it with VAM scores calculated for the study:

- First, we present quantitative evidence about overall TE and the allocation of teachers and students. We describe trends over time in overall TE (i.e., did the teacher workforce in each site become more effective during the IP initiative?).
- Then we present information about the allocation of teachers and students, focusing on LIM students' access to effective teachers:
  - We describe SLs' responses to survey items about the ways in which students are assigned to classes. These responses offer the

principals' perspective on their efforts to consciously allocate LIM students to effective teachers.
- We describe trends in LIM students' access to effective teachers (i.e., we examine the associations between TE and the proportion of students in each teacher's classes and school who are LIM students) and examine how those relationships have changed over time.
- Finally, we explore the mechanisms through which LIM students' access to effective teachers might have changed over time.

Changes in the Distribution of TE

In this section, we look at the overall teacher workforce in each of the three districts, and we examine whether the distribution of TE, as measured by the study, improved during the IP initiative. We do not present the results for the CMOs because they have considerably fewer teachers and because we are missing a year of CMO test data from when California switched standardized tests in 2014. Additionally, most of the CMOs are overwhelmingly composed of LIM students. Given that the empirical models in this chapter are estimated based on variation in student-body LIM proportions, we would have had difficulty identifying differences in access even if we had sufficient numbers of teachers and scores for all years.

Our Measure of TE

We measure TE in terms of VAM—that is, an estimate of the increase in student test scores that can be attributed to teacher inputs rather than to other factors. In Appendix A, we describe the VAM method we used. We measure these VAM estimates in national standard deviations of student achievement.[1] Thus, if a teacher has a VAM of 0.1, this implies that a student assigned to this teacher will experience an

[1] We do so by rescaling the tests based on state and national NAEP scores, as explained in Appendix A, in the section "Estimation of Teacher Value Added."

increase in test scores of 0.1 standard deviations in the national student achievement distribution compared with what the student's performance would have been had he or she been assigned to the average teacher. Although standard deviations are the typical measure of VAM, they can be difficult to interpret. To aid interpretation, Table 12.1 shows the relationship between standard deviations in student achievement and the corresponding percentile growth from the median student.[2] For example, a VAM of 0.1 standard deviations corresponds to the given teacher increasing the average student performance 4 percentile points (i.e., from the 50th percentile to the 54th percentile) over what the average teacher would do (i.e., no change and remaining at the 50th percentile).

Table 12.1
Relationship Between Standard Deviations of Student Achievement and Percentile Growth from the Median Student

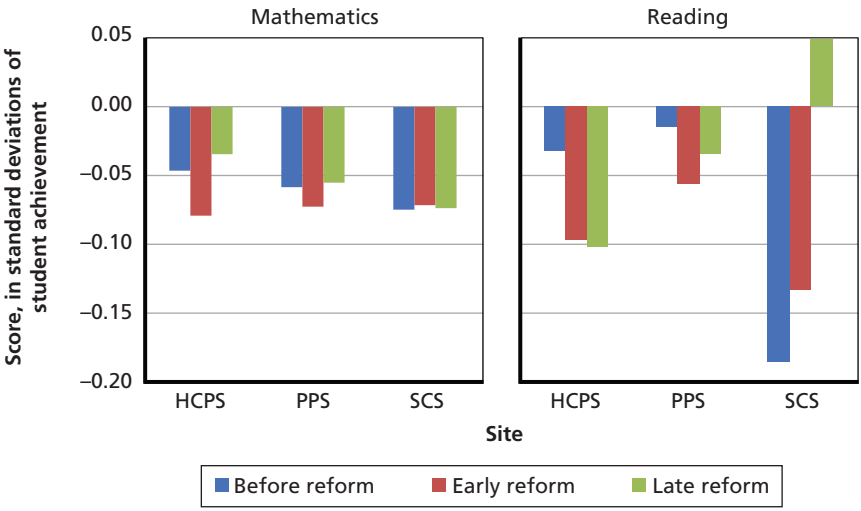
Standard Deviation	Percentile Growth from the Median
0.01	0.4
0.02	0.8
0.05	2.0
0.1	4.0
0.2	7.9
0.4	15.5
0.6	22.6
0.8	28.8
1	34.1

[2] Treatment effects in standard deviations of student achievement are converted to percentile gains at the mean by assuming a normal distribution, and the equation that percentile growth = 100 × (Φ(standard deviation) – 0.5), where Φ(•) is the standard normal cumulative density function.

Distribution of TE

We start by showing the change in overall teacher VAM for each site. We present the distributions of teacher value added during three intervals: before reform (school years 2007–2008, 2008–2009, and 2009–2010, where available), early reform (2010–2011, 2011–2012, and 2012–2013), and late reform (2013–2014 and 2014–2015). Figure 12.1 shows the average TE measured in VAM, and Table Q.1 in Appendix Q presents the means and standard deviations.[3]

Figure 12.1
District Average Teacher VAM Scores, over Time



RAND *RR2242-12.1*

[3] A word of caution is in order when interpreting these means and standard deviations. It is tempting to compare the average VAM estimates reported here with the estimated overall impact on test scores reported in Chapter Thirteen. Although both sets of estimates use student test scores and control for student characteristics, two important differences make this comparison difficult. First, the overall impact estimates measure the test score level, relative to average test scores during the prereform period, whereas VAM estimates measure the change in test score due to the teachers' contributions during the year. In a rough sense, the VAM estimates are the annual change in the overall impact estimates, with greater-than-average VAM being roughly equivalent to increasing impact. Second, the overall impact estimates compare test scores in IP schools with those in similar schools in the same state. The VAM estimates, on the other hand, are adjusted for score growth in all other schools in

In HCPS, average mathematics teachers' effectiveness significantly increased during the IP initiative by about 0.5 percentile point, while reading teachers' effectiveness significantly decreased by about 3 percentile points. Although HCPS had teacher value added that scored below the national public school average each year (below 0, given the normalization), average mathematics teachers' effectiveness was highest in the late reform period, at –0.035 standard deviations—equivalent to a decrease of between 0.8 and 2.0 percentile points from the 50th percentile (see Table 12.1). Compared with TE before the reform, this marks an improvement of 0.012 standard deviations—equivalent to a gain of about 0.5 percentile point. As shown in Table Q.1 in Appendix Q, TE became more varied across time: In the late reform years, there was a larger difference in student learning between the most-effective and least effective teachers. For reading teachers' effectiveness, there was no widening of the VAM distribution over time, but there was a steady decrease in average VAM (–0.069 standard deviations), equivalent to a loss of between 2 and 4 percentile points.

In PPS, the effectiveness of mathematics teachers improved slightly, while that of reading teachers declined, although only the change in reading was statistically significant. PPS saw slight, and not significant, improvements in mathematics teachers' effectiveness, with an increase of 0.003 standard deviations (0.1 percentile point). Reading teachers' effectiveness declined –0.02 standard deviations, which also is not large. The standard deviations were also roughly the same over time.

SCS reading teachers were much more effective, on average, in 2015–2016 than before the IP initiative. The very small increase in effectiveness for mathematics teachers was statistically significant but of negligible importance. SCS achieved a dramatic increase in average VAM across time for reading and a tiny increase for math-

---

the nation. IP schools have a larger LIM population than the nation as a whole. The VAM estimates will confound nationwide movement in LIM students' test scores relative to the average for all students with changes in VAM within an IP district.

ematics.[4] The mathematics change (0.001 standard deviations) was not of meaningful size, but the reading change was both statistically significant and large, at 0.234 standard deviations or the equivalent of about a 9-percentile-point gain. The very low average VAM for SCS reading in the pre- and early reform periods, followed by a high average VAM in the postreform period, is consistent with the impact estimates for SCS reported in Chapter Thirteen. Those results show that achievement in the pre- and early reform periods was decreasing compared with that in schools with similar demographics, then rebounded in the late reform period. The standard deviation of the reading VAM decreased slightly since the earlier periods, while average student achievement steadily improved.

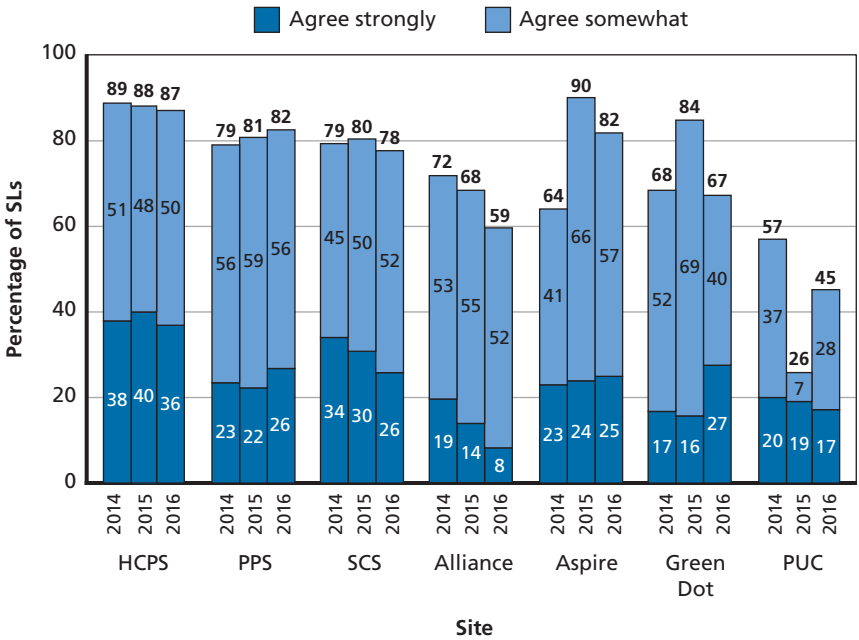
SL Perceptions About Student Access to Effective Teachers

In this section, we present SL survey findings that have bearing on students' access to effective teachers—particularly within respondents' schools.

Most SLs said that they had taken steps to ensure that students with the greatest needs were taught by the most effective teachers and that their school did a good job matching students with teachers. In nearly every site, most SLs indicated that they had taken steps to ensure that students with the greatest needs were taught by the most effective teachers (see Figure 12.2). Most SLs also said that their school did "a good job of matching students with teachers in ways that benefit the most students." (See Figure Q.1 in Appendix Q.) In the three districts, there was little change over time (from 2014 through 2016) in the percentages of SLs agreeing with these statements; in the

[4] We use only the legacy-MCS schools for our analysis of SCS teacher value added in this chapter, so as not to conflate changes in the population of students and schools with changes in the valued added of teachers in particular schools.

Figure 12.2
SLs' Agreement That They Had Taken Steps to Ensure That Students with the Greatest Needs Were Taught by the Most Effective Teachers, Springs 2014–2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly."
RAND *RR2242-12.2*

CMOs, there was no systematic change over time.[5] Principals were generally more likely than APs to agree with both statements.[6]

---

[5] As shown in Figure 12.1 here and Figure Q.1 in Appendix Q, PUC had a dip in 2015 in the percentage of SLs agreeing with both statements. On both, the decrease from 2014 to 2015 was statistically significant ($p < 0.01$), but the increase from 2015 to 2016 was not ($p < 0.05$). We are not sure what caused these swings in perception. It is worth noting that fewer than 20 SLs in PUC responded to the survey each year, so the volatility in the percentages could be due, in part, to the small number of respondents.

[6] For "I have taken steps to ensure that students with the greatest needs are taught by the most effective teachers," principals were significantly ($p < 0\ 0.05$) more likely than APs to agree in HCPS (2016, 2015, and 2014), SCS (2016, 2015, and 2014), PPS (2016, 2015,

Although a minority of SLs said that the highest-achieving students in their schools were typically taught by the "best" teachers, SLs in HSs were more likely to say this was the case. Less than half of SLs indicated that, in their school, "the highest-achieving students typically get the best teachers." However, SLs at HSs were generally more likely than those at elementary schools and MSs to agree that this occurred; the difference was starkest in HCPS, in which 63 percent of HS leaders agreed, compared with about only 30 percent of SLs at elementary schools and MSs (see Figure 12.3, which shows 2016; results for 2014 and 2015 were similar).[7]

In HCPS and SCS, most SLs said that teacher-evaluation results would be used to assign teachers to classes or students, but few teachers in the two sites shared this perception. In HCPS and SCS, but not in PPS or the CMOs, nearly all SLs indicated that teacher-evaluation results would be used at least to a small extent "to assign teachers to classes/students (within the school)," and about 80 percent said that teacher-evaluation results would be used for this purpose to a large or moderate extent (see Figure 12.4). However, few teachers shared this perception; in *every* site, in fact, very few teachers thought that their evaluation results would be used even to a moderate extent to "determine what classes or students you will teach next year" (see Figure 12.5). Although teachers in HCPS and SCS were slightly more likely than teachers in the other sites to say that their evaluation results would be used for this purpose, they were far less likely to say this based on what one might expect from the SL results. We are not sure why HCPS and SCS SLs and teachers differed in their perceptions. SLs might have been using evaluation results in assignment decisions but not making teachers aware of this. In addition, evalua-

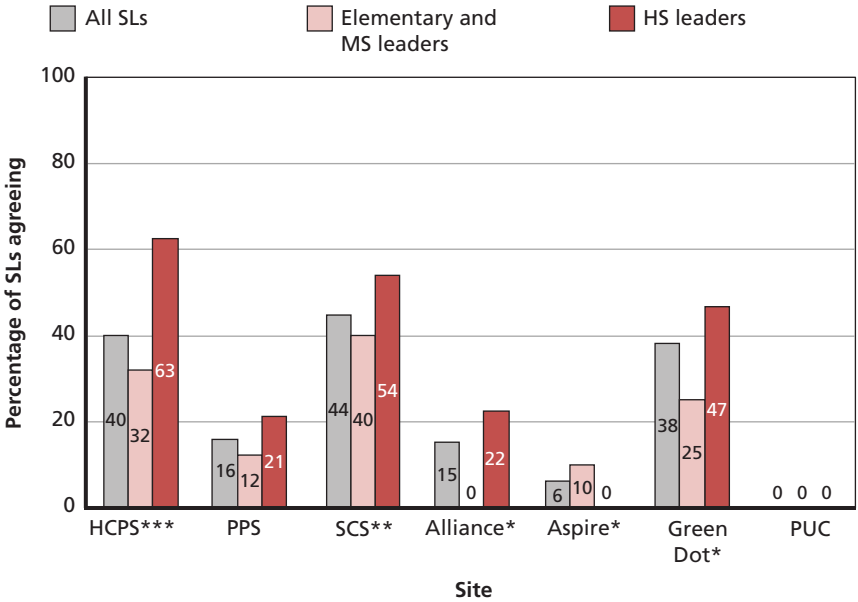
---

and 2014), Alliance (2016 and 2015), Aspire (2014), Green Dot (2015 and 2014), and PUC (2014).

For "My school does a good job of matching students with teachers in ways that will benefit the most students," principals were significantly ($p < 0.05$) more likely than APs to agree in HCPS (2016, 2015, and 2014), SCS (2016, 2015, and 2014), PPS (2016 and 2015), Alliance (2015 and 2014), Aspire (2015 and 2014), and Green Dot (2015 and 2014).

[7] In both 2014 and 2015, in HCPS, SCS, Alliance, and Green Dot, HS leaders were significantly ($p < 0.05$) more likely to agree than elementary school or MS leaders.

Figure 12.3
SLs' Agreement That, "In My School, the Highest-Achieving Students Typically Get the Best Teachers," by School Level, Spring 2016

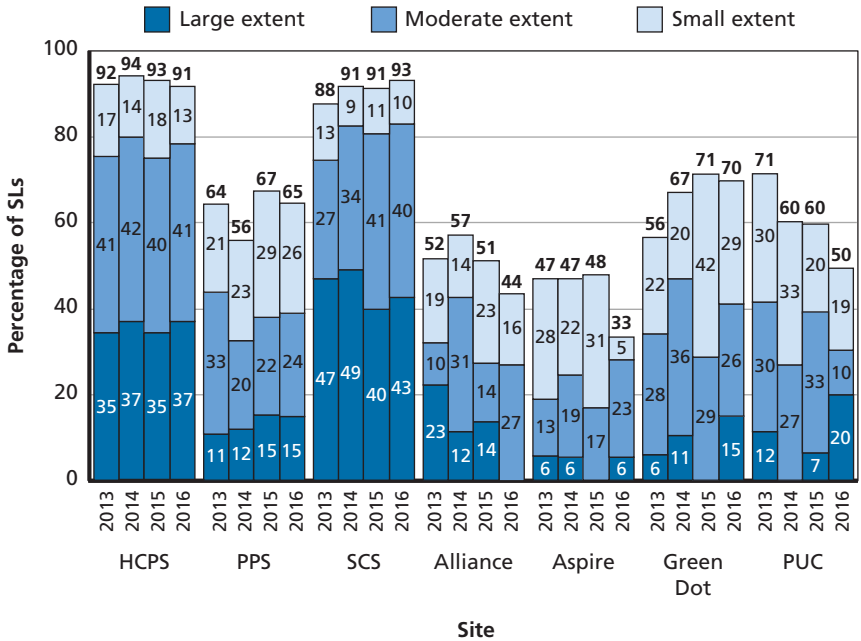


NOTE: * denotes that the difference between elementary school and MS and HS percentages was significant at $p < 0.05$. ** denotes that the difference was significant at $p < 0.01$. *** denotes that the difference was significant at $p < 0.001$. We combined the original response categories of "agree strongly" and "agree somewhat." Omitted response categories are "disagree somewhat" and "disagree strongly." In PUC, no SLs agreed with the statement, and, in Aspire, no HS leaders agreed with the statement.
RAND *RR2242-12.3*

tion results might have affected assignments for relatively few teachers; those teachers might have been aware of it but the rest of the teachers might not have been.

In some sites and years, more SLs indicated that effective teachers were given preference to teach honors courses than indicated that effective teachers were given preference to teach remedial courses. In MSs and HSs in which SLs indicated that the school offered honors and remedial classes, honors and other classes, or remedial and other classes, we asked the SL how much preference was given to effective teachers to teach honors classes and, separately, remedial

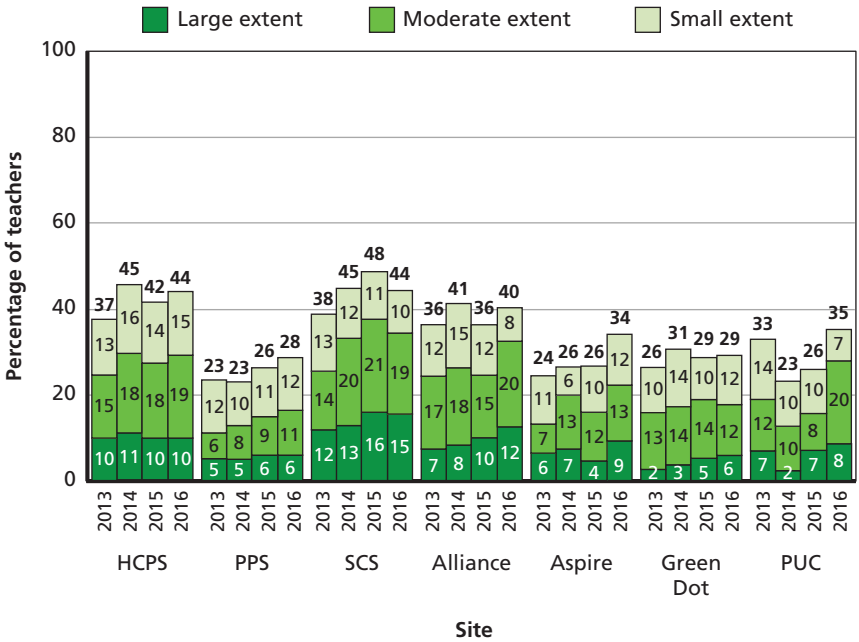
Figure 12.4
SLs Reporting That Teacher-Evaluation Results Would Be Used "to Assign Teachers to Classes/Students" (Within Their School), Springs 2013–2016



NOTE: Omitted response categories are "not at all" and "don't know."
RAND *RR2242-12.4*

classes. In 2016, SLs in all but two sites were more likely to indicate that effective teachers were given preference to teach the honors classes than they were to indicate that effective teachers were given preference to teach the remedial classes (see Figure 12.6). The same was true in 2014, but, in 2015, this was the case in only three sites. In none of the three years did more than 40 percent of SLs in any site agree that "teachers who are effective with high-achieving students would probably be less effective with low-achieving students" (see the last row of Figure Q.1 in Appendix Q).

Figure 12.5
Teachers Reporting That Their Evaluation Results Would Be Used "to Determine What Classes or Students Within Your School You Will Teach Next Year," Springs 2013–2016



NOTE: Omitted response categories are "not at all" and "don't know."
RAND *RR2242-12.5*

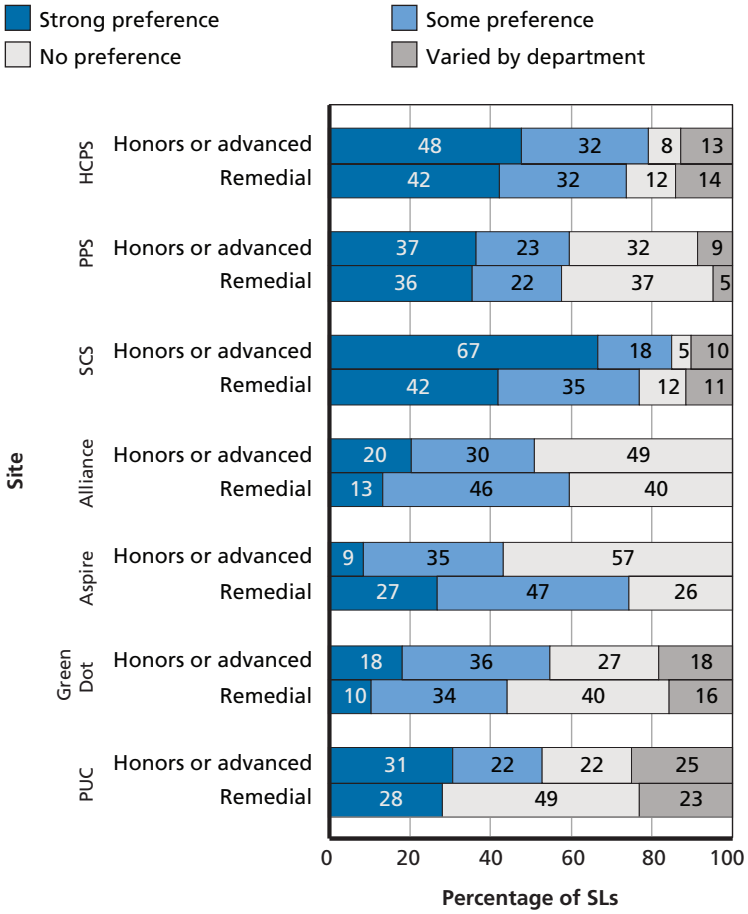
Estimates of LIM Students' Access to Effective Teachers

We also evaluate LIM students' access to effective teachers. In the three districts, in which sufficient data are available, we examine this association during the three school years before the IP initiative commenced (2007–2008 through 2009–2010), as well as during the five years of implementation (2010–2011 through 2014–2015).

Our Measure of LIM Students' Access

We empirically examine the relationship between TE and LIM students' assignments at the three district sites using the study-calculated

Figure 12.6
SLs' Responses About the Preference Given to Teachers with Higher Effectiveness Ratings in Assigning Them to Teach Honors and Remedial Classes, Spring 2016



NOTE: We asked these two questions only of SLs who reported that their schools included MS or HS grade levels. Only leaders who reported that their schools offered both honors and remedial courses saw both questions; some saw only one or the other (or neither), depending on what course offerings they reported.

RAND *RR2242-12.6*

VAM measures of effectiveness rather than the sites' composite TE measures. We do this for several reasons. The sites' measures were not

available for the years before the IP initiative, which would prevent our examining changes in access that occurred at the onset of the initiative. We were also very concerned that the sites' measures included classroom observations, which have been shown to be biased against teachers in classrooms with high proportions of LIM students, thereby threatening the very purpose of this analysis. We use these study-calculated VAM measures of effectiveness in three ways.

First, we evaluate the overall change in LIM students' access to effective teachers. We also decompose this overall change into two parts: whether LIM students have access to the most-effective teachers within their schools and whether LIM students are likely to attend schools with the most-effective teachers. We refer to this analysis as the *within- and between-school decomposition*. Our method, described in Appendix R, measures LIM students' overall access to effective teaching, the between-school access, and the within-school access. A coefficient estimate of 0.2, for example, implies that an increase of 25 points to a teacher's percentage of LIM students (e.g., from 50 percent of the teacher's students who are LIM to 75 percent) is associated with an increase of $0.2 \times 0.25 = 0.05$ standard deviations of student achievement. As shown in Table 12.1, this increase of 0.05 standard deviations is associated with a 2-percentile-point gain from such a teacher. In this example, a student at median or 50th-percentile achievement in a class made up of 50 percent LIM students could be expected to perform at the 52nd percentile if moved to a class with 75 percent LIM students because they obtained more-effective teachers.

Second, we estimate interrupted time-series regressions to see how access has changed over time, in particular from before to after the initiative. This analysis tests whether the initiative had any significant impact on the levels of access that LIM students have to effective teachers.

Third, we examine the mechanisms by which any changes in overall access have occurred. If LIM students' access is increasing, is it because their teachers are becoming more effective, because more-effective teachers are newly assigned to classes with many LIM students, or because the number of LIM students in the classes of effective teachers is increasing? We label these mechanisms *improve*, *replace*,

and *reassign*, respectively. Details of the methodology can be found in Appendix R and in Baird et al., 2016. Here, we provide a high-level summary of the findings.

LIM Students' Access to Effective Teaching

Overall, LIM students did not have consistently greater access to effective teachers in any of the three public school districts before or during the reform. Figures 12.7 through 12.9 show that, in general, LIM students have less access than non-LIM students within their schools to effective teachers, but they are more likely to be in the schools with the most-effective teachers. The greater access due to the fact that LIM students and effective teachers are located in a given school (the between effect) is larger than the lack of access within schools (the within effect), leading to slightly greater access overall. However, these patterns were present before the IP reforms, suggesting that the reforms had little impact on access.

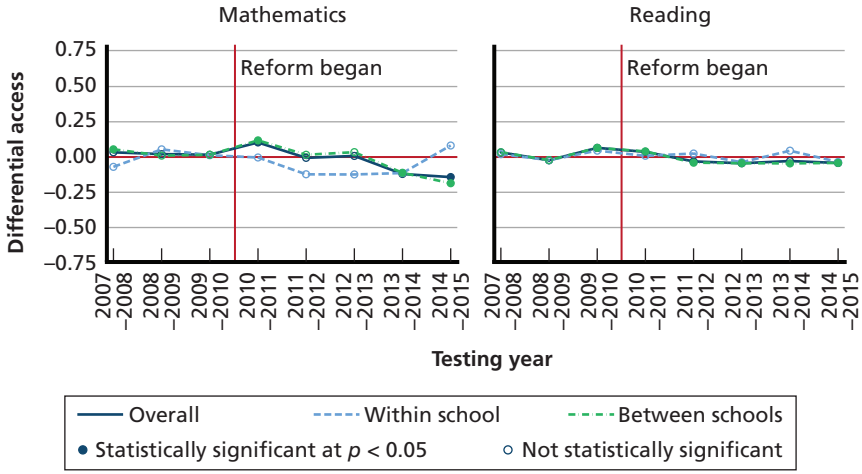
In HCPS, we observe differential access (i.e., sorting) that is generally small, especially for reading. By 2015, there is a slight overall disadvantage in access to effective teaching for LIM students (Figure 12.7). We see a one-year nonsignificant improvement in LIM students' within-school access to effective mathematics teachers from 2014 to 2015 after several years of favoring non-LIM students. However, this improvement in 2015 was offset by continued change in between-

How to Read the Access Figures

The figures in this section show how LIM students' access to effective teaching has changed over time. The following is a guide to how to read them:

- *The vertical red line* denotes when the reform began.
- *There are three access coefficients:* overall access, within-school access, and between-school access.
- *Positive values on the* y-*axis (differential access)* denote the allocation or sorting of effective teachers that favors LIM students. *Negative values* denote the allocation that favors non-LIM students. The larger the value, the greater the difference in effectiveness between teachers who instruct high proportions of LIM students and those who teach low proportions of LIM students. A value of 0.1, for example, means that an increase in LIM proportion of 50 percentage points (e.g., going from a teacher assigned 25 percent LIM students to one assigned 75 percent) is associated with the assignment of a teacher who is 0.1 × 0.5 = 0.05 standard deviations of student achievement—or 2 percentile points—better.

Figure 12.7
Difference in Access to Effective Teaching in HCPS Between LIM Students and Students Who Are Not LIM Students



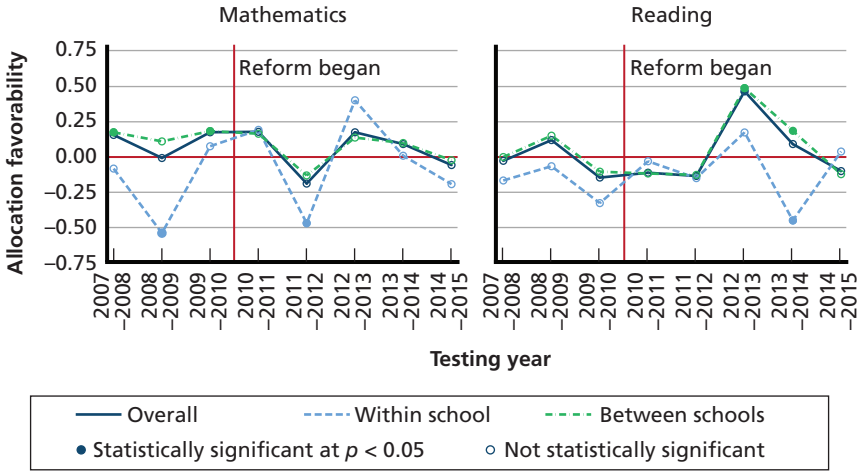
NOTE: A positive value indicates allocation favoring LIM students. A negative value indicates allocation favoring non-LIM students.

RAND *RR2242-12.7*

school access that favored non-LIM students, leading to little change in overall access to effective mathematics teachers from 2014 to 2015.

In PPS, within-school sorting tends to be worse than between-school or overall sorting for LIM students (Figure 12.8). However, only a few of the results are statistically significant. For 2015, we cannot say that any access parameter differs from 0 in reading or in mathematics. The distribution of mathematics teachers was more favorable to non-LIM students in 2015 than in 2014 for all three measures of access. The results regarding the change in the allocation of reading teachers in recent years are mixed. Although none of the differences in access between LIM and non-LIM students was statistically significant in 2015, reading saw improvements in within-school LIM-student access to effective teachers from 2014 to 2015, while between-school and overall access worsened. This could happen if several effective teachers moved from teaching classes with few LIM students in high-LIM schools to teaching classes with the most LIM students in low-LIM

Figure 12.8
Difference in Access to Effective Teaching in PPS Between LIM Students and Students Who Are Not LIM Students



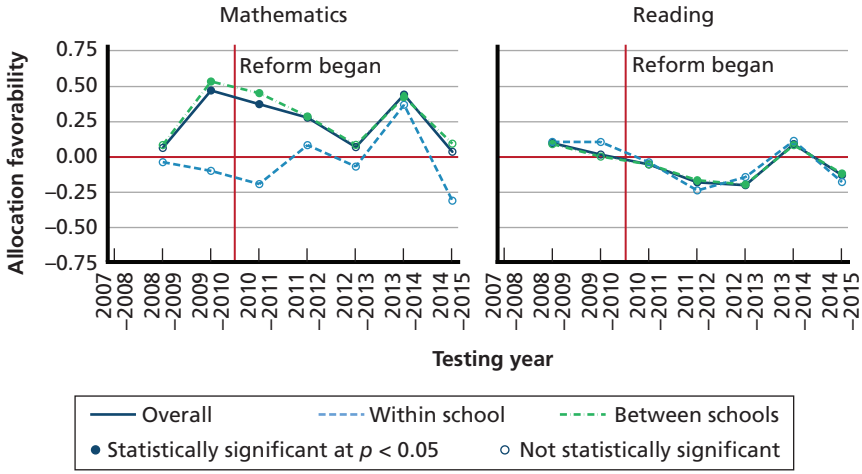
NOTE: A positive value indicates allocation favoring LIM students. A negative value indicates allocation favoring non-LIM students.

RAND *RR2242-12.8*

schools. The overall and between-school access would get worse, but the within-school access would improve.

In SCS, we observe larger coefficients and, in mathematics, a wider separation between within-school sorting (more negative) and between-school sorting (more positive; see Figure 12.9). The overall sorting in mathematics is consistently favorable to LIM students, whereas, for reading, it tends to be unfavorable for LIM students, with an exception in 2014. We also see notable year-to-year trends. All three access coefficients in both mathematics and reading dropped from 2014 to 2015. In 2015, we see that overall, between-school, and within-school access to effective reading teaching favors non-LIM students. For overall sorting in mathematics in 2015, differential access reduced LIM students' achievement by 0.05 standard deviations—or 2 percentile points—relative to non-LIM students. Meanwhile, within-school access to effective mathematics teaching favored LIM students in 2014 then reversed course in 2015. The allocation for mathematics over-

Figure 12.9
Difference in Access to Effective Teaching in SCS Between LIM Students and Students Who Are Not LIM Students



NOTE: A positive value indicates allocation favoring LIM students. A negative value indicates allocation favoring non-LIM students.

RAND *RR2242-12.9*

all and between schools was approximately neutral in 2015, having decreased slightly.

The Impact of the IP Reform on LIM Students' Access

With the exception of HCPS, in which overall access declined in both subjects, there was no statistically significant difference in LIM students' access to effective teachers from before the initiative to after the initiative. Table 12.2 presents the parameters that represent the change in the access coefficients from before (2009–2010 and earlier) to after (2013–2014 and later) the IP initiative. Positive values represent improvements, with asterisks denoting results that were statistically significant. In HCPS, there was a statistically significant decrease from prereform to postreform in LIM students' overall and between-school access to effective teachers for both mathematics and reading. In PPS and SCS, no differences were statistically significant between before the initiative and after, although the magnitude was

Table 12.2
Interrupted Time-Series Estimates of Change in LIM Students' Access to Effective Teaching from Before the Initiative to After the Initiative

District	Subject	Overall	Between	Within	Observations	Teachers	Schools
HCPS	Mathematics	–0.151***	–0.163***	–0.053	8,169	3,800	199
		(0.040)	(0.044)	(0.051)			
	Reading	–0.072***	–0.083***	–0.005	9,469	4,314	196
		(0.023)	(0.024)	(0.042)			
PPS	Mathematics	–0.046	–0.063	0.121	917	396	59
		(0.086)	(0.089)	(0.210)			
	Reading	0.023	0.031	–0.035	1,165	515	59
		(0.077)	(0.074)	(0.166)			
SCS	Mathematics	–0.032	–0.049	0.052	3,164	1,699	291
		(0.076)	(0.081)	(0.146)			
	Reading	–0.053	–0.044	–0.084	3,764	2,078	295
		(0.043)	(0.046)	(0.116)			

NOTE: The prereform period is 2009–2010 and earlier; postreform is 2013–2014 and later. *** denotes that the difference is statistically significant at $p < 0.01$. Regressions are weighted using the standard errors, clustered at school level, of the VAM measures. We adjusted p-values for multiple hypotheses using Benjamini–Hochberg correction.

large in some cases, and lower statistical significance is likely primarily driven by the fact that the samples were smaller. (Appendix R provides estimation details.)

Mechanisms for Changing LIM Students' Access to Effective Teaching

There is no evidence that the sites took steps to assign LIM students to more-effective teachers, nor is there evidence that teachers of LIM students increased their effectiveness more than teachers of non-LIM students did. Although there was no consistent pattern to the change in access over time, sites might have been making changes to improve access that were offset by other changes. Therefore, we analyze individual components of the change in access that were more or less under the sites' control. As discussed in Appendix R and Baird et al., 2016, we decompose changes in overall access into three portions that reflect the mechanisms for changing access: (1) differential changes in TE, holding assignments to LIM constant (Δ improve); (2) reassignment of teachers to different LIM students' proportions, holding TE constant (Δ reassign); and (3) replacement of teachers (Δ replace). For comparison, we present these three decompositions in the figures, as well as the overall change in access; we weight the changes by either p or $1 - p$, where p is the proportion of teachers who are retained, this stemming from the decomposition methodology described in Appendix R.

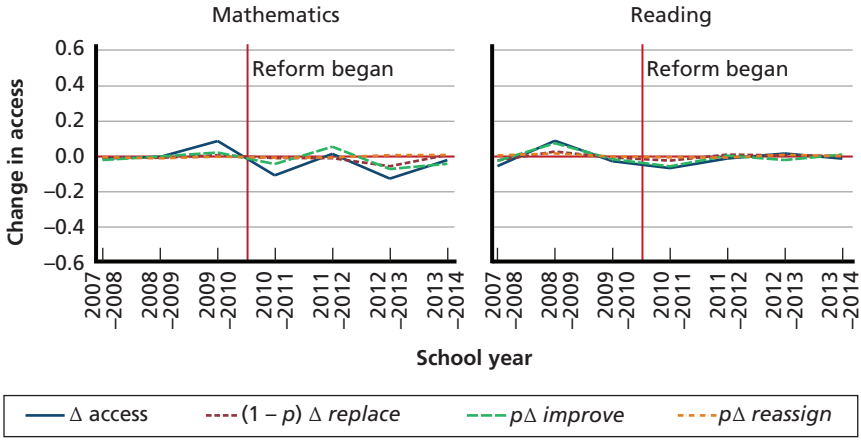
Before examining the results, we note three things. First, the first two components pertain to teachers who continue to teach, who usually constitute a much larger fraction of the workforce than those who are replaced each year. Second, the first and third components both reflect the site-wide change in the VAM score between two years, so they will tend to move together. Third, the process for estimating confidence intervals for these estimates is very complicated, and we do not include it here. These estimates of the importance of the three mechanisms should be viewed as guidance regarding their relative levels and magnitudes, presented for exploratory purposes. In that spirit, we find little evidence that our earlier finding of little overall change in access is the result of offsetting changes in any of the decomposition trends associated with the mechanisms *improve*, *reassign*, or *replace*.

First, we look at the decomposition trends for HCPS (Figure 12.10). For both mathematics and reading, there is little to no effect due to reassignment. For mathematics, in which there was improvement in access after the initiative, it tended to be driven by the *improve* mechanism. Thus, HCPS seems to be characterized by little change in how teachers were assigned to LIM students but with some improvements driven by increased effectiveness among teachers assigned to high-LIM classrooms (or decreases in the quality of teachers assigned to non-LIM classrooms).

Next, we look at PPS (Figure 12.11). We find some evidence for movement due to reassignment, with 0 or positive movement coming from changes in the effectiveness of teachers assigned to LIM students in each year after the initiative. The other mechanisms continue to have stronger effects, though, and act as the drivers of changes in overall sorting of effective teachers to LIM students.

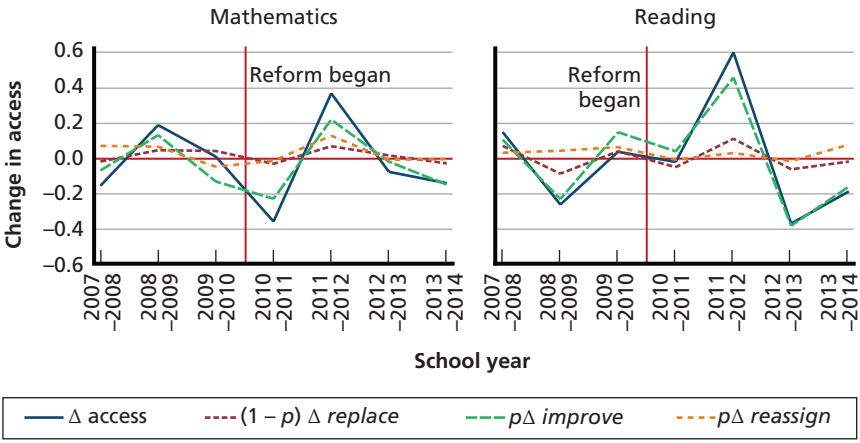
Finally, we look at SCS (Figure 12.12) and again find virtually no effect stemming from changes in assignments, holding effectiveness constant. The improvements in mathematics and reading are both

Figure 12.10
HCPS Decomposition of Change in Access Coefficients



RAND *RR2242-12.10*

Figure 12.11
PPS Decomposition of Change in Access Coefficients



RAND *RR2242-12.11*

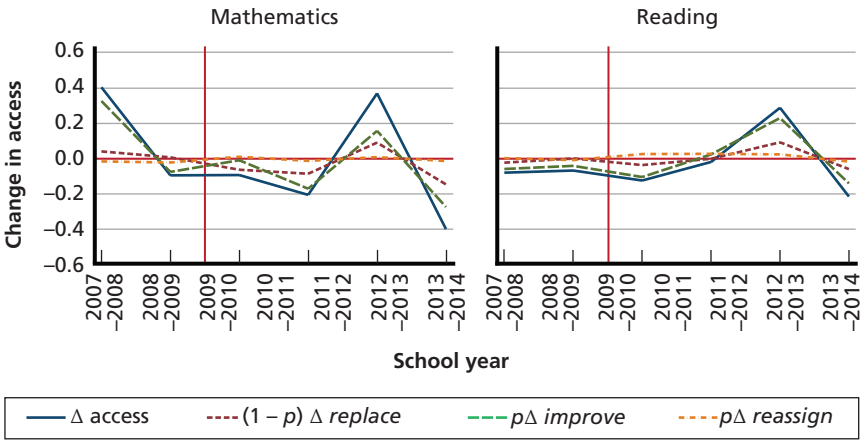
driven primarily by changes in effectiveness but also by replacement of teachers. However, the direction of change due to each component switches from year to year.

Summary

In this chapter, we evaluate the changes in TE and LIM students' access to effective teachers across time. To measure effectiveness, we used our study-calculated VAM measure rather than the sites' composite TE measures because the sites' measures were not available for the years before the IP initiative and because we wanted to guard against any drift that might have crept into their measures. We anchored our VAM measure to national criteria, as captured by NAEP and as described in Appendix A, in order to have a fixed yardstick by which to measure effectiveness.

We found mixed results in changes in overall TE. In HCPS, we found statistically significant but small increases in the effectiveness of mathematics teachers but decreases in that of reading teachers. The changes in effectiveness for PPS reading teachers were small, nega-

Figure 12.12
SCS Decomposition of Change in Access Coefficients



RAND *RR2242-12.12*

tive, and statistically significant, and, for math teachers, the changes in effectiveness were very small but not significant. The trend for SCS was more positive: Mathematics teachers were more effective but only a small amount, and reading teachers were much more effective.

We examined overall LIM students' access to effective teaching, as well as access within their current schools and access to schools with the most-effective teaching. We examined access each year during the initiative, and we assess the initiative's overall impact by comparing access before the initiative with access during the initiative using an interrupted time-series approach. We found no evidence of consistent improvement during the IP initiative in the measures of LIM students' access to effective teaching. In fact, in HCPS, there was evidence of decreased overall LIM students' access to effective teaching in both reading and mathematics, driven primarily by a decrease in access to schools with effective teaching.

When we decompose the mechanisms through which access might change, we find that changes in teacher assignments played little or no role in observed changes in access. This suggests that, even with the aim of improving access, these districts did little in terms of reas-

signing teachers to improve LIM students' access to effective teachers. This is consistent with our finding in Chapter Seven that effective teachers who taught LIM students were not receiving greater salaries. This situation might have occurred because the districts found it difficult to change teacher assignments, especially to reassign teachers across schools. Even within schools, some SLs might have been reluctant to reassign HE teachers to classes with more LIM students, especially given our finding that some SLs believed that teachers who were effective with high-achieving students would probably be less effective with low-achieving students.[8]

---

[8] Recent research (e.g., Loeb, Soland, and Fox, 2014) has pointed out that an effective teacher is not necessarily still effective when moved to a new context. It might be tempting to suggest that our finding of little change in LIM students' access to effective teachers is an example of such changing effectiveness associated with a change in classroom composition. However, our analysis of mechanisms explicitly rejects this hypothesis. The $p\Delta$ *reassign* measure is near 0 throughout the years and for all districts; this measure captures how much the access coefficients change as a result of changing assignments, after holding fixed the previous year's VAM measure of TE. For the "new context leads to reduced value added" hypothesis to be true, we would have to see an increase in the estimated *reassign* mechanism as a result of the most-effective teachers being assigned higher-LIM classes on average (and consequently, the lower-effectiveness teachers being assigned, on average, lower-LIM classes). However, we did not find this result.

CHAPTER THIRTEEN

The Initiative's Impact on Student Outcomes

In Brief

As a whole, did the IP initiative improve student outcomes?

Both survey responses from teachers and comparative analyses of student outcomes indicate that the initiative did not meet its goal of dramatic improvement in outcomes for LIM students. In all the sites, most SLs thought that the initiative would benefit students in the long run, but they became less likely to think this over time. Teachers in the three districts also became less likely over the course of the initiative to think that the initiative would benefit students in the long run. Teachers in the CMOs were more likely than district teachers to think that the reform would benefit students.

Our analyses of student test results and graduation rates showed that, six years after the IP initiative began, there is no evidence of widespread positive impact of the initiative on student outcomes. In 2014–2015, like in previous years, the estimated impacts in the IP sites were mostly not statistically significant across grades and subjects, although there were significant positive effects for HS ELA in PPS and the CMOs and significant negative effects in mathematics in grades 3 through 8 in the CMOs.

Introduction

As we described in Chapter One, the IP initiative's goals were to produce dramatic improvement in student outcomes (including achievement, graduation, and college-going), particularly among LIM students. This chapter presents two kinds of information about the initiative's impact

on student outcomes: (1) teachers' and SLs' perceptions of effects on student and teacher performance and (2) statistical analyses of measured student outcomes (both achievement test results and nontest outcomes, such as dropout rates).[1] Although we place greater stock in the statistical analyses as measures of overall impact, it is informative to see how the educators who are the initiative's targets are predicting and gauging effects. We begin with educators' perspectives on the initiative's effects and then turn to our impact findings.

Teachers' and SLs' Perceptions of Impact

How teachers and SLs perceive the IP initiative's impact is important for at least two reasons. First, if educators believe that the initiative is having a positive impact, their efforts might, in fact, help bring about a positive impact (i.e., the power of positive thinking); conversely, if they do not believe that the reforms are having a positive impact, they might subconsciously (or even consciously) work to undermine the initiative's goals. Thus, a belief can become a self-fulfilling prophecy. Second, teachers and SLs might define *impact* in broader terms than the metrics that program evaluators use (e.g., measurable indicators, such as test scores and graduation rates). The standard metrics are crucial to examine, but few would argue that they capture the full spectrum of potential positive impact. The following are key findings about educators' perceptions of the initiative's impact.

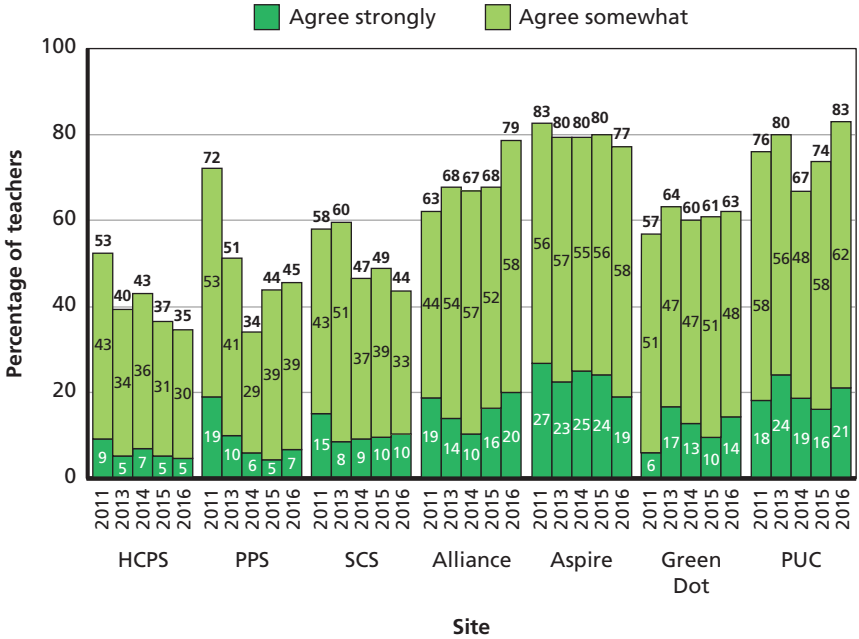
Teachers in the districts became less likely over the course of the initiative to think that the initiative would benefit students in the long run; teachers in the CMOs were more likely than district teachers to think that the reform would benefit students, with no consistent time trend. As part of our annual surveys, we asked teachers and SLs whether they agreed or disagreed (on a four-point scale) that, in the long run, students would benefit from their site's teacher-evaluation system. (Although the teacher-evaluation system was only

[1] Because there were no publicly available data on college-going rates by HS, at least for the sites included in this study, we did not assess the initiative's impact on college-going.

one part of each site's reform efforts, it was arguably the most central part, which, in turn, influenced nearly all other parts.) Figure 13.1 shows the teacher responses, and Figure 13.2 shows the SL responses.

In the three districts, typically less than 50 percent of teachers agreed that the evaluation system would ultimately benefit students, and the percentages declined over the years of the reform.[2] Although the overall trend in PPS was negative, perceptions began to rebound after 2014.[3]

Figure 13.1
Teachers' Agreement That, "in the Long Run, Students Will Benefit from the Teacher-Evaluation System," Springs 2011–2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly." We did not survey teachers in 2012.

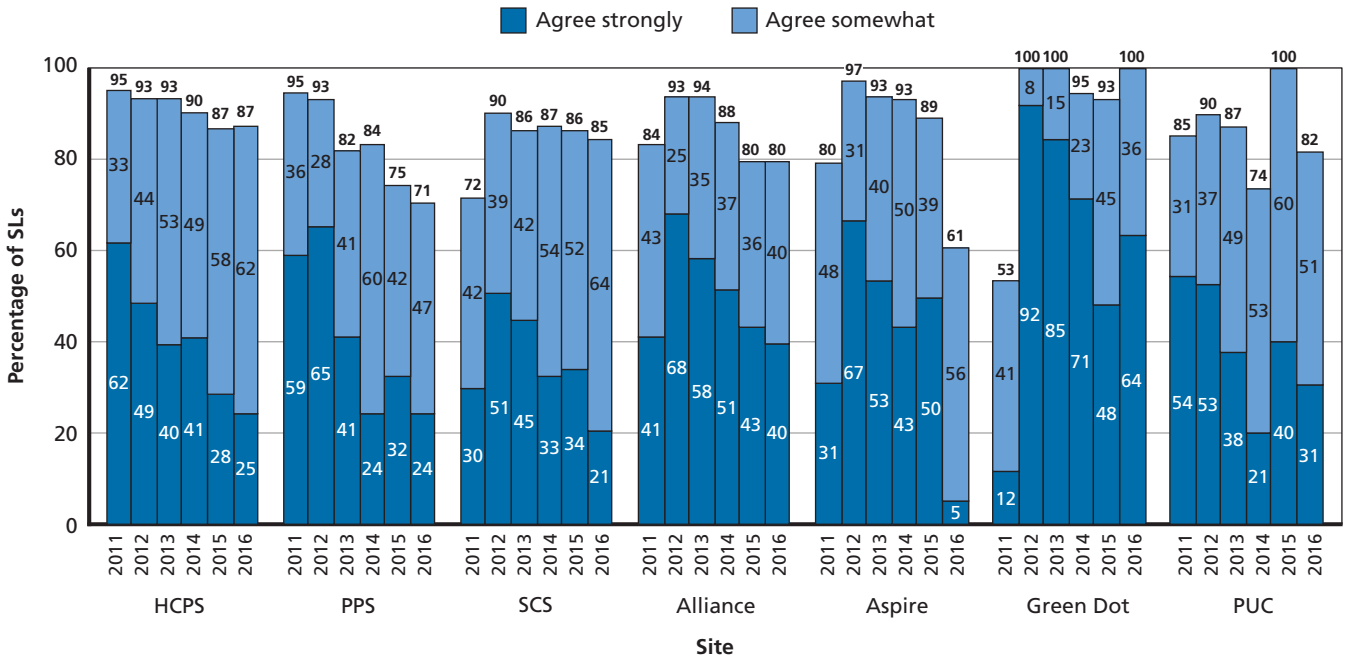
RAND *RR2242-13.1*

---

[2] In all three districts, the decline in the percentage agreeing (strongly and somewhat) from 2011 to 2016 was statistically significant ($p < 0.05$).

[3] The increase from 2014 to both 2015 and 2016 was statistically significant ($p < 0.05$).

Figure 13.2
SLs' Agreement That, "in the Long Run, Students Will Benefit from the Teacher-Evaluation System," Springs 2011–2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly."

RAND *RR2242-13.2*

In the CMOs, agreement was notably higher. In Aspire and Green Dot, the percentage agreeing remained essentially the same over time, while Alliance teachers became more likely to believe that the system would benefit students.[4] In PUC, the percentage fell initially and then rebounded.[5]

In general, novice teachers were more likely than more-experienced teachers to agree that the evaluation system would benefit students in the long run.[6] In addition, teachers who had received a high effectiveness rating were generally more likely to agree than teachers with a mid-level rating.[7] In 2014, high-rated teachers were also more likely than low-rated teachers to agree,[8] but this was not the case in the other years.

SLs were more likely than teachers to think that the reform would benefit students in the long run, but, like teachers in the districts, they became less likely to agree over time. In SCS and the CMOs, SL agreement increased significantly from the spring of 2011 (before the TE measure was available) to the spring of 2012 (in the first year it was calculated). But, after 2012, agreement began to decline, as seen in the overall percentages for five sites and in the "agree strongly" percentages for all seven. In other words, after their initial enthusiasm when the evaluation systems were first implemented, SLs became less likely to believe that the system would have long-term positive effects on students.

---

4 In both Aspire and Green Dot, there were no pairs of years in which the percentage agreeing differed significantly ($p < 0.05$) between the two years. In Alliance, the increase from every prior year to 2016 was statistically significant ($p < 0.05$).

5 The 2014 percentage was significantly lower ($p < 0.05$) than the percentage in 2011, 2013, and 2016. The 2016 percentage was significantly higher ($p < 0.05$) than the 2015 percentage.

6 In HCPS (2013, 2014, 2015, and 2016), SCS (2015 and 2016), Alliance (2013, 2014, and 2016), Aspire (2016), and PUC (2014), novice teachers were significantly ($p < 0.05$) more likely than experienced teachers to agree. We did not do the comparison for 2011.

7 High-rated teachers were significantly ($p < 0.05$) more likely to agree than mid-rated teachers in HCPS (2015 and 2016), SCS (2014), PPS (2014 and 2016), Alliance (2014), and Green Dot (2013 and 2016).

8 Significant in PPS, Alliance, and Aspire.

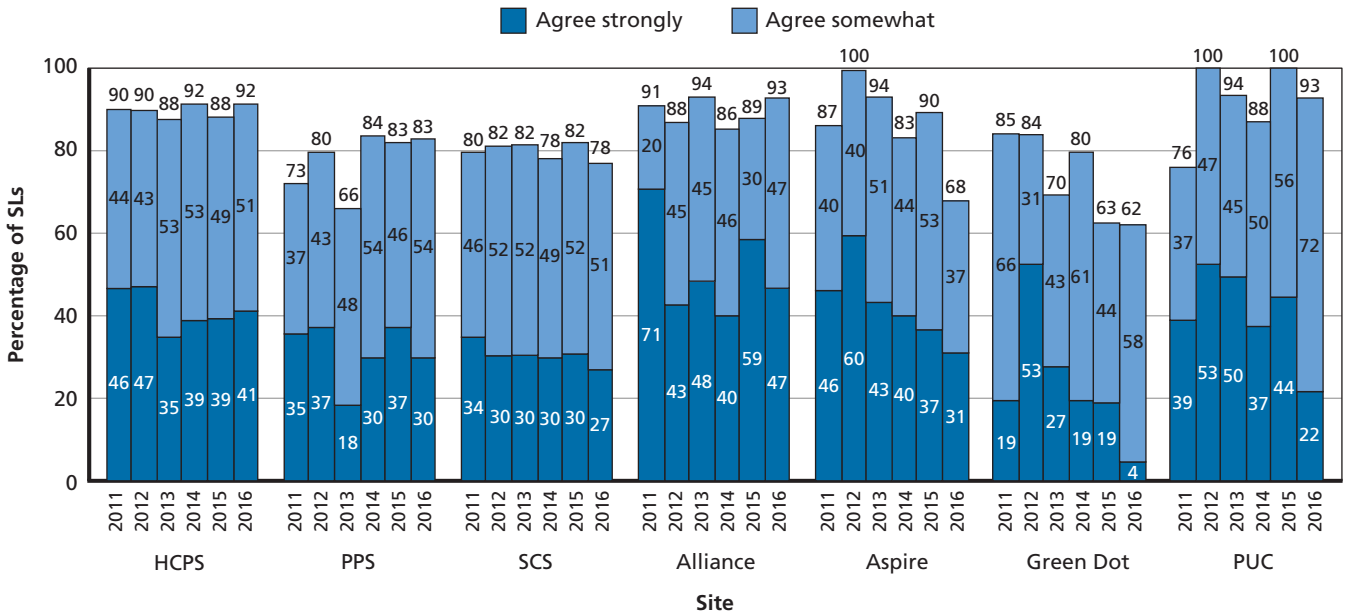
Most SLs indicated that they were satisfied with the quality of their current teaching staff; across sites, no consistent time trends emerged. Each year, we asked SLs whether they were satisfied with the quality of their current teaching staff. Figure 13.3 presents the results. In both HCPS and Alliance, about 90 percent of SLs agreed that they were satisfied, with little change over time. In SCS, about 80 percent agreed, a level that was also very stable across the years. The other sites had greater variability from year to year, but, comparing 2011 with 2016, we see that agreement grew in PPS and PUC while it declined in Aspire and Green Dot.

In interpreting these findings, several considerations should be kept in mind. One might hope that the quality of schools' teaching staffs would have improved over the course of the initiative and that SL perception of the quality would correspondingly improve, as was the case in PPS and in PUC.[9] However, it could also be that SLs' *standards* of quality increased during the initiative, potentially causing a decline in their *perceptions* of quality. Moreover, context might have played a role, particularly in the CMOs, in which teacher turnover was high, new schools were opening, and it was difficult to recruit high-quality teachers because of a tight labor market (see Chapter Four).

SLs tended to think that most, but not all, of the teachers in their school had the skills needed to foster meaningful student learning. Starting in 2013, we asked SLs roughly what proportion of teachers in their school had "the skills needed to foster meaningful student learning," as well as various other skills related to improving student outcomes. Nearly all SLs said that half or more of their teachers had the skills needed to foster meaningful student learning. However, very low percentages of leaders—often 20 percent or fewer—said that "all or nearly all" of the teachers had these skills (see Figure 13.4). Like with the previous question, percentages were essentially constant over

[9] As discussed in Chapter Ten, the likelihood that low-TE teachers would remain teaching dropped significantly in PPS during these years, which could help explain the increase in the percentage of PPS SLs reporting satisfaction with their teaching staff. On the other hand, several other sites (HCPS, Alliance, Aspire, and Green Dot) also all had decreases in the likelihood of low-TE teachers remaining in teaching, yet SLs did not become more likely to express satisfaction with the quality of their teaching staff.

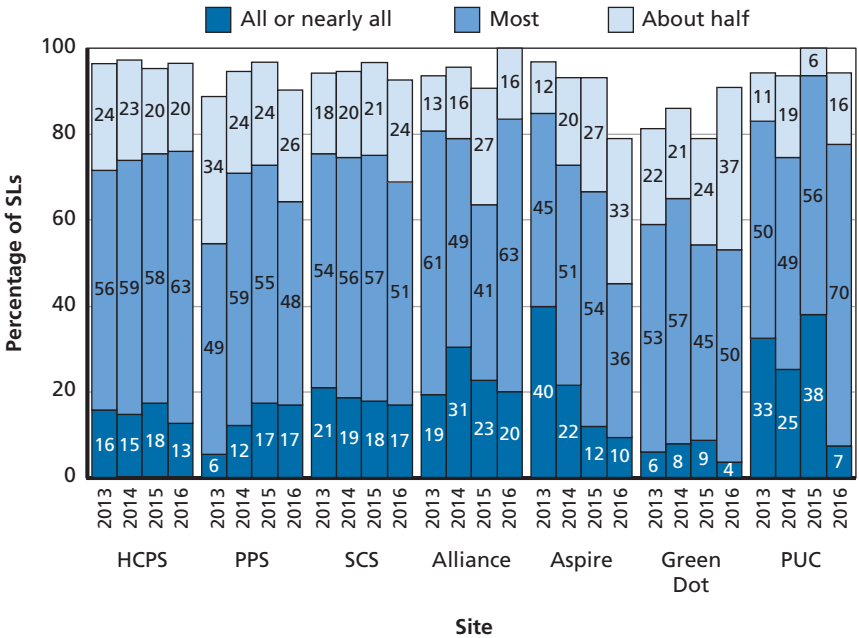
Figure 13.3
SLs' Agreement That, "Overall, I'm Satisfied with the Quality of the *Whole Teaching Staff* in My School This Year," Springs 2011–2016



NOTE: Omitted response categories are "disagree somewhat" and "disagree strongly." In 2011 and 2012, the question as it appeared in the survey also had a "not applicable" option, but we coded the very few responses of "not applicable" as missing in the calculation of the percentages shown.

RAND *RR2242-13.3*

Figure 13.4
SLs' Responses About the Proportion of Teachers in Their School Who Had "the Skills Needed to Foster Meaningful Student Learning," Springs 2013–2016



NOTE: Omitted response categories are "a few" and "none or almost none." We did not ask this question in 2011 or 2012.

RAND *RR2242-13.4*

time in HCPS and SCS but declined in Aspire. (For results for SLs' perceptions of the other teacher skills, see Figure S.1 in Appendix S.)

Impact on Measured Student Outcomes

In this section, we present our analysis of the IP initiative's effects on reading and math scores on state-administered achievement tests; on nontest outcomes that are informative about TE, including graduation, dropout, and attendance rates (when available); and on the achievement of student subgroups (when available). Using statistical methods,

we estimate how students are faring in the IP sites compared with how they would likely have performed without the IP initiative.

Methodology

Our estimates of impact are based on a DiD methodology that lets us compare outcomes in schools participating in the IP initiative with outcomes in similar schools that were not part of the initiative. We compare outcomes before and after the start of the initiative while adjusting for differences in the demographics (e.g., ethnicity, recipients of FRPL) of the comparison groups at both the school level and the site level over time.

We also take into account each school's proficiency levels and each site's overall average achievement before the start of the IP initiative.

In these analyses, we define the preinitiative period as the school years through 2008–2009, before the initial grants for the initiative were received; and we define the postinitiative period as the school years starting in 2009–2010. The preinitiative period is used to forecast what the students' outcomes would have been had the initiative not taken place. Accordingly, we consider the initiative to have started when the sites' grants were awarded.[10]

How to Read Difference-in-Differences Figures

This chapter includes several figures depicting DiD effects on test scores. The following is a guide to how to read them:

- *The horizontal axis* indicates the spring of the school year when students were tested (e.g., *2010* represents the spring of the 2009–2010 school year).
- *The solid red vertical line* at 2009 marks the year before the initial grants from the initiative were received (i.e., school year 2008–2009).
- *The solid black lines* depict the initiative's estimated average effect across grades. Values greater than 0 on the vertical scale indicate that initiative schools had greater achievement gains than comparison schools did. (The magnitudes along the vertical scale are in standard deviations of the student test score distribution.)
- *The dashed blue lines* depict the 95-percent confidence intervals for the estimated effect. If 0 is not included in the confidence interval, we consider the effect statistically significant—meaning that we can attribute the effect to the initiative rather than to natural variation in scores.

[10] To avoid confusion, it is worth noting that many other sections of this report regard the start of the initiative as when the reforms were implemented, starting in 2010–2011.

Essentially, the DiD methodology tests whether student achievement in IP schools improved in comparison with its predicted value by more than in similar schools in the same state. We refer to the difference in improvement between the two groups as *gains* associated with participating in the initiative (although sometimes the difference is negative). A more detailed description of our methodology is in Appendix T and in Gutierrez, Weinberger, and Engberg, 2016.

Findings for the Three Districts

This section presents detailed results for each of the three IP districts; later we present results for the CMOs. Our discussion focuses on one key academic outcome: standardized student scores on state assessments in reading and math. For simplicity's sake, we first present the initiative's effect on scores averaged across grades within two groups—lower grades (grades 3 through 8, when available) and HS (grades 9 through 12, when available). Then, we comment briefly on the variability of the estimates by grade, subject, and year. We also comment on the initiative's effects on achievement by student subgroups (black, Hispanic, or low income) and on nontest outcomes (dropout, graduation, and attendance rates) when available. Appendix U contains more detail, including separate results for each grade and student subgroup, and for available nontest outcomes. In most cases, we focus our analysis on student outcomes for 2014–2015, the latest year with available data, based on the hypothesis that the longer the reforms have been in place, the more likely we are to see effects on student outcomes. However, we also sometimes comment on results from previous years, for example, to highlight trends or reversals in prior estimated impacts.

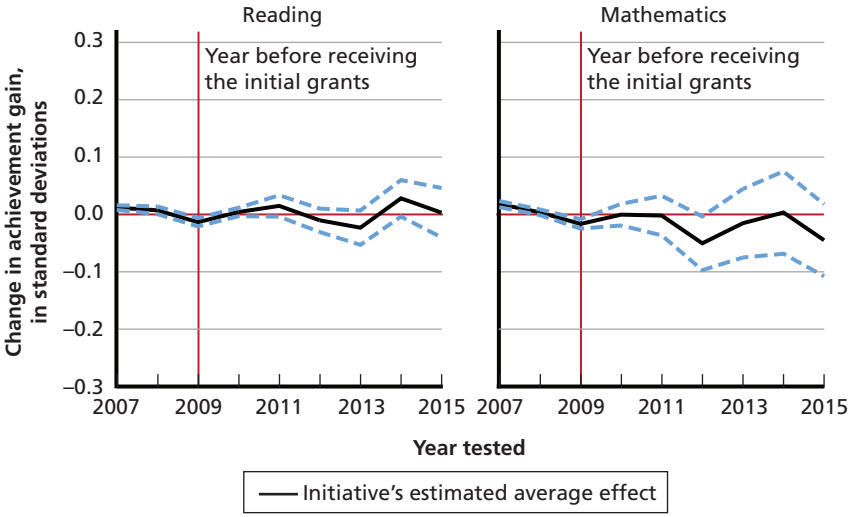
HCPS

The IP initiative had no effect on lower grades' academic achievement in HCPS over the course of the initiative. The initiative was associated with negative effects on HS achievement initially, but, for 2014–2015, we found no effects on academic achievement or dropout rates. In HCPS, the IP initiative had no statistically significant impacts on math or reading achievement in lower grades in most years since the initiative began. The estimated effect on lower

grades' average reading achievement in 2014–2015 was small and not statistically significant, as shown in Figure 13.5. The effect on lower grades' average math achievement was negative in 2014–2015 but not statistically significant. For subgroups in lower grades (black, Hispanic, and high-poverty students), we found no statistically significant effects on average achievement in 2014–2015, as shown in Table U.1 in Appendix U.

Over the course of the entire initiative, the estimated impacts in lower grades varied depending on the subject, year, and grade, as shown in Table 13.1. (Red indicates statistically significant negative effects, gray indicates non–statistically significant effects, and green indicates positive significant effects.) For example, we observe more negative effects in MS achievement (i.e., grades 6 through 8), particu-

Figure 13.5
Estimates of the Initiative's Effect on Average Achievement in HCPS Grades 3 Through 8



NOTE: The years are the spring terms, so *2010* represents the spring of the 2009–2010 school year. A positive value indicates that the initiative had greater achievement gains than comparison schools. A negative value indicates that the initiative had lower achievement gains than comparison schools. The dashed blue lines indicate the 95-percent confidence interval.

RAND *RR2242-13.5*

Table 13.1
Visualization of the Initiative's Impacts on Achievement in HCPS, by Year, Grade, and Subject

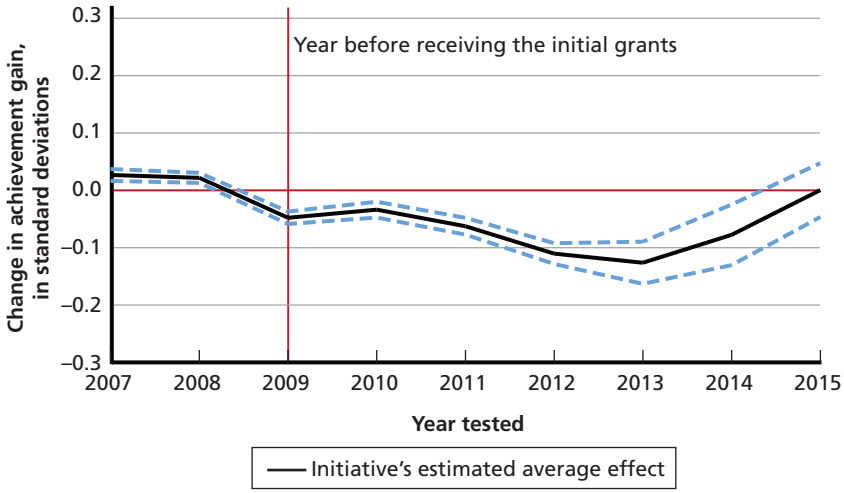
	Reading						Mathematics					
Grade	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
3	0.02	0.057	0.029	–0.01	0.076	0.007	–0.01	0.016	–0.04	–0.03	0.024	–0.05
4	–0.02	0.058	0.035	0.017	0.033	–0.01	0.027	–0.01	–0.07	–0.06	–0.06	–0.07
5	0.01	0.008	–0.01	0.002	0.06	0.024	–0	–0.01	–0.05	–0.01	–0.04	–0.05
6	–0.02	–0.04	–0.11	–0.09	–0.12	–0.09	–0.03	0.02	–0.10	–0.07	–0.07	–0.08
7	–0.02	–0.06	–0.09	–0.12	–0.08	0.068	–0.04	–0.05	–0.04	–0.03	–0	0.107
8	0.018	–0.07	–0.10	–0.08	–0.09	–0.07	0.007	–0.01	–0.03	0.115	0.163	–0.27
3–8	**0.003**	**0.013**	**–0.01**	**–0.03**	**0.026**	**0.001**	**–0**	**–0**	**–0.05**	**–0.02**	**0.002**	**–0.05**
9	–0.03	–0	–0.11	–0.13	–0.13	0.027	N/A	N/A	N/A	N/A	N/A	N/A
10	–0.08	–0.17	–0.12	–0.16	–0.10	–0.07	N/A	N/A	N/A	N/A	N/A	N/A
HS	**–0.03**	**–0.06**	**–0.11**	**–0.013**	**–0.08**	**0.000**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**

NOTE: Red indicates statistically significant negative effects, gray indicates non–statistically significant effects, and green indicates positive significant effects. (Criterion for statistical significance is $p < 0.05$.) The "3–8" row is from an analysis that combines these grades, weighted by enrollment in each grade. Similarly, the "HS" row is from a weighted analysis (based on enrollment) that combines grades 9 and 10.

larly in reading, during the early years of the initiative. In contrast, the effects on reading achievement in elementary school were mostly nonsignificant or even positive in some years (see Table U.1 in Appendix U for more details).

Figure 13.6 shows the initiative's estimated impacts on achievement in HS reading (grades 9 and 10).[11] As shown in Figure 13.6, after a drop in HS reading achievement compared with similar Florida dis-

Figure 13.6
Estimates of the Initiative's Effects on Average Reading Achievement in HCPS HSs



NOTE: The years are the spring terms, so *2010* represents the spring of the 2009–2010 school year. A positive value indicates that the initiative had greater achievement gains than comparison schools. A negative value indicates that the initiative had lower achievement gains than comparison schools. The dashed blue lines indicate the 95-percent confidence interval.

RAND *RR2242-13.6*

[11] In Florida, before 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 end-of-course (EOC) assessments. The FCAT 2.0 continued to test students in reading through grade 10 but does not administer a mathematics exam in grade 9 or 10. As a result, we have excluded HS mathematics from our analysis.

tricts, we observe improvements in the final three years of the initiative. We find a similar improvement in the estimates for black, Hispanic, and high-poverty students, as shown in Table U.1 in Appendix U.

For dropout rates, the initiative's effect in 2014–2015 was negative (i.e., an increase in dropouts) but small and not statistically significant. However, it was an improvement over 2012–2013, when dropout rates worsened by a statistically significant amount (for more details, see Table U.1 in Appendix U). Dropout rates also provide information that helps us judge whether TE improved. If achievement improved but the dropout rate worsened, the improvement in achievement might have been due to the departure of struggling students rather than increased TE. Improvements in both HS reading achievement and dropout rates from 2012–2013 to 2014–2015 suggest that the initiative improved TE from that in earlier years.[12]

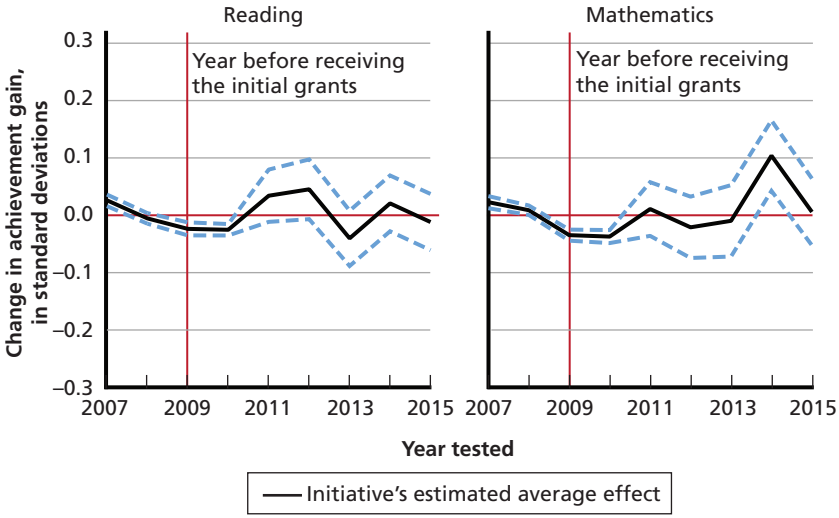
PPS

In school year 2014–2015, the initiative had no effects on average student achievement in reading or mathematics in the lower grades in PPS. However, it had a positive effect on HS (grade 11) reading achievement in three of the four most-recent years. In the 2014–2015 school year, PPS schools experienced about the same achievement gains in reading and mathematics in grades 3 through 8 as comparable schools in other Pennsylvania districts did, as shown in Figure 13.7. The estimated average impact is negative for reading in lower grades and positive for math, but both estimates are small and not statistically significant. Table U.2 in Appendix U also shows no statistically significant impacts on achievement by student subgroups (black and high-poverty students) in 2014–2015.

We also found mostly nonsignificant effects of the initiative on the average achievement across lower grades in prior years. The only

[12] The estimates of the initiative's effects on dropout rates for 2012–2013 and 2013–2014 differ from those previously reported in Gutierrez, Weinberger, and Engberg, 2016, because we updated the data on school demographics using information from the Common Core of Data (NCES, undated [b]) rather than the Florida Department of Education. The estimates of the initiative's effects on test scores in Gutierrez, Weinberger, and Engberg, 2016, already reflect this change.

Figure 13.7
Estimates of the Initiative's Effect on Average Achievement in PPS Grades 3 Through 8



NOTE: The years are the spring terms, so *2010* represents the spring of the 2009–2010 school year. A positive value indicates that the initiative had greater achievement gains than comparison schools. A negative value indicates that the initiative had lower achievement gains than comparison schools. The dashed blue lines indicate the 95-percent confidence interval.

RAND *RR2242-13.7*

exception was math in 2013–2014, for which we found significant positive effects. However, these positive effects were not sustained in the next year. As Table 13.2 shows, there was some variation in the estimated effects across years, grades, and subjects, but most estimates were not statistically significant.

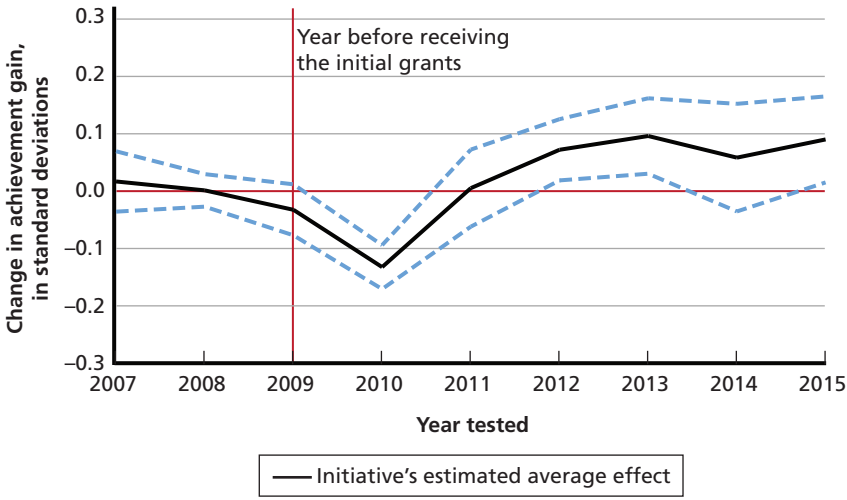
Table 13.2
Visualization of the Initiative's Impacts on Achievement in Pittsburgh Public Schools, by Year, Grade, and Subject

	Reading						Mathematics					
Grade	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
3	–0.07	–0.12	–0.09	–0.08	–0.01	–0.02	–0.07	–0.07	–0.10	0.11	0.14	0.05
4	–0.05	0.04	0.04	–0.06	0.04	0.06	–0.07	–0.03	–0.07	–0.07	0.11	0.03
5	0.04	0.11	0.15	–0.06	0.01	–0.04	0.00	–0.01	0.01	–0.02	0.00	–0.01
6	–0.05	0.03	0.08	–0.01	–0.08	–0.06	–0.11	–0.14	–0.08	–0.16	–0.03	–0.16
7	0.03	0.02	0.08	–0.11	0.05	–0.05	0.03	0.00	–0.11	–0.09	0.12	–0.01
8	0.03	–0.01	0.07	–0.08	–0.03	–0.01	0.00	0.00	0.02	–0.13	0.05	0.06
3–8	**–0.03**	**0.04**	**0.05**	**–0.04**	**0.02**	**–0.01**	**–0.04**	**0.01**	**–0.02**	**–0.01**	**0.10**	**0.01**
11	–0.13	0.01	0.07	0.10	0.06	0.09	N/A	N/A	N/A	N/A	N/A	N/A

NOTE: Red indicates statistically significant negative effects, gray indicates non–statistically significant effects, and green indicates positive significant effects. (Criterion for statistical significance is $p < 0.05$.) The "3–8" row is from an analysis that combines these grades, weighted by enrollment in each grade.

Figure 13.8 shows the initiative's impact on achievement in HS reading.[13] The analysis shows a significant positive effect on overall achievement gains in HS reading (grade 11) in three of the four most-

Figure 13.8
Estimates of the Initiative's Effect on Reading Achievement in PPS HSs



NOTE: The years are the spring terms, so *2010* represents the spring of the 2009–2010 school year. A positive value indicates that the initiative had greater achievement gains than comparison schools. A negative value indicates that the initiative had lower achievement gains than comparison schools. The dashed blue lines indicate the 95-percent confidence interval.

RAND *RR2242-13.8*

[13] In 2013, Pennsylvania introduced the Keystone Exams, which replaced the PSSA in HSs. In contrast to the PSSA, which was 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 grade 11 PSSA reading test, whereas the algebra I Keystone test cannot be used as a comparable replacement for the grade 11 PSSA math test. Therefore, for the HS grades, we evaluate the initiative's impact only on reading achievement.

recent school years, including 2014–2015.[14] For black students, the estimated positive impacts in reading are also statistically significant in each of the final four years of the initiative, as shown in Table U.2 in Appendix U (grade 11 new model estimates).

For PPS dropout rates, the initiative was associated with a reduction in 2014–2015, although the effects were not statistically significant.[15]

SCS

After negative effects in the first years of the initiative, there were no statistically significant effects on math and reading achievement in lower grades in 2014–2015.[16] We find that schools in SCS

---

[14] These estimates are based on a different model from the one we used in Gutierrez, Weinberger, and Engberg, 2016. Our original standard model used only HSs that were in operation by 2008–2009. However, only a small number of these HSs remained open, without changes, in 2014–2015. We excluded from the original standard DiD analysis HSs that merged or opened after 2008–2009. Our revised DiD model includes these schools. We think that the revised model better captures district-wide effects of the initiative because of the high attrition of HSs in the original sample, a phenomenon that we do not see in the other IP sites. The original standard model that excluded new HSs in PPS found a negative effect on student achievement in grade 11 reading compared with that in similar schools in other Pennsylvania districts (see Table U.2 in Appendix U).

[15] The Pennsylvania methodology for calculating graduation rates changed in school year 2014–2015. To maintain consistency, we have updated estimates only on dropout rates, not graduation rates, for that year. Previous estimates of the initiative's effect on graduation rates are available in Gutierrez, Weinberger, and Engberg, 2016. We also updated the estimates of the initiative's effect on dropout rates in earlier years to reflect that dropout rates are being weighted by grade enrollment in each school. We did this to standardize the methodology used across all initiative sites.

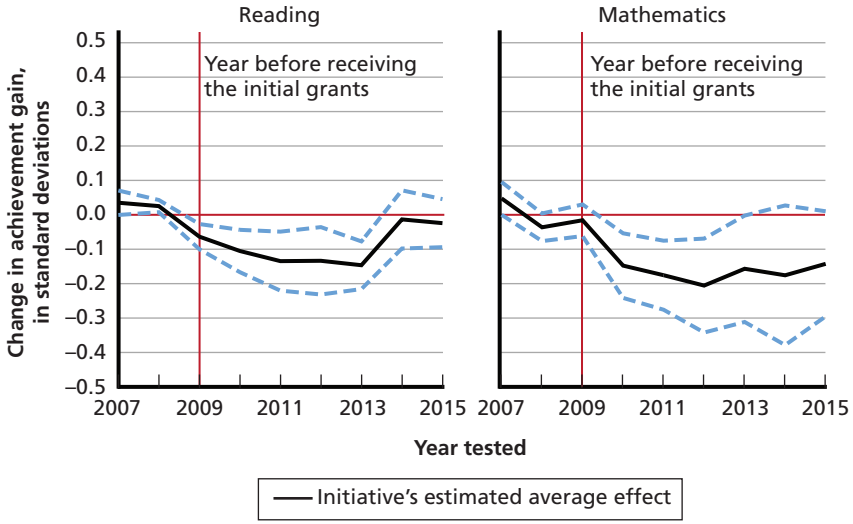
The DiD model for dropout rates does not limit the sample of HSs to those that were opened before the initiative began.

[16] Legacy MCS merged with legacy SCS just before the 2013–2014 school year. Schools that were originally in legacy SCS were not part of the initiative until after the merger. Thus, we do not expect any impact of the initiative on these legacy SCS schools for the 2013–2014 school year and probably only a small impact in 2014–2015. Therefore, we excluded these schools from the impact analyses for all years. In other words, we analyze the initiative's impact only on schools that were originally in the legacy MCS district. This restriction also implies that our analysis is not affected by the changes in the district boundaries that occurred from 2013–2014 to 2014–2015, with many of the suburbs of legacy SCS leaving the newly merged district and creating their own districts. It is worth noting that this restriction

experienced fewer gains in average math achievement in lower grades in the 2014–2015 school year than comparable schools in other Tennessee districts did, but the effect was not statistically significant. As Figure 13.9 shows, the estimated impacts on math achievement improved after a large dip in the first three years after the initiative began. A similar dip and rebound occurred for reading achievement in lower grades.

The impact estimates for SCS should be interpreted with caution because of the difficulty in finding similar schools and districts in the

Figure 13.9
Estimates of the Initiative's Effect on Average Achievement in SCS Grades 3 Through 8



NOTE: The years are the spring terms, so *2010* represents the spring of the 2009–2010 school year. A positive value indicates that the initiative had greater achievement gains than comparison schools. A negative value indicates that the initiative had lower achievement gains than comparison schools. The dashed blue lines indicate the 95-percent confidence interval.

RAND *RR2242-13.9*

(i.e., focusing only on legacy MCS schools) was not made in other data-collection activities of the IP evaluation, such as the surveys of teachers and SLs, because they did not focus on impact.

state, as indicated in Gutierrez, Weinberger, and Engberg, 2016. For example, more than 80 percent of the students in the schools included in the analysis were black—a far higher proportion than the 15 percent in the rest of the state. We should also note that we cannot calculate the initiative's impact for black students or for other subgroups (e.g., low-income students) because Tennessee does not provide data on average performance by subgroup in each school, grade, and subject.

Unlike in the the other IP sites, we find no statistically significant positive effect on achievement in SCS in any grade, subject, or year, as shown in Table 13.3. Nevertheless, Table 13.3 shows a rebound in the estimated impacts—from negative effects to 0 or nonsignificant effects—across all grades 3 through 8.

We could not estimate the initiative's impact on HS state test scores because changes to EOC tests disrupted the measurement of HS achievement (Gutierrez, Weinberger, and Engberg, 2016). However, the initiative's effects were mixed for HS attendance, graduation, and dropout rates in 2014–2015. Although attendance rates improved, graduation rates declined, and there was no significant impact on the dropout rate.

For lower grades, we found no significant effects on attendance and promotion rates for 2014–2015. As with lower-grade achievement, these estimates should be interpreted with caution because of the difficulty of finding adequate comparison schools and districts.[17]

Findings for the CMOs

In this section, we present the initiative's impact on student outcomes in the CMOs. We did not include the CMOs in our prior report (Gutierrez, Weinberger, and Engberg, 2016) because California, where

[17] There are a few very small changes to previously reported estimates of the initiative's impact on SCS test scores, dropout rates, graduation rates, and attendance rates for all years. The revisions reflect two changes to the analysis sample: First, we excluded all schools that were part of legacy SCS before it merged with legacy MCS. Second, because schools that transferred into the state-run ASD were subject to the initiative only until the transfer, we excluded these schools after they were transferred. In Gutierrez, Weinberger, and Engberg, 2016, we discussed these data issues, but we did not incorporate them into the school sample selection.

Table 13.3
Visualization of the Initiative's Impact on Achievement in SCS, by Year, Grade, and Subject

	Reading						Mathematics					
Grade	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
3	−0.17	−0.21	−0.15	−0.12	0.03	0.00	−0.28	−0.16	−0.13	−0.12	−0.03	−0.10
4	−0.05	−0.10	−0.12	−0.16	−0.11	−0.07	−0.03	−0.17	−0.08	−0.02	−0.12	−0.09
5	−0.13	−0.18	−0.09	−0.20	0.04	−0.07	−0.17	−0.21	−0.32	−0.26	−0.04	−0.14
6	−0.09	−0.19	−0.19	−0.11	−0.22	−0.10	−0.19	−0.31	−0.30	−0.14	−0.30	−0.18
7	−0.11	−0.24	−0.16	−0.15	−0.03	−0.14	−0.14	−0.27	−0.22	−0.10	−0.27	−0.13
8	−0.01	−0.16	−0.08	−0.10	−0.08	−0.03	−0.10	−0.24	−0.29	−0.22	−0.22	−0.17
3–8	**−0.10**	**−0.13**	**−0.13**	**−0.14**	**−0.01**	**−0.02**	**−0.15**	**−0.27**	**−0.20**	**−0.16**	**−0.17**	**−0.14**

NOTE: Red indicates statistically significant negative effects, gray indicates non–statistically significant effects, and green indicates positive significant effects. (Criterion for statistical significance is $p < 0.05$.) The "3–8" row is from an analysis that combines these grades, weighted by enrollment in each grade.

most of the CMO schools are located, changed its student assessment system in the 2013–2014 school year and did not publish test scores for that year.[18] Some of the CMOs have too few schools to sustain a credible statistical analysis on their own, so we group the four CMOs together for this analysis. However, we also conduct a separate analysis for the two larger CMOs in the initiative: lower-grade achievement for Aspire, which manages the most elementary schools and MSs, and higher-grade achievement for Green Dot, which manages the most HSs included in the analysis sample.

A difference between the CMOs and the traditional districts that is important for our evaluation is that CMOs are more likely than districts to gain or lose schools from year to year.[19] CMOs are more likely than districts to create new schools, acquire district-run schools, and close schools for noncompliance with their goals. Thus, CMOs' average performance is more likely than that in traditional school districts to be subject to changes in the mix of schools. To avoid this complication and to keep the size and composition of schools relatively stable over time, we include in the analysis only CMO schools that were open by the 2008–2009 school year. Like in the rest of the IP sites, we compare the CMO schools in our sample with similar schools in the state (both traditional and charter), controlling for demographic composition and preinitiative proficiency levels.[20] Another difference from the districts is

---

[18] Alliance has schools only in California; Aspire has schools in California and Tennessee; Green Dot has schools in California, Tennessee, and Washington; and PUC has schools in California and New York. We include only the CMOs' California schools in this evaluation because they are the clear majority of the CMOs' schools and are the schools for which we can evaluate the impact using a common metric (i.e., the California statewide student achievement system).

Given that the reforms were implemented gradually, we considered that having data up to only school year 2012–2013 was not enough to evaluate the initiative's impact in the CMOs. Therefore, we did not report CMO estimates in our earlier report.

[19] For instance, when Locke High School converted into a Green Dot School, the CMO's student body increased more than 33 percent and changed substantially in student composition (Rubin and Blume, 2007).

[20] In an alternative analysis, we compared the CMO schools only with other charter schools in California. The estimated results are similar to the ones presented in this chapter, which compare the CMO schools with all other schools in California.

that some of the CMOs operate schools in multiple states. Our analysis uses only the CMO schools in California, which contains the vast majority of these CMOs' schools.

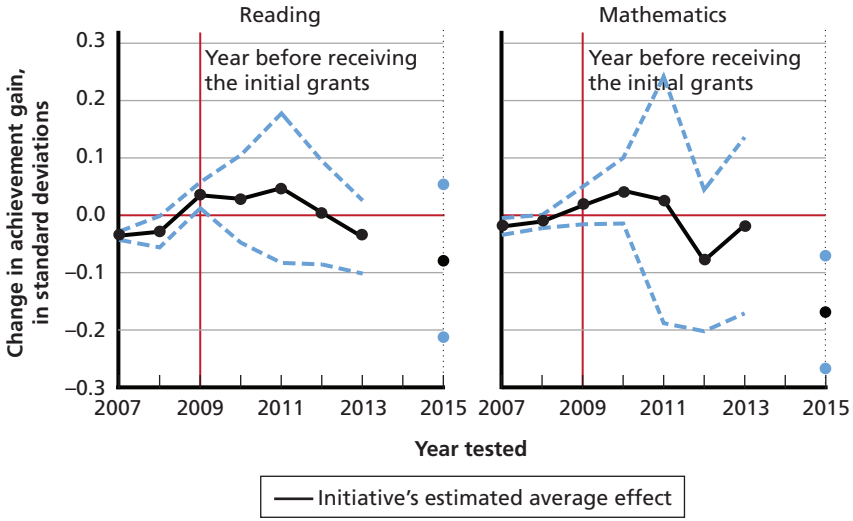
CMOs Combined

In 2014–2015, the initiative had negative effects on average achievement in lower-grade math but positive effects on HS reading achievement in the CMOs. The effect on average achievement in lower-grade reading was also negative but not statistically significant. California changed its standardized testing in 2013–2014.[21] Test scores under the new testing system were first released in 2014–2015. Figure 13.10 indicates small, mostly negative effects of the initiative on average reading and math achievement in the CMOs in lower grades compared with those in similar California schools before the change in tests and then a large drop in the initiative's estimated impact after the new test was implemented. The large drop might be a continuation of a trend that started earlier, or it might reflect differential performance due to the shift in tests. Although we have no reason to think that the rest of the state adapted curriculum and instruction to the new standards faster or differently from the CMOs, that is one possible interpretation of the results for 2014–2015 that we cannot rule out. It is also important to keep in mind that these estimates refer only to CMO schools that were open by the 2008–2009 school year and not those that opened or were acquired by the CMOs after the initiative started.

Figure 13.10 shows positive preinitiative trends that indicate that student outcomes at CMO schools were improving before the initiative compared with those at similar schools in California. After the initiative, these positive trends continued for a few years and then reversed. What do the preinitiative trends suggest about the initiative's impact? On the one hand, the positive preinitiative trends might reflect true

[21] In the 2013–2014 school year, California revamped its assessment testing for all grades from the Standardized Testing and Reporting system to the new California Assessment of Student Performance and Progress. This assessment, California's version of the Smarter Balanced assessment, was created to evaluate new standards that emphasize analytical thinking, problem solving, and communication skills (i.e., the Common Core State Standards).

Figure 13.10
Estimates of the Initiative's Effect on Average Achievement in CMO Schools Grades 3 Through 8



NOTE: We show the 2014–2015 test results as disconnected points because there were no test results released in 2013–2014. The years are the spring terms, so *2010* represents the spring of the 2009–2010 school year. A positive value indicates that the initiative had greater achievement gains than comparison schools. A negative value indicates that the initiative had lower achievement gains than comparison schools. The dashed blue lines indicate the 95-percent confidence interval.

RAND *RR2242-13.10*

underlying structural improvements, so we would have expected them to continue in the absence of the initiative. On the other hand, the positive preinitiative trends might reflect only short-term, temporary, random deviations, in which case we would expect an eventual return to preinitiative average levels (i.e., a reversion to the mean). Although we cannot determine with certainty which scenario is true, we think that the size of the negative estimates, especially in school year 2014–2015, is too large to be explained by only a reversion-to-the-mean scenario.[22]

[22] We also investigated models that used pre-2010 district and CMO trends in addition to levels to predict later outcomes. We did not find that modeling these early trends added to the models' predictive power, further lending support to our assumption that the early trends

Estimated impacts were also negative in 2014–2015 for lower-grade achievement in math and reading for black, Hispanic, and high-poverty students. Although the impact estimates in all subgroups have the same patterns as those for the overall student population, they are generally larger, as shown in Table U.4 in Appendix U.

Before the 2014–2015 school year, most of the initiative's estimated impacts on achievement in different grades, subjects, and years were either positive or not statistically significant, especially in reading, as shown in Table 13.4. In 2014–2015, the effects were negative across all grades for math and in grades 4 through 6 for reading. As discussed earlier, because no information is available for 2013–2014, it is difficult to know whether the lower results in 2014–2015 reflect the continuation of existing negative trends or differences in proficiency in the types of knowledge measured by the new state assessments.

In contrast with what we saw in lower grades, the initiative appears to have had a positive impact on HS reading achievement (grade 11) in the CMOs, as shown in Figure 13.11.[23] The initiative's impact increased from a small, negative, and non–statistically significant effect in 2012–2013 to a positive and statistically significant effect in 2014–2015. This positive effect on HS reading in 2014–2015 held true for black students. For Hispanic and high-poverty students, the effects were positive but not statistically significant (see Table U.4 in Appendix U).

Although the initiative had a positive impact on HS reading, we find a negative effect on the graduation rate in 2014–2015 and a positive but not significant increase in the dropout rate (see Table U.4 in Appendix U). One possible interpretation is that the improvement in reading achievement is due to an increase in the number of less academically successful students dropping out.

We also find, for 2014–2015, a positive and significant impact of the initiative on the CAHSEE math results and a small and nonsig-

---

reflect random variation rather than structural changes. Therefore, we do not model pre-IP trends in any of the districts or CMOs.

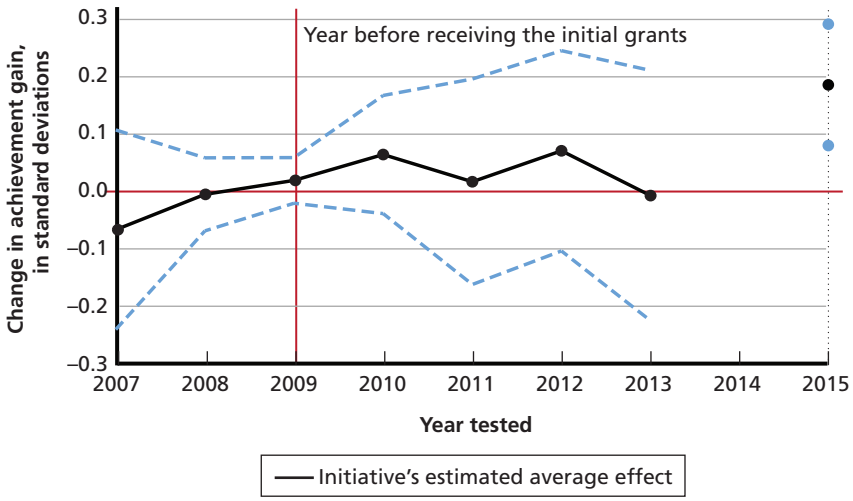
[23] Except for the CAHSEE tests, on which we report separately, math testing was not conducted for HS grades before 2015.

Table 13.4
Visualization of the Initiative's Impacts on Achievement in CMO Schools, by Year, Grade, and Subject

	Reading						Mathematics					
Grade	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
3	0.11	0.17	0.18	0.11	N/A	–0.01	0.00	–0.07	–0.01	0.06	N/A	–0.11
4	0.17	0.11	0.02	0.12	N/A	–0.21	0.03	–0.07	–0.17	–0.03	N/A	–0.23
5	0.10	0.09	–0.02	–0.01	N/A	–0.14	0.05	0.06	–0.16	–0.17	N/A	–0.23
6	–0.01	0.01	0.01	–0.05	N/A	–0.13	–0.02	0.02	–0.05	–0.15	N/A	–0.29
7	0.04	0.02	0.00	0.01	N/A	–0.07	0.20	0.06	–0.08	0.03	N/A	–0.24
8	–0.03	0.07	0.04	–0.16	N/A	–0.22	N/A	N/A	N/A	N/A	N/A	N/A
3–8	**0.03**	**0.05**	**0.01**	**–0.04**	**N/A**	**–0.08**	**0.04**	**0.03**	**–0.08**	**–0.02**	**N/A**	**–0.17**
11	0.07	0.02	0.07	–0.01	N/A	0.19	N/A	N/A	N/A	N/A	N/A	N/A

NOTE: Red indicates statistically significant negative effects, gray indicates non–statistically significant effects, and green indicates positive significant effects. (Criterion for statistical significance is $p < 0.05$.) The "3–8" row is from an analysis that combines these grades, weighted by enrollment in each grade. Grade 8 math tests in California are EOC, rather than end-of-grade, assessments.

Figure 13.11
Estimates of the Initiative's Effects on Reading Achievement in CMO HSs



NOTE: We show the 2014–2015 test results as disconnected points because there were no test results released in 2013–2014. The years are the spring terms, so *2010* represents the spring of the 2009–2010 school year. A positive value indicates that the initiative had greater achievement gains than comparison schools. A negative value indicates that the initiative had lower achievement gains than comparison schools. The dashed blue lines indicate the 95-percent confidence interval.

RAND *RR2242-13.11*

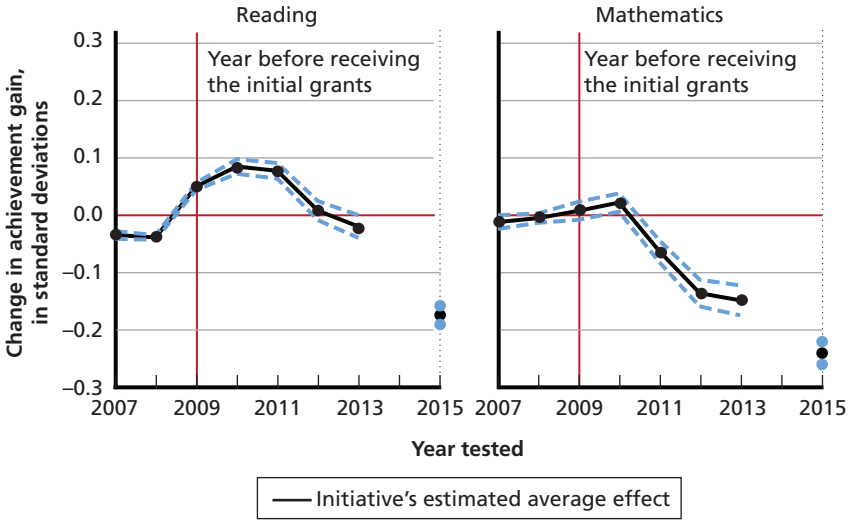
nificant negative effect on CAHSEE reading results (see Table U.4 in Appendix U).[24]

Aspire

In 2014–2015, the initiative had negative effects on average math and reading achievement in lower grades at Aspire, which we analyze separately as the largest operator of elementary schools and MSs among the CMOs. We find a negative trend in the impact estimates for Aspire schools with large negative impacts for 2014–2015, as shown in Figure 13.12. It is worth noting that, although we observe a similar downward trend in the estimates as when analyzing all four

[24] The CAHSEE is an achievement test that was a graduation requirement for students in California public schools. This test was suspended effective January 1, 2016.

Figure 13.12
Estimates of the Initiative's Effect on Average Achievement in Aspire Grades 3 Through 8

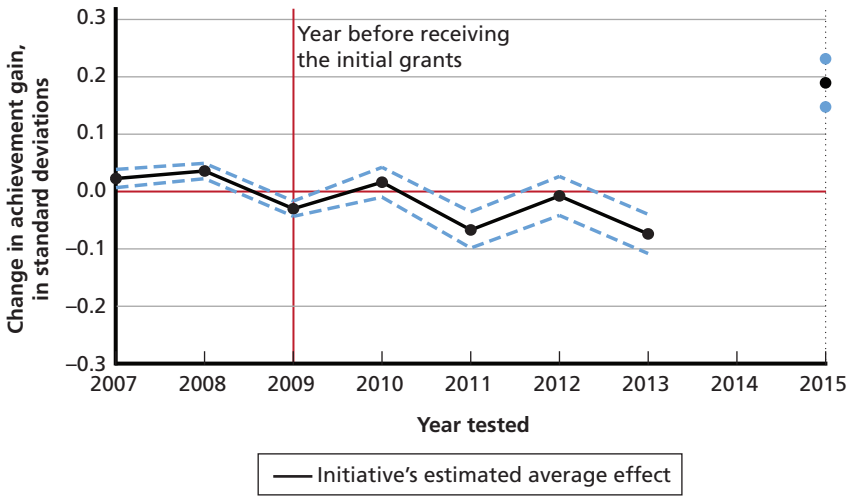


NOTE: We show the 2014–2015 test results as disconnected points because there were no test results released in 2013–2014. The years are the spring terms, so *2010* represents the spring of the 2009–2010 school year. A positive value indicates that the initiative had greater achievement gains than comparison schools. A negative value indicates that the initiative had lower achievement gains than comparison schools. The dashed blue lines indicate the 95-percent confidence interval.

RAND *RR2242-13.12*

CMO schools together (Figure 13.13), the confidence intervals in Figure 13.11 are tighter, and most of the estimated effects are statistically significant for Aspire. This highlights the limitations of modeling student outcomes for the CMOs as one group: The CMOs vary in their implementation of policies and levers associated with the initiative, and some might be subject to shocks (e.g., changes in leadership) not shared by other CMOs. In this light, the combined estimates for CMO lower grades suggest that the other CMOs are faring better than Aspire, but we do not have enough data to precisely estimate how much better they are performing.

Figure 13.13
Estimates of the Initiative's Effects on Reading in Green Dot HSs



NOTE: We show the 2014–2015 test results as disconnected points because there were no test results released in 2013–2014. The years are the spring terms, so *2010* represents the spring of the 2009–2010 school year. A positive value indicates that the initiative had greater achievement gains than comparison schools. A negative value indicates that the initiative had lower achievement gains than comparison schools. The dashed blue lines indicate the 95-percent confidence interval.

RAND *RR2242-13.13*

Table U.5 in Appendix U shows that the impact estimates across all subgroups of Aspire students in 2014–2015 are negative, relatively large, and statistically significant.

The negative effect on lower-grade math achievement began shortly after the start of the initiative, as shown in Table 13.5. In comparison, the effects on reading achievement were more mixed across grades and years, before the change in California's assessment tests. For instance, for 2012–2013, we found positive effects on reading achievement in grades 3 and 4 and negative effects in grades 6 and 8.

Green Dot

For 2014–2015, the initiative had mixed effects on HS outcomes at Green Dot; there were positive effects on reading achievement (grade 11) and the CAHSEE math results, but there were negative

Table 13.5
Visualization of the Initiative's Impacts on Achievement in Aspire, by Year, Grade, and Subject

	Reading						Mathematics					
Grade	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
3	0.11	0.17	0.18	0.11	N/A	–0.01	0.00	–0.07	–0.01	0.06	N/A	–0.11
4	0.17	0.11	0.02	0.12	N/A	–0.21	0.03	–0.07	–0.17	–0.03	N/A	–0.23
5	0.10	0.09	–0.02	–0.01	N/A	–0.14	0.05	0.06	–0.16	–0.17	N/A	–0.23
6	0.03	0.01	0.02	–0.11	N/A	–0.17	–0.02	–0.12	–0.14	–0.33	N/A	–0.32
7	0.08	0.05	–0.07	0.01	N/A	–0.18	0.20	–0.04	–0.22	–0.20	N/A	–0.34
8	–0.02	–0.02	–0.07	–0.23	N/A	–0.38	N/A	N/A	N/A	N/A	N/A	N/A
3–8	**0.09**	**0.08**	**0.01**	**–0.02**	**N/A**	**–0.17**	**0.02**	**–0.06**	**–0.14**	**–0.15**	**N/A**	**–0.24**

NOTE: Red indicates statistically significant negative effects, gray indicates non–statistically significant effects, and green indicates positive significant effects. (Criterion for statistical significance is $p < 0.05$.) The "3–8" row is from an analysis that combines these grades, weighted by enrollment in each grade. Grade 8 math tests in California are EOC, rather than end-of-grade, assessments.

effects on the CAHSEE reading results and the graduation rate. In addition, the dropout rate increased. We analyze Green Dot separately because it was the largest operator of HSs among the CMOs. Figure 13.13 shows the initiative's estimated impact on achievement in HS (grade 11) reading for Green Dot. Contrary to the findings when we pooled the CMO schools together (Figure 13.10), we find negative impacts of the IP initiative in the years shortly after the initiative began. However, we estimate a positive impact for 2014–2015. We also found this positive impact for black students, as shown in Table U.6 in Appendix U. For Hispanic and high-poverty students, the estimated effects in 2014–2015 were also positive but not statistically significant.

The evidence on other HS outcomes is mixed. For 2014–2015, the initiative had a positive impact on the CAHSEE math results but a negative effect on CAHSEE reading results. We also find a negative effect on the graduation rate and an increase in the dropout rate (see Table U.6 in Appendix U). Again, this could suggest that the increase in the average achievement of grade 11 students might be a result of more low-performing students dropping out.

Summary

This chapter presents teachers' and SLs' opinions about the initiative's effect on student performance and statistical analyses of the impact of the initiative on measured student outcomes (both achievement test results and nontest outcomes, such as graduation rates). Neither source of information indicates any clear evidence of positive effects of the initiative in the IP sites.

In the three districts, typically less than 50 percent of teachers agreed that the teacher-evaluation system would ultimately benefit students, and that percentage declined over the years of the reform. In the CMOs, agreement was notably higher. SLs were more likely than teachers to think that the reform would benefit students in the long run, but they also became less likely to agree over time. Also, although most SLs indicated that they were satisfied with the quality

of their current teaching staff, there was not a consistent increase in the reported perception of quality since the start of the initiative.

In terms of measured student outcomes, although there were some positive and statistically significant effects for some outcomes in some sites and in some years, the overall pattern does not suggest that the initiative led to broad-based, sustained gains in any of the sites. Table 13.6 summarizes the initiative's impact across sites for school year 2014–2015, the most recent year analyzed in this chapter.

For HCPS, we estimate nonsignificant negative impacts of the initiative on math achievement in lower grades and virtually zero effect on reading for the school year 2014–2015. We also find nonsignificant effects of the initiative on lower-grade achievement in prior years, except for statistically negative effects on math in 2011–2012. Regarding HS, we find zero effect on reading achievement in 2014–2015, although it is important to note that there was a trend over the three final years of sustained increases in the estimated effects, eliminating the negative impacts estimated for earlier years.

For PPS, our estimates indicate that there were no statistically significant persistent effects on average achievement gains in lower-grade reading and math. Except for a few grades in some years, most of the estimates were either nonsignificant or negative. In the case of reading achievement in HS, we find that, after an initial drop in the years after

Table 13.6
Summary of the Initiative's Impacts on Achievement Across Sites, 2014–2015, in Standard Deviations of State Test Scores

	Grades 3–9		HS	
Site	Mathematics	Reading	Mathematics	Reading
HCPS	–0.046	0.001	N/A	0
PPS	0.006	–0.011	N/A	0.091
SCS	–0.14	–0.022	N/A	N/A
CMOs	–0.167	–0.079	N/A	0.186

NOTE: Red indicates statistically significant negative effects, gray indicates non–statistically significant effects, and green indicates positive significant effects. (Criterion for statistical significance is $p < 0.05$.)

the initiative began, we see a positive trend in PPS schools compared with other schools in Pennsylvania. In 2014–2015, students in PPS had larger achievement gains in HS reading than the rest of the state. This effect was higher for black students. We also find a reduction in the dropout rate in some years after the initiative, although the effects were not statistically significant in 2013–2014 and 2014–2015.

For SCS, we find strong negative effects of the initiative on lower-grade math and reading achievement in the years after the initiative began but also improvements in recent years. For 2014–2015, we find no statistically significant effects on academic achievement in lower grades. We note that these results should be interpreted with caution because of the difficulty in finding schools in the rest of the state that were similar in demographic characteristics (which could serve as a good comparison group).

For the CMOs, we find positive but not significant effects on average lower-grade reading and math achievement in the initial years after the initiative began. But we also find a reversal of these positive estimates, which is more noticeable in 2014–2015, after the change in the student assessment system. For 2014–2015, we find negative effects of the initiative on average achievement in lower-grade reading and math, although the effect is significant only for math. For HS, we find a large positive effect on achievement in HS reading in 2014–2015, after obtaining not statistically significant effects for prior years. However, at the same time, we also find a negative effect on the graduation rate in 2014–2015.

In conclusion, six years after the IP initiative began, a widespread positive impact on student outcomes is not in evidence. The impacts that the broad set of reforms in teacher evaluation and workforce management embodied in the IP initiative has had on student outcomes have varied considerably across sites, subjects, and school levels. For the most recent year, 2014–2015, the impacts have been mostly zero or negative or, in some cases, mixed across grades and subjects.

There are three important caveats in interpreting the results on students' outcomes presented in this chapter.

First, it might take longer than reported here for the initiative to affect student outcomes as the sites continue implementing, and adjust-

ing, the reforms designed to increase effective teaching. No benchmarks are available from similar district-level interventions to guide expectations regarding the size or timing of 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 models, such as Direct Instruction, School Development Program, and Success for All. The authors found that their effect sizes were similar in schools that had implemented these reforms for up to four years (on average, around 0.15 standard deviations). But the effects were almost double for schools that had implemented these reforms for five or six years (0.25 standard deviations) and increased to more than 2.5 times (0.39 standard deviations) for schools with seven years of implementation. Thus, given that most sites have taken multiple years to implement the broad set of reforms since the launch of the initiative in 2009–2010, the results we present in this report might not fully capture the long-run effects of combining rigorous teacher evaluation with changes in workforce management practices. Those effects might be larger in the future, and the foundation has decided to continue this part of the study for two more years to see whether there are increases during this period.

Second, some state-level policy changes during this period influenced both the IP sites and other sites in those states, and our estimates of impacts cannot determine the initiative's effect had these changes not occurred. The Bill & Melinda Gates Foundation and other organizations have worked at the state and national levels to promote HC reforms to improve the quality of education and student outcomes. In fact, since the start of the IP initiative, most states have implemented legislation requiring student achievement to be incorporated into teacher evaluation (Gutierrez, Weinberger, and Engberg, 2016). Two of the IP 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-evaluation measures. The impact estimates presented in this report should be interpreted as the improvements in student performance that can be attributed to the IP initiative over and above

any improvement resulting from other state- or national-level policy changes.

Third, evidence from other district-level interventions that have targeted HC improvement, use of data, and technical assistance seems to suggest that expected impacts in terms of student test scores are relatively modest. For example, a RAND team evaluated the New Leaders program, which recruited and trained school principals in ten districts (S. Gates et al., 2014) and found moderate effects on mathematics and reading scores for students who attended, for three years, a school led by a New Leaders principal. Those authors found no significant effects in HS achievement in math or reading. Other studies (Carlson, Borman, and Robinson, 2011; Strunk and McEachin, 2014) have also found small (and sometimes not statistically significant) effects of district-wide interventions aimed at encouraging data assessments and providing training and assistance.[25] Given that these previous experiences seem to suggest that the effects of district-level interventions are moderate at best, it is important to acknowledge that our estimates have confidence intervals that are wide enough that we cannot rule out small to moderate positive (or negative) effects in most cases.

---

[25] More details about these district-level interventions and their impacts on student outcomes, as well as the impacts of school-level interventions, are available in our previous report (Gutierrez, Weinberger, and Engberg, 2016).

CHAPTER FOURTEEN

Summary and Conclusions

In Brief

The IP initiative, designed and funded by the Bill & Melinda Gates Foundation, was a major effort to dramatically improve student outcomes—particularly HS graduation and college attendance among LIM students—by increasing students' access to effective teaching. The core of the theory underlying the initiative was the development of a robust measure of TE that each participating site could use to improve its teacher development and teacher management efforts. Seven sites—three school districts and four CMOs—participated in the initiative from 2009–2010 through 2015–2016, and we followed their efforts for six years starting in 2010–2011.

Our research shows that the sites enacted TE measures that combined systematic classroom observation, teachers' contributions to student achievement growth, and other factors, and many educators and site leaders reported benefits from doing so. For example, majorities of teachers we surveyed reported receiving useful feedback from observations and using the information from the TE measures to change their instruction. The sites also invested in new IT to enable the creation and use of the measures of effective teaching. In addition, the sites modified other HR policies to align with or take advantage of their newly developed TE measures, including policies related to recruitment, hiring, placement, and transfer; dismissal and tenure; PD; and compensation and CLs.

However, the sites did not implement these other aspects of the initiative as fully as the developers might have expected. For example, all teachers received TE ratings, but very few teachers were classified as ineffective; the sites struggled to deliver evaluation-linked PD; they offered relatively

small performance-based bonuses to relatively large proportions of eligible teachers; and although they created some specialized leadership roles, none created fully developed CLs.

Overall, the initiative did not achieve its stated goals for students, particularly LIM students. By the end of 2014–2015, student outcomes were not dramatically better than outcomes in similar sites that did not participate in the IP initiative. Furthermore, in the sites where these analyses could be conducted, we did not find improvement in the effectiveness of newly hired teachers relative to experienced teachers; we found very few instances of improvement in the effectiveness of the teaching force overall; we found no evidence that LIM students had greater access than non-LIM students to effective teaching; and we found no increase in the retention of effective teachers, although we did find declines in the retention of ineffective teachers in most sites.

We also examined variation in implementation and outcomes across sites. Although sites varied in context and in the ways in which they approached the levers, these differences did not translate into differences in ultimate outcomes. Although the sites implemented the same levers, they gave different degrees of emphasis to different levers, and none of the sites achieved strong implementation or outcomes across the board.

Unfortunately, the evaluation cannot identify the reasons the IP initiative did not achieve its student outcome goals by 2014–2015. It is possible that the reforms are working but we failed to detect their effects because insufficient time has passed for effects to appear. It is also possible that the other schools in the same states we use for comparison purposes adopted similar reforms, limiting our ability to detect effects. However, if the findings of no effect are valid, the results might reflect a lack of successful models on which sites could draw in implementing the levers, problems in making use of teacher-evaluation measures to inform key HR decisions, the influence of state and local context, or insufficient attention to factors other than teacher quality.

Introduction

This chapter summarizes the key elements and impact of the IP initiative and explores possible explanations for the results to date.[1] The initiative focused on improving sites' HR systems—their policies and practices related to teacher evaluation (in particular, measuring TE), staffing (recruitment, hiring, placement, tenure, and dismissal), PD, and compensation and CLs. As an initial step, each site was expected to develop a robust composite measure of effective teaching that included, at a minimum, a direct measure of teaching practice and a measure of a teacher's contribution to student achievement growth. The composite measure would then be used to inform better policies related to the other levers: staffing, PD, and compensation and CLs.

The underlying theory of action was informed by research on measuring TE and improving student outcomes, but not all elements of the initiative had been modeled successfully in practice; some required the sites to create new policies or procedures. For example, sites could draw on existing observation instruments (e.g., C. Danielson, 2013), stakeholder surveys (e.g., Tripod), and ways to combine evidence into a robust composite measure (e.g., Kane and Staiger, 2012), but less was understood about using a composite measure to identify struggling teachers, provide evaluation-linked PD, create incentives to retain effective teachers or set guidelines for dismissing ineffective ones, and increase LIM students' access to effective teachers.

What Changes Did the Sites Make?

Most sites began implementing most levers by 2011–2012, and all sites had implemented all levers to some degree by 2013–2014, although the specific new policies and procedures varied in terms of scale (i.e., the

[1] The RAND/AIR team will continue to monitor implementation and impact through 2017–2018, and we will report on longer-term results at that point.

proportion of staff affected), scope (i.e., the range of potential practices included), and quality. Specifically, we note the following:

- Each site adopted an observation rubric that established a standardized view of effective teaching. Sites devoted considerable time and effort to training and certifying classroom observers and to observing teachers on a regular basis (Chapter Three).
- Each site implemented a composite measure of TE that included scores from (1) direct observations of teaching guided by the selected rubric, (2) measures of growth in student achievement, and, in some cases, (3) feedback from students or parents. The composite measure combined the component measures using weights, and it was then used to inform HR decisions. Over time, fewer teachers were rated as ineffective; by the end of the initiative, only 1 to 2 percent were classified as ineffective. This might reflect actual improvement in teaching, but there is some evidence that it is due to other factors, such as increasingly generous ratings on subjective components (e.g., classroom observations) (Chapter Three).

To varying degrees, each site modified its HR policies based on its conception of effectiveness and its TE measure. Specifically, we note the following:

- Although sites varied in the extent to which they modified recruitment, hiring, and transfer policies, most expanded recruitment efforts and streamlined their hiring process. Sites varied in the extent to which they incorporated standards for effective teaching into their candidate review processes (Chapter Four).
- The three districts set specific criteria based on their new evaluation systems to identify low-performing teachers who might be denied tenure, placed on improvement plans, or considered for dismissal or nonrenewal of their contracts. In contrast, the CMOs, which did not offer tenure, took the new evaluation results into consideration but, except for Green Dot, did not establish specific

criteria that triggered action for teachers with low performance (Chapter Five).

- In all the sites, very few teachers were rated ineffective and even fewer were dismissed; for example, about 1 percent of the teacher workforce was dismissed in 2015–2016 in the five sites for which we have information (Chapter Five).
- Principals and other staff recommended PD based on teachers' evaluation results, but sites did not require teachers to participate in recommended PD, nor did they put systems in place to monitor teachers' participation or determine whether teachers' effectiveness improved after participating in specific PD. Over time, they tried to individualize support by putting more emphasis on individual coaching or mentoring as an improvement strategy, particularly for new and struggling teachers, and on using the observation process as a mechanism to frame conversations about instructional improvement (Chapter Six).
- Most of the sites implemented yearly bonus programs, but not permanent salary increases, based on TE ratings; most bonuses were small relative to salary, and a majority of eligible teachers in the sites offering bonuses met the criteria to earn them. One of the CMOs and PPS adopted effectiveness-based salary schedules and planned to continue their use (Chapter Seven).[2]
- All the sites added specialized roles to give teachers more opportunities for leadership, but they did not adopt fully articulated, fully developed CLs as the backers of the initiative envisioned. The CMOs had many such roles, with many teachers taking on specialized tasks for a year; the districts had fewer roles, mostly as peer coaches and mentors for new teachers (Chapter Eight).

[2] The effectiveness-based salary schedule in PPS applied only to teachers hired after 2010–2011. As explained in Chapter Two, it was scheduled to expire in 2015 but was extended for two additional years, as of the writing of this report.

What Outcomes Were Achieved?

The overarching goals of the initiative were to improve the effectiveness of the teacher workforce, increase LIM students' access to effective teaching, and, as a result, produce dramatic gains in student success. However, as of 2014–2015, the desired outcomes had not yet materialized. Our analyses found the following:

- Over time, higher percentages of teachers in each site were classified as E or HE. There is no objective yardstick by which we can determine whether this apparent improvement was real or a result of other factors (Chapter Three).
- Although most SLs reported that sites' recruitment and hiring efforts worked well, and these perceptions improved in some sites, the measured effectiveness of newly hired teachers did not improve over time relative to that of experienced teachers (Chapter Ten).
- More-effective teachers were retained at higher rates than less effective teachers; however, this was true before the start of the initiative, and the retention rate for effective teachers did not improve. On the other hand, the retention rate of less effective teachers declined over time in many of the sites, which is consistent with the goals of the initiative (Chapter Eleven).
- There was very small improvement in effectiveness among mathematics teachers in HCPS and SCS and larger improvement among reading teachers in SCS, but there were also significant declines in effectiveness among reading teachers in HCPS and PPS (Chapter Twelve).
- LIM students did not gain increased access to effective teachers. Although LIM students attended schools with slightly more-effective teachers than non-LIM students did, within those schools, LIM students had slightly less access to effective teachers than non-LIM students did (Chapter Twelve).
- Schools in each of the IP sites, on average, did not have better outcomes (e.g., student achievement, graduation) than similar non-participating schools in their state (Chapter Thirteen). Table 14.1 shows the effects on student achievement by grade span and site

Table 14.1
The Initiative's Estimated Impact on Student Achievement in 2014–2015, by Grade Span and Site

	Grades 3–8		HS	
Site	Mathematics	Reading	Mathematics	Reading
HCPS	–0.05	0.00	N/A	0.00
PPS	0.01	–0.01	N/A	0.09
SCS	–0.14	–0.02	N/A	N/A
The four CMOs	–0.17	–0.08	N/A	0.19

NOTE: We report the values in standard deviations of state test scores. We could not estimate the impact on HS mathematics because students did not take the same secondary mathematics tests. Red indicates statistically significant ($p < 0.05$) negative effects, gray indicates non–statistically significant effects, and green indicates positive significant effects.

for the spring of 2015, the final year for which we have data. In most sites and grade-level ranges, the initiative did not have a significant impact on student achievement in mathematics or reading (the gray cells).[3]

How Much Did the Sites Vary in Implementation and Outcomes?

The previous sections highlight findings that apply broadly to the initiative as a whole. However, because sites exercised considerable control over the design and implementation of their policies and procedures, it is also appropriate to take a site-specific perspective. As described in Chapters Three through Eight, there was variation among the sites in most aspects of implementation. One example for each lever illustrates the kinds of variation that occurred:

- TE measures: HCPS, PPS, SCS, Alliance, and Aspire continued to use their composite TE measures through at least 2015–2016,

[3] As shown in Chapter Thirteen, impact estimates vary by site, subject, and year.

the final year of the initiative, but Green Dot stopped including achievement in its measure, and PUC stopped using a composite measure altogether.

- recruitment, hiring, placement, and transfer: The CMOs and SCS made use of residency programs to recruit and train new teachers, whereas HCPS and PPS did not. For teachers transferring among schools (relevant mainly in the districts), seniority was a consideration in PPS but not in HCPS, and it ceased to be a consideration in SCS as a result of a change in state law.
- tenure and dismissal: In concert with statewide changes, HCPS and SCS modified the requirements for earning tenure, whereas PPS did not (although the definition of *satisfactory performance* changed). The CMOs did not offer tenure.
- PD: The sites differed in the ways and extent to which they used online resources and forums for evaluation-linked PD. HCPS and SCS developed online catalogs in which teachers could easily look up PD opportunities linked to specific dimensions of the observation rubric; SCS and Aspire created extensive repositories of online videos.
- compensation: Only Alliance, Aspire, and PPS adopted effectiveness-based salary schedules, and, in PPS, the new salary schedule applied only to a small proportion of teachers (those hired after July 2010).
- CLs: All the sites implemented specialized instructional leadership positions for teachers, but, in HCPS, less than 1 percent of teachers held such positions, while, in Aspire, more than 50 percent of teachers did.

As these examples show, most of the differences among the sites in how they implemented the IP reforms were in the details. On a broader scale, the initiative looked generally similar in all the sites; they all attended to the same set of levers, and none implemented reforms that differed substantially from the others. Moreover, there is no clear evidence that, overall, some sites were more successful than others in implementing the collective set of IP reforms.

When it comes to intermediate outcomes—the hiring and retention of effective teachers, improving TE, and increasing LIM students' access to effective teachers—we found generally similar outcomes across the sites. Although some positive effects appeared in some sites in some years, we found few consistent patterns indicating more-positive results for some sites than others. For example, across sites, there is no evidence that the effectiveness of newly hired teachers improved during the initiative relative to experienced teachers or that effective teachers were more likely than ineffective ones to be retained. Similarly, in most of the sites, the retention of ineffective teachers declined over the course of the initiative—a success story but not one that points to differences among the sites. With respect to TE, the study-calculated VAM measure of TE improved in HCPS in math but did not improve in reading. In SCS, overall effectiveness improved slightly in mathematics, and it improved considerably in reading. In PPS, there was little change in TE in either mathematics or reading.[4] In most of the sites, LIM students' access to effective teaching did not improve.

In terms of final outcomes, most of the estimates of the initiative's impact on achievement were nonsignificant, although there were significant positive effects on some outcomes in some years in PPS, SCS, Aspire, and Green Dot and significant negative effects on some outcomes in some years in HCPS, SCS, and Aspire. Thus, although the sites varied in context and in the specific ways in which they approached the levers, these differences did not translate into differences in ultimate outcomes. The results imply that, across the board, no site was more effective than any of the others, either in implementing the reforms or in achieving positive outcomes.

[4] We do not present overall effectiveness results for the CMOs because they have considerably fewer teachers and because we are missing a year of CMO test data from when California switched standardized tests in 2014.

Are the IP Levers Being Sustained?

Although the sites' efforts after 2015–2016 are not the focus of this report, we are continuing to gather information from the sites about the extent to which the components of the IP initiative are being sustained. As of 2016–2017, the sites have maintained many of the elements of the initiative, and we take this as an indication that they find these features to be beneficial. The key elements that have become established policy in the sites include the following:

- All the sites incorporated more-systematic teacher evaluation into their culture. They continue to use structured classroom observations to gather information about teaching practice and to structure conversations about improvement. All but two continue to compute a composite TE measure that includes observation scores and a student achievement growth measure (and other factors in some sites) and to use this measure to identify teachers in need of improvement.
- All the sites modified their recruitment and hiring policies to improve the potential effectiveness of new teachers and offer them more initial support. The sites' visions of effective teaching were incorporated in various ways into their reviews of new candidates and their hiring decisions.
- All the sites are retaining or building on the new data systems they put in place to improve teacher evaluation and other HR functions.
- Many of the sites continue to support the specialized instructional positions they created in place of hierarchical CLs, particularly those pertaining to mentoring and coaching.
- Some of the sites are retaining central-office staff positions and departments created to support the initiative.

Other elements of the initiative are not being continued because they were not perceived to provide sufficient benefits or were too costly. Most sites have reduced the number, frequency, or length of classroom observations to reduce the burden on administrators and, in some

cases, to expedite the rapid sharing of information with teachers. Most sites continue various types of performance-based bonuses, but they are not large and are usually available only to teachers in high-need schools. None of the sites created true CLs with positions arranged in a defined hierarchy of increasing responsibility and salary, and none appears interested in doing so.

Potential Limitations of the Methods Used to Estimate Impact

Before turning to potential explanations of why the initiative did not achieve its intended effects, it is worth reiterating two features of our study that might affect the interpretation of the results.[5] First, we studied changes in outcomes over a five-year period, but it might take longer for the elements of the IP initiative to affect student outcomes. Borman et al. (2003) found that large-scale, comprehensive school reforms had substantially larger effects after five and six years than after four years or less. For the districts, we report outcome data from the first to the fifth year after the grants were awarded (for the CMOs, there are fewer estimates because of gaps in the state testing program), but, as we noted in previous chapters, not all elements of the initiative were implemented immediately, and some might take longer to have or manifest an impact. We are continuing to examine outcomes and will issue a report covering the 2015–2016 and 2016–2017 school years at a future date.

Second, the methods we used to estimate impact compare the performance of students who attended schools in the IP sites with an estimate of what their performance would have been if their site had not participated in the initiative. We estimated the counterfactual condition—how they would have performed in the absence of the initiative—based on the performance of similar schools in the same

[5] In Chapter Thirteen and Appendix T, we discuss the methods we used to assess the initiative's impact. We also describe them in more detail in our previous publication (Gutierrez, Weinberger, and Engberg, 2016).

state that did not receive Gates Foundation funding and support to implement comprehensive HR reforms. This estimate assumes that those schools engaged in "business as usual" during this period—that is, that they did not reform their teacher-evaluation systems in ways similar to those intended by the IP initiative. However, some statewide policy changes affected both the IP sites and the schools we used as a comparison, and our estimates cannot remove these influences. In particular, new teacher-evaluation measures with consequences were enacted in three of the states. Tennessee won a federal RTT grant in the first round and Florida in the second, and the grants required both states to implement stronger teacher-evaluation systems. In addition, Pennsylvania enacted tougher teacher-evaluation requirements in 2012, which went into effect in 2013–2014. Thus, our impact estimates reveal how well the IP initiative improved student outcomes *over and above* these statewide efforts.

Discussion

The IP initiative heightened the sites' attention to TE; however, measuring effectiveness and using it as the basis for teacher management and incentives did not appear to lead to gains in student achievement or graduation rates. The evaluation does not tell us why these outcomes were not achieved, but we are willing to speculate—informed by our observations of the sites and the foundation during the past several years—about potential factors that might explain the lack of impact. In some instances, we offer recommendations to address the identified issues. These recommendations might be of value to the foundation, the sites, or others who might be contemplating similar large-scale reform efforts in the future.

Incomplete Implementation and Lack of Successful Models

It is possible that the new policies might not have been implemented with sufficient quality, intensity, or duration to achieve their potential full effect. None of the main policy levers—staffing, PD, or compensation or CLs—was implemented fully or as initially envisioned.

Implementing the levers required the sites to develop specific policies and practices to translate general ideas from the theory of action into operational realities. For many of the levers, the sites had to do this in the absence of successful models. This practical gap was most apparent in the sites' efforts to implement evaluation-linked PD; they did not find successful models (e.g., from other districts) they could observe, adopt, or adapt. Instead, they had to develop their own systems, and they encountered some practical problems in doing so—problems that had still not been fully resolved by the end of the grant period.

The sites' experiences with evaluation-linked PD (and other levers) suggests to us the value of distinguishing between innovation (developing new practices) and implementation (putting developed practices into place). At the beginning of the evaluation, many members of the research team thought that the main challenge of the IP initiative would be effective implementation (i.e., carefully and systematically enacting well-developed ideas). Effective implementation might include such things as explaining new practices to stakeholders, adapting them to the local context, and reallocating resources to support new systems or positions. However, we found that, for many of the levers, the main challenge in enacting the IP theory of action was innovation. The sites had to create new methods or approaches in the absence of functioning models. Our experience suggests that the capacities needed to implement might differ from those needed to innovate. In addition, implementation requires less time than innovation. For example, had there been time, the sites might have allocated staff or hired consultants to analyze existing data to help them understand how well PD efforts were working, revised their PD infrastructure to collect better information about participation and perceived quality, or tried different ways of linking PD to effectiveness to see which worked best. Alternatively, the foundation might have worked with other sites to develop and test prototypes before incorporating them into a larger reform. This is essentially what the foundation did with the MET project; it conducted a research and development effort to determine the best combination of measures to use for measuring

TE. The IP sites would have benefited if similar preparatory work had been done on systems to link effectiveness measures to HR policies.[6]

Problems in Making Use of Teacher-Evaluation Measures

Teacher evaluation was at the core of the initiative, and the sites were committed to using the measures to inform key HR decisions. But, as we described in Chapters Three through Eight, the sites encountered two problems related to these intended uses of the TE measures. First, it was difficult for the sites to navigate the underlying tension between using evaluation information for professional improvement and using it for high-stakes decisions. Second, some sites encountered unexpected resistance when they tried to use effectiveness scores for high-stakes personnel decisions; this occurred despite the fact that the main stakeholder groups had given their support to the initiative in general terms at the outset.

Tension Between Different Purposes for Measures

Researchers distinguish between measures used for summative purposes (to make overall judgments about quality) and measures used for formative purposes (to improve conditions or practices). For a variety of reasons, it is hard to design a measure that is equally good for both purposes because the characteristics of good summative measures differ from those of good formative measures. For example, because summative measures are used to make consequential decisions about people or programs, they need to meet high standards of technical quality (i.e., reliability and validity). Formative measures, which are used to guide improvement, do not need to be as technically sound. Because of the need for quality, summative measures are usually longer and adminis-

[6] Developing and validating methods to increase the value of observation feedback for teacher improvement could be a particularly generative focus for research and development. In particular, much could be learned by helping districts gather systematic data on the recommendations for PD that flow from teacher evaluations, the PD in which teachers actually participate, and changes in their evaluation ratings associated with participation. Sites might also conduct small-scale experiments, randomly assigning teachers to receive different forms of PD based on evaluation results (e.g., more- or less-intensive coaching).

tered only occasionally. Formative measures can be shorter, more frequent, embedded in curriculum, and so forth.

The IP sites hoped that their TE measures could be used both for improvement and for accountability purposes. But these two goals were often in conflict. For example, because effectiveness was going to be used in tenure and dismissal decisions, it was essential that it meet high standards for reliability and validity. Thus, all classroom observers had to be trained and had to pass certification tests to ensure that they were scoring lessons accurately. Furthermore, each teacher had to be observed multiple times during the year for full lessons. On the other hand, sites found that shorter, more-frequent observations could be more useful for improvement purposes because they allowed for more-immediate feedback to teachers. Many sites changed the structure of the observations to reduce the time burden and to better support improvement goals. Sites also found greater support for effectiveness-based improvement policies that involved low or no stakes (e.g., coaching or mentoring informed by measured effectiveness) and greater resistance to policies that threatened compensation or employment.

Resistance to Using Teacher Evaluation for High-Stakes Decisions

During the initiative, the sites adapted their evaluation systems, either formally or informally, to avoid having to dismiss many teachers. The IP initiative was launched with great fanfare and with the endorsement of each site's board, administration, teachers, and local community. During the first two years, most stakeholders had positive attitudes toward it. A change occurred when new policies threatened some teachers with loss of employment or reduction in salary. Teacher organizations, most notably in PPS, began to object and mount public campaigns against the effectiveness measures when high stakes were due to be attached and larger numbers of teachers were threatened. Over time, in most of the sites, fewer and fewer teachers were identified as lower performing, possibly because local norms and expectations influenced how ratings were assigned or where performance levels were set. The lesson we draw from this experience is that reformers should not underestimate the resistance that can arise if changes have major negative consequences for staff employment.

Changes in the State and Local Contexts

It is also possible that changes in the local or state context interfered with the sites' ability to implement the policy levers fully. In Chapter Two, we provided an overview of the context in each site, and, in Chapters Three through Eight, we described some external conditions that hampered implementation of the initiative. The most-unexpected and most-problematic shifts came from political decisions relating to governance and testing and from changes in local leadership. With respect to testing, each of the four states changed its statewide test during the course of the initiative, which necessitated adjustments to the sites' achievement and growth measures and caused some concerns about comparability of scores across years. Other significant but unexpected changes included the merger of legacy MCS with legacy SCS; the Pennsylvania budget crisis that led to teacher furloughs, elimination of the teacher academies, and increasingly contentious district–union relations; the HCPS board's abrupt decision to remove the superintendent; and the California education budget cuts that significantly reduced funding to the CMOs.

In addition, every site except Green Dot and PUC had a turnover in top leadership during the IP initiative. In some places, the new superintendent or director maintained the focus on TE, but, in other sites, the new leaders eliminated parts of the IP reform, slowed implementation, or established other priorities.

Insufficient Attention to Other Factors

Finally, we should mention that the IP initiative might not have achieved its dramatic goals because improvement on that scale requires attention to a broader set of factors. The initiative was appealing, in part, because of its tight focus on TE. Research that informed the initiative's design suggested that attention to TE is necessary to improve student outcomes. Although teachers remain the most salient in-school factor in determining student outcomes, and thus improving teaching is a plausible lever for improvement, differences among teachers explain only a small percentage of the variation in student achieve-

ment.[7] Perhaps a near-exclusive focus on TE is insufficient to dramatically improve student outcomes. Many other factors might need to be addressed, ranging from early childhood education to students' social and emotional competencies, the school learning environment, and family support. We suspect that dramatic improvement in outcomes, particularly for LIM students, will require attention to many of these factors as well.

The Importance of Measuring Implementation

Apart from suggesting lessons about the reform of teacher HC systems, our work also has a few implications for the evaluation of large-scale initiatives. The IP initiative was a multifaceted reform that addressed interrelated policies and procedures in each site. In change efforts such as this, it is important to measure the extent to which each of the new policies and procedures is implemented in order to understand how the specific elements of the reform are related to outcomes. We think it is particularly important to measure "dosage" (i.e., how much particular activities occur, over what period of time, and with what intensity). In the present study, we tried to document the extent to which various levers were enacted in each site. We also obtained the evaluation ratings that teachers received, and we were able to link teachers' ratings to their attitudes, compensation, and retention. However, we could not measure other aspects of dosage (e.g., the extent of feedback teachers received, such as duration of meetings with observers; the number and types of PD activities in which teachers participated; teachers' placement on improvement status). Even with the data we had, we could not disentangle the effects of levers that were enacted simultaneously on the same set of teachers (e.g., the impact of effectiveness-based compensation, separate from that of CLs, on teachers' retention decisions). It seemed to us that the initiative did not place sufficient emphasis from the start on developing the data systems that would have been required to track implementation systematically. In particular, most of the IP

[7] Estimates of the percentage of the variation in student achievement growth that is attributable to variation in teachers range from 1 to 14 percent (American Statistical Association, 2014).

sites did not generally collect information about access to, participation in, or quality of PD, and we could not gather that information on our own. Knowing more about PD dosage would have increased our ability, as well as the sites' ability, to judge the effectiveness of the sites' emerging effectiveness-linked PD efforts. We think that this is a lesson worth considering in future reforms.

Final Thought

A favorite saying in the educational measurement community is that one does not fatten a hog by weighing it. The IP initiative might have failed to achieve its goals because it succeeded more at measuring teaching effectiveness than at using the information to improve student outcomes. Contrary to the developers' expectations, and for a variety of reasons described in the report, the sites were not able to use the information to improve the effectiveness of their existing teachers through individualized PD, CLs, or coaching and mentoring. In the end, the sites were able to measure effectiveness but not increase it.

References

Aaronson, Daniel, Lisa Barrow, and William Sander, "Teachers and Student Achievement in the Chicago Public High Schools," *Journal of Labor Economics*, Vol. 25, No. 1, January 2007, pp. 95–135.

Abshere, Erin, *The Relationship Between Employee Value Propositions and Teacher Employment Decisions*, Los Angeles, Calif.: University of Southern California, Rossier School of Education, dissertation, May 2016. As of February 5, 2018: http://cdm15799.contentdm.oclc.org/cdm/ref/collection/p15799coll40/id/232831

Achievement School District, home page, undated. As of March 1, 2018: http://achievementschooldistrict.org/

Adnot, Melinda, Thomas Dee, Veronica Katz, and James Wyckoff, "Teacher Turnover, Teacher Quality, and Student Achievement in DCPS," *Educational Evaluation and Policy Analysis*, Vol. 39, No. 1, 2017, pp. 54–76.

Allen, Michael B., *Eight Questions on Teacher Recruitment and Retention: What Does the Research Say?* Denver, Colo.: Education Commission of the States, September 2005. As of February 5, 2018: https://eric.ed.gov/?id=ED489332

Alliance College-Ready Public Schools, "Financials," undated. As of January 17, 2017: http://www.laalliance.org/apps/pages/index.jsp?uREC_ID=46278&type=d&termREC_ID=&pREC_ID=159758&hideMenu=1

American Federation of Teachers, "Reclaiming the Promise: Career Ladders," Washington, D.C., November 2013. As of March 28, 2018: http://www.aft.org/sites/default/files/rtp_careerladders.pdf

American Statistical Association, "ASA Statement on Using Value-Added Models for Educational Assessment," Washington, D.C., April 8, 2014. As of February 5, 2018: http://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf

Angrist, Joshua D., and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton, N.J.: Princeton University Press, 2008.

ASD—*See* Achievement School District.

Aspire Public Schools, home page, undated. As of November 13, 2017: https://aspirepublicschools.org

Atteberry, Allison, Susanna Loeb, and James Wyckoff, *Do First Impressions Matter? Improvement in Early Career Teacher Effectiveness*, Cambridge, Mass.: National Bureau of Economic Research, Working Paper 19096, June 2013. As of September 20, 2017:
http://www.nber.org/papers/w19096

Baird, Matthew D., John Engberg, Gerald Hunter, and Benjamin Master, *Improving Teaching Effectiveness: Access to Effective Teaching—The Intensive Partnerships for Effective Teaching Through 2013–2014*, Santa Monica, Calif.: RAND Corporation, RR-1295/4-BMGF, 2016. As of February 5, 2018:
https://www.rand.org/pubs/research_reports/RR1295z4.html

Baker, Bruce D., David G. Sciarra, and Danielle Farrie, *Is School Funding Fair? A National Report Card*, 4th ed., Newark, N.J.: Rutgers Graduate School of Education, Education Law Center, Spring 2015. As of February 5, 2018:
https://drive.google.com/file/d/0BxtYmwryVI00cVZueUhhYlBsdE0/view

Baker, Eva L., Paul E. Barton, Linda Darling-Hammond, Edward Haertel, Helen F. Ladd, Robert L. Linn, Diane Ravitch, Richard Rothstein, Richard J. Shavelson, and Lorrie A. Shepard, *Problems with the Use of Student Test Scores to Evaluate Teachers*, Washington, D.C.: Economic Policy Institute, Briefing Paper 278, August 29, 2010. As of March 28, 2018:
https://www.epi.org/publication/bp278/

Baratz-Snowden, Joan, *Fixing Tenure: A Proposal for Assuring Teacher Effectiveness and Due Process*, Washington, D.C.: Center for American Progress, June 2009. As of February 5, 2018:
https://www.americanprogress.org/issues/education-k-12/reports/2009/06/25/6242/fixing-tenure/

Bell, Courtney A., Drew H. Gitomer, Daniel F. McCaffrey, Bridget K. Hamre, Robert C. Pianta, and Yi Qi, "An Argument Approach to Observation Protocol Validity," *Educational Assessment*, Vol. 17, No. 2–3, 2012, pp. 62–87.

Betebenner, Damian W., National Center for the Improvement of Educational Assessment, *New Directions for Student Growth Models*, briefing slides presented to unknown audience, December 13, 2010.

Bill & Melinda Gates Foundation, *Intensive Partnerships to Empower Effective Teachers: Invitation-Only Request for Proposals*, Seattle, Wash., 2009.

Bireda, Saba, *Devil in the Details: An Analysis of State Teacher Dismissal Laws*, Washington, D.C.: Center for American Progress, June 3, 2010. As of February 5, 2018:
https://www.americanprogress.org/issues/education-k-12/reports/2010/06/03/7886/devil-in-the-details/

Booker, Kevin, and Steven Glazerman, *Missouri's Teacher Career Ladder Program*, Nashville, Tenn.: Peabody College, Vanderbilt University, National Center on Performance Incentives, Working Paper 2008-15, 2008.

Borjas, George J., and Glenn T. Sueyoshi, "A Two-Stage Estimator for Probit Models with Structural Group Effects," *Journal of Economics*, Vol. 64, No. 1–2, September–October 1994, pp. 165–182.

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, 2003, pp. 125–230.

Braun, Henry I., *Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models*, Princeton, N.J.: Educational Testing Service, policy information perspective, 2005. As of September 20, 2017: http://www.ets.org/research/policy_research_reports/publications/report/2005/cxje

Briggs, Derek, and Ben Domingue, *Due Diligence and the Evaluation of Teachers: A Review of the Value-Added Analysis Underlying the Effectiveness Rankings of Los Angeles Unified School District Teachers by the "Los Angeles Times,"* Boulder, Colo.: National Education Policy Center, February 2011. As of March 28, 2018: https://eric.ed.gov/?id=ED516008

Brill, Sam, and Abby McCartney, "Stopping the Revolving Door: Increasing Teacher Retention," *Politics and Policy*, Vol. 36, No. 5, October 2008, pp. 750–774.

California Department of Education, "Schools Chief Tom Torlakson Leads 'Strong Schools for a Strong Economy' Tour," news release 12-91, October 4, 2012.

Campbell, Donald T., "Assessing the Impact of Planned Social Change," *Evaluation and Program Planning*, Vol. 2, No. 1, 1976, pp. 67–90.

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, 2011, pp. 378–398.

Carnegie Forum on Education and the Economy, *A Nation Prepared: Teachers for the 21st Century—The Report of the Task Force on Teaching as a Profession*, Hyattsville, Md., May 1986.

Carnegie Foundation for the Advancement of Teaching, "Carnegie Cost Calculator," undated. As of December 19, 2017: https://carnegiefoundation.co1.qualtrics.com/jfe/form/SV_8oylGoEjGCQotsF

Carver-Thomas, Desiree, and Linda Darling-Hammond, *Addressing California's Growing Teacher Shortage: 2017 Update*, Palo Alto, Calif.: Learning Policy Institute, February 2017. As of February 6, 2018:
https://learningpolicyinstitute.org/product/
addressing-californias-growing-teacher-shortage-2017-update-report

Casabianca, Jodi, and John Engberg, *The Impact of Professional Development Reforms on Teacher Performance in the Intensive Partnership Sites: A Conceptual Framework and a Review of Data Availability*, Santa Monica, Calif.: RAND Corporation, unpublished research, September 2013, not available to the public.

Castellano, Katherine E., and Andrew D. Ho, *A Practitioner's Guide to Growth Models*, Council of Chief State School Officers, February 2013. As of March 28, 2018:
https://scholar.harvard.edu/andrewho/publications/
practitioners-guide-growth-models

Center for Education Policy Research, Harvard University, "SDP Provides New Findings on Teacher Placement and Retention Patterns," press release, July 18, 2012. As of September 20, 2017:
http://cepr.harvard.edu/news/
sdp-releases-new-findings-teacher-placement-and-retention-patterns

Center for Public Education, "Wanted: Good Teachers," October 8, 2008.

Chait, Robin, *Removing Chronically Ineffective Teachers: Barriers and Opportunities*, Washington, D.C.: Center for American Progress, March 10, 2010. As of February 6, 2018:
https://www.americanprogress.org/issues/education-k-12/reports/2010/03/10/7525/
removing-chronically-ineffective-teachers/

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff, *Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood*, Cambridge, Mass.: National Bureau of Economic Research, Working Paper 19424, revised May 2014. As of February 6, 2018:
http://www.nber.org/papers/w19424

Clement, Mary C., "Teachers Hiring Teachers," *Educational Leadership*, Vol. 71, No. 2, October 2013.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor, "Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects," *Journal of Human Resources*, Vol. 45, No. 3, 2010, pp. 655–681.

Coggshall, Jane G., Claudette Rasmussen, Amy Colton, Jessica Milton, and Catherine Jacques, *Generating Teacher Effectiveness: The Role of Job-Embedded Professional Learning in Teacher Evaluation*, Washington, D.C.: National Comprehensive Center for Teacher Quality, research and policy brief, May 2012. As of February 6, 2018:
https://eric.ed.gov/?id=ED532776

Cohen, David K., and James P. Spillane, "Policy and Practice: The Relations Between Governance and Instruction," *Review of Research in Education*, Vol. 18, 1992, pp. 3–50.

Courageous Conversation and Pacific Educational Group, "Courageous Conversations About Race," undated. As of March 31, 2018: https://courageousconversation.com/about/

Cowen, Joshua M., and Marcus A. Winters, "Do Charters Retain Teachers Differently? Evidence from Elementary Schools in Florida," *Education Finance and Policy*, Vol. 8, No. 1, Winter 2013, pp. 14–42.

Cullen, Julie Berry, Cory Koedel, and Eric Parsons, *The Compositional Effect of Rigorous Teacher Evaluation on Workforce Quality*, Washington, D.C.: Center for Analysis of Longitudinal Data in Education Research, American Institutes for Research, Working Paper 168, 2016.

Curtis, Rachel, and Ross Wiener, *Means to an End: A Guide to Developing Teacher Evaluation Systems That Support Growth and Development*, Washington, D.C.: Aspen Institute, January 5, 2012. As of February 6, 2018: https://www.aspeninstitute.org/publications/means-end-guide-developing-teacher-evaluation-systems-support-growth-development/

Danielson, C., *The Framework for Teaching Evaluation Instrument, 2013 Edition: The Newest Rubric Enhancing the Links to the Common Core State Standards, with Clarity of Language for Ease of Use and Scoring*, 2nd ed., Princeton, N.J.: Charlotte Danielson Group, 2013.

Danielson Group, "The Framework," undated. As of November 13, 2017: https://www.danielsongroup.org/framework/

Danielson, Lana, "Developing and Retaining Quality Classroom Teachers Through Mentoring," *Clearing House*, Vol. 75, No. 4, March–April 2002, pp. 183–185.

David, Jane L., "What Research Says About . . . Teacher Recruitment Incentives," *Educational Leadership*, Vol. 65, No. 7, April 2008, pp. 84–86.

Decker, Geoff, "State: New York City's Top-Rated Teachers Less Likely to Serve Black and Hispanic Students," *Chalkbeat*, April 17, 2015. As of September 20, 2017:
http://www.chalkbeat.org/posts/ny/2015/04/17/state-releases-new-evaluation-data-showing-uneven-spread-of-top-rated-teachers/

Dee, Thomas S., "Teachers, Race, and Student Achievement in a Randomized Experiment," *Review of Economics and Statistics*, Vol. 86, No. 1, February 2004, pp. 195–210.

Dee, Thomas S., and Benjamin J. Keys, "Dollars and Sense: What a Tennessee Experiment Tells Us About Merit Pay," *Education Next*, Vol. 5, No. 1, Winter 2005, pp. 61–67.

Dee, Thomas, and James Wyckoff, *Incentives, Selection, and Teacher Performance: Evidence from IMPACT*, Cambridge, Mass.: National Bureau of Economic Research, Working Paper 19529, October 2013. As of February 6, 2018: http://www.nber.org/papers/w19529

———, "Incentives, Selection, and Teacher Performance: Evidence from IMPACT," *Journal of Policy Analysis and Management*, Vol. 34, No. 2, Spring 2015, pp. 267–297.

Desimone, Laura M., "Improving Impact Studies of Teachers' Professional Development: Toward Better Conceptualizations and Measures," *Educational Researcher*, Vol. 38, No. 3, 2009, pp. 181–199.

District of Columbia Public Schools, "IMPACT: An Overview," undated. As of November 13, 2017:
https://dcps.dc.gov/page/impact-overview

Doherty, Kathryn M., and Sandi Jacobs, *State of the States 2015: Evaluating Teaching, Leading and Learning*, Washington, D.C.: National Council on Teacher Quality, November 2015. As of September 20, 2017:
http://www.nctq.org/dmsView/StateofStates2015

Donaldson, Morgaen, and Casey D. Cobb, "Implementing Student Learning Objectives and Classroom Observations in Connecticut's Teacher Evaluation System," in Jason A. Grissom and Peter Youngs, eds., *Improving Teacher Evaluation Systems: Making the Most of Multiple Measures*, New York: Teachers College Press, 2015, pp. 131–142.

Dowling, Jane, Sheila Murphy, and Baofeng Wang, *The Effects of the Career Ladder Program on Student Achievement*, Phoenix, Ariz.: Sheila Murphy Associates, evaluation report, 2007.

Doyle, Daniela, "Leadership and Lattices: New Pathways Across the Teaching Profession," Washington, D.C.: Center on Great Teachers and Leaders, American Institutes for Research, February 2015. As of February 6, 2018:
https://gtlcenter.org/products-resources/
leadership-and-lattices-new-pathways-across-teaching-profession

Duan, Naihua, "Smearing Estimate: A Nonparametric Retransformation Method," *Journal of the American Statistical Association*, Vol. 78, No. 383, September 1983, pp. 605–610.

Dynarski, Mark, *Teacher Observations Have Been a Waste of Time and Money*, Washington, D.C.: Brookings Institution, December 8, 2016. As of September 20, 2017:
https://www.brookings.edu/research/
teacher-observations-have-been-a-waste-of-time-and-money/

Ebmeier, Howard, and Ann Weaver Hart, "The Effects of a Career-Ladder Program on School Organizational Process," *Educational Evaluation and Policy Analysis*, Vol. 14, No. 3, Autumn 1992, pp. 261–281.

Egalite, Anna J., and Brian Kisida, "The Effects of Teacher Match on Students' Academic Perceptions and Attitudes," *Educational Evaluation and Policy Analysis*, Vol. 40, No. 1, 2018, pp. 1–23.

Eger, Andrea, "TPS: New Teacher-Evaluation System Cost Estimates Misrepresented," *Tulsa World*, December 14, 2011. As of February 4, 2018: http://www.tulsaworld.com/news/education/tps-new-teacher-evaluation-system-cost-estimates-misrepresented/article_b6a88708-44d7-5de9-9c43-2ea4b71d1fd3.html

Ehrenberg, Ronald G., and Dominic J. Brewer, "Did Teachers' Verbal Ability and Race Matter in the 1960s? *Coleman* Revisited," *Economics of Education Review*, Vol. 14, No. 1, March 1995, pp. 1–21.

Elfers, Ana M., Margaret L. Plecki, and Michael S. Knapp, "Teacher Mobility: Looking More Closely at 'the Movers' Within a State System," *Peabody Journal of Education*, Vol. 81, No. 3, 2006, pp. 94–127.

EngageNY, "Pedagogical Shifts Demanded by the Common Core State Standards," undated. As of April 1, 2018: https://www.engageny.org/file/4981/download/common-core-shifts.pdf

Finn, Chester E., Jr., "The End of Teacher Tenure?" *Defining Ideas*, Stanford, Calif.: Hoover Institution, May 23, 2017. As of February 6, 2018: https://www.hoover.org/research/end-teacher-tenure

Florida Senate, Education Personnel, Senate Bill 736, passed March 24, 2011. As of February 9, 2018: https://www.flsenate.gov/Session/Bill/2011/0736

Florida Statutes, Title XLVIII, K–20 Education Code, Chapter 1012, Personnel, Section 1012.01, Definitions, Paragraph 2, Instructional Personnel. As of February 9, 2018: http://www.leg.state.fl.us/statutes/index.cfm?App_mode=Display_Statute&URL=1000-1099/1012/Sections/1012.01.html

Fryer, Roland G., "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools," *Journal of Labor Economics*, Vol. 31, No. 2, April 2013, pp. 373–407.

Garet, Michael S., Andrew C. Porter, Laura Desimone, Beatrice F. Birman, and Kwang Suk Yoon, "What Makes Professional Development Effective? Results from a National Sample of Teachers," *American Educational Research Journal*, Vol. 38, No. 4, 2001, pp. 915–945.

Garet, Michael S., Andrew J. Wayne, Seth Brown, Jordan Rickles, Mengli Song, and David Manzeske, *The Impact of Providing Performance Feedback to Teachers and Principals*, Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, NCEE 2018-4001, December 2017. As of February 6, 2018: https://ies.ed.gov/ncee/pubs/20184001/

Gates, Bill, "Shame Is Not the Solution," *New York Times*, opinion, February 22, 2012. As of February 9, 2018:
http://www.nytimes.com/2012/02/23/opinion/
for-teachers-shame-is-no-solution.html

Gates, Susan M., Laura S. Hamilton, Paco Martorell, Susan Burkhauser, Paul Heaton, Ashley Pierson, Matthew D. Baird, Mirka Vuollo, Jennifer J. Li, Diana Catherine 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

Gilson, Timothy, "Educational Leadership: Are We Busy Yet?" *American Secondary Education*, Vol. 36, No. 3, Summer 2008, pp. 84–97.

Glazerman, Steven, Susanna Loeb, Dan Goldhaber, Douglas Staiger, Stephen Raudenbush, and Grover Whitehurst, *Evaluating Teachers: The Important Role of Value-Added*, Washington, D.C.: Brookings Institution, November 17, 2010. As of February 6, 2018:
https://www.brookings.edu/research/
evaluating-teachers-the-important-role-of-value-added/

Goe, Laura, Kietha Biggers, and Andrew Croft, *Linking Teacher Evaluation to Professional Development: Focusing on Improving Teaching and Learning*, Washington, D.C.: National Comprehensive Center for Teacher Quality, research and policy brief, May 2012. As of February 6, 2018:
https://www.air.org/resource/linking-teacher-evaluation-professional-development-focusing-improving-teaching-and

Goldhaber, Dan, and Michael Hansen, "Using Performance on the Job to Inform Teacher Tenure Decisions," *American Economic Review*, Vol. 100, No. 2, May 2010, pp. 250–255.

Goldhaber, Dan, Michael Hansen, and Joe Walch, *Time to Tenure: Does Tenure Reform Affect Teacher Absence Behavior and Mobility?* Washington, D.C.: National Center for Analysis of Longitudinal Data in Education Research, American Institutes for Research, Working Paper 172, December 2016. As of February 6, 2018:
https://eric.ed.gov/?id=ED573188

Goldhaber, Dan, Vanessa Quince, and Roddy Theobald, *Reconciling Different Estimates of Teacher Quality Based on Value Added*, Washington, D.C.: National Center for Analysis of Longitudinal Data in Education, American Institutes for Research, Brief 14, December 2016. As of January 19, 2017:
http://www.caldercenter.org/sites/default/files/Full%20Text_0.pdf

Goldhaber, Dan, and Joe Walch, "Teacher Tenure: Fog Warning," *Phi Delta Kappan*, Vol. 97, No. 6, 2016, pp. 8–15.

Goldrick, Liam, *Improving Teacher Evaluation to Improve Teacher Quality*, Washington, D.C.: National Governors' Association, issue brief, December 2002. As of March 28, 2018:
https://eric.ed.gov/?id=ED480159

Green Dot Public Schools, "About Us," undated (a). As of February 13, 2018:
http://greendot.org/about/

———, "Careers," undated (b). As of November 13, 2017:
http://careers.greendot.org/

———, "Green Dot Public Schools Education Team Sponsored Professional Development," undated (c).

Grissom, Jason A., and Susanna Loeb, "Assessing Principals' Assessments: Subjective Evaluations of Teacher Effectiveness in Low- and High-Stakes Environments," *Education Finance and Policy*, Vol. 12, No. 3, Summer 2017, pp. 369–395.

Grissom, Jason A., Susanna Loeb, and Benjamin Master, "Effective Instructional Time Use for School Leaders: Longitudinal Evidence from Observations of Principals," *Educational Researcher*, Vol. 42, No. 8, 2013, pp. 433–444.

Guarino, Cassandra, Mark Reckase, Brian Stacy, and Jeffrey Wooldridge, "A Comparison of Student Growth Percentile and Value-Added Models of Teacher Performance," *Statistics and Public Policy*, Vol. 2, No. 1, 2015, pp. 1–11.

Guarino, Cassandra M., Lucrecia Santibañez, and Glenn A. Daley, "Teacher Recruitment and Retention: A Review of the Recent Empirical Literature," *Review of Educational Research*, Vol. 76, No. 2, 2006, pp. 173–208.

Gurin, Patricia, Biren A. Nagda, and Gretchen E. Lopez, "The Benefits of Diversity in Education for Democratic Citizenship," *Journal of Social Issues*, Vol. 60, No. 1, 2004, pp. 17–34.

Guthrie, James W., and Cynthia D. Prince, *Paying for and Sustaining a Performance-Based Compensation System: Guide to Implementation—Resources for Applied Practice*, Washington, D.C.: Center for Educator Compensation Reform, U.S. Department of Education, Office of Elementary and Secondary Education, 2008.

Gutierrez, Italo A., Gabriel Weinberger, and John Engberg, *Improving Teaching Effectiveness: Impact on Student Outcomes—The Intensive Partnerships for Effective Teaching Through 2013–2014*, Santa Monica, Calif.: RAND Corporation, RR-1295/3-1-BMGF, 2016. As of December 13, 2016:
http://www.rand.org/pubs/research_reports/RR1295z3-1.html

Hannaway, Jane, "Motivate Teachers with Incentives," February 5, 2006. As of February 6, 2018:
https://www.urban.org/research/publication/motivate-teachers-incentives

Hanushek, Eric A., "The Trade-Off Between Child Quantity and Quality," *Journal of Political Economy*, Vol. 100, No. 1, February 1992, pp. 84–117.

———, "Teacher Deselection," in Daniel Goldhaber and Jane Hannaway, eds., *Creating a New Teaching Profession*, Washington, D.C.: Urban Institute Press, 2009, pp. 165–180.

———, "School Human Capital and Teacher Salary Policies," *Journal of Professional Capital and Community*, Vol. 1, No. 1, 2016, pp. 23–40.

Harris, Douglas N., and Tim R. Sass, "Teacher Training, Teacher Quality and Student Achievement," *Journal of Public Economics*, Vol. 95, No. 7–8, August 2011, pp. 798–812.

Hock, Heinrich, and Eric Isenberg, *Methods for Accounting for Co-Teaching in Value-Added Models*, Princeton, N.J.: Mathematica Policy Research, working paper, June 2012. As of March 28, 2018:
https://eric.ed.gov/?id=ED533144

Holley, Marc, *A Teacher Quality Primer for Michigan School Officials, State Policymakers, Media and Residents*, 1st ed., Midland, Mich.: Mackinac Center for Public Policy, June 30, 2008. As of February 6, 2018:
https://www.mackinac.org/S2008-05

Hoyer, Kathleen Mulvaney, *Making the Hidden Visible: A Cost Analysis of Teacher Incentive Fund (TIF)–Supported Educator Incentive Projects*, College Park, Md.: Digital Repository of the University of Maryland, University of Maryland, 2015. As of September 20, 2017:
http://hdl.handle.net/1903/16606

Ingersoll, Richard, "The Problem of Underqualified Teachers in American Secondary Schools," *Educational Researchers*, Vol. 28, No. 2, March 1999, pp. 26–37.

———, with Lisa Merrill, and Daniel Stuckey, *Seven Trends: The Transformation of the Teaching Force*, Philadelphia, Pa.: Consortium for Policy Research in Education, University of Pennsylvania, CPRE Research Report RR-80, April 2014. As of February 6, 2018:
http://www.cpre.org/7trends

Ingersoll, Richard, and Lisa Merrill, *A Quarter Century of Changes in the Elementary and Secondary Teaching Force: From 1987 to 2012*, Washington, D.C.: U.S. Department of Education, National Center for Education Statistics, Statistical Analysis Report, NCES 2017-092, April 2017. As of February 6, 2018:
https://eric.ed.gov/?id=ED573526

Isenberg, Eric, Jeffrey Max, Philip Gleason, Matthew Johnson, Jonah Deutsch, and Michael Hansen, *Do Low-Income Students Have Equal Access to Effective Teachers? Evidence from 26 Districts*, Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, NCEE 2017-4008, October 2016. As of January 19, 2017:
https://ies.ed.gov/ncee/pubs/20174008/pdf/20174008.pdf

Jacob, Brian A., *The Effect of Employment Protection on Worker Effort: Evidence from Public Schooling*, Cambridge, Mass.: National Bureau of Economic Research, Working Paper 15655, February 2010. As of February 6, 2018:
http://www.nber.org/papers/w15655

———, "Do Principals Fire the Worst Teachers?" *Educational Evaluation and Policy Analysis*, Vol. 33, No. 4, 2011, pp. 403–434.

Jacobs, Sandi, "Improve Tenure with Better Measures of Teacher Effectiveness," *Phi Delta Kappan*, Vol. 97, No. 6, 2016, pp. 33–37.

James Irvine Foundation, "Leadership Awards: Don Shalvey," undated (award date 2006). As of November 15, 2017:
https://www.irvine.org/leadership-awards/recipients/don-shalvey

Johnson, Susan Moore, Jill Harrison Berg, and Morgaen L. Donaldson, *Who Stays in Teaching and Why: A Review of the Literature on Teacher Retention*, Cambridge, Mass.: Harvard Graduate School of Education, Project on the Next Generation of Teachers, February 2005. As of February 6, 2018:
https://projectngt.gse.harvard.edu/publications/
who-stays-teaching-and-why-review-literature-teacher-retention

Johnson, Susan Moore, Susan M. Kardos, David Kauffman, Edward Liu, and Morgaen L. Donaldson, "The Support Gap: New Teachers' Early Experiences in High-Income and Low-Income Schools," *Education Policy Analysis Archives*, Vol. 12, No. 61, October 29, 2004. As of February 6, 2018:
https://epaa.asu.edu/ojs/article/view/216/342

Johnson, Susan Moore, Matthew A. Kraft, and John Papay, "How Context Matters in High-Need Schools: The Effects of Teachers' Working Conditions on Their Professional Satisfaction and Their Students' Achievement," *Teachers College Record*, Vol. 114, No. 10, 2012, pp. 1–39.

Jones, Leticia M., "First STEPs on Rethinking Student Teacher Placement and Training," *Getting Smart*, March 3, 2017. As of September 20, 2017:
http://www.gettingsmart.com/2017/03/
first-steps-on-rethinking-student-teacher-placements-and-training

Jones, Michael D., "How Do Teachers Respond to Tenure?" *IZA Journal of Labor Economics*, Vol. 4, No. 8, December 2015.

Kane, Thomas J., and Douglas O. Staiger, *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains*, Seattle, Wash.: Bill & Melinda Gates Foundation, Measures of Effective Teaching project, research paper, January 2012. As of February 5, 2018:
https://eric.ed.gov/?id=ED540960

Kane, Thomas J., Eric S. Taylor, John H. Tyler, and Amy L. Wooten, *Identifying Effective Classroom Practices Using Student Achievement Data*, Cambridge, Mass.: National Bureau of Economic Research, Working Paper 15803, March 2010. As of February 6, 2018:
http://www.nber.org/papers/w15803

Kennedy, Mary M., "How Does Professional Development Improve Teaching?" *Review of Educational Research*, Vol. 86, No. 4, 2016, pp. 945–980.

Kini, Tara, and Anne Podolsky, *Does Teaching Experience Increase Teacher Effectiveness? A Review of the Research*, Palo Alto, Calif.: Learning Policy Institute, June 3, 2016. As of February 6, 2018:
https://learningpolicyinstitute.org/product/
does-teaching-experience-increase-teacher-effectiveness-review-research

Koedel, Cory, Jiaxi Li, Matthew G. Springer, and Li Tan, *Do Evaluation Ratings Affect Teachers' Professional Development Activities?* September 2015. As of April 19, 2018:
https://my.vanderbilt.edu/matthewspringer/files/2013/02/
Koedel-et-al-Evaluation-and-PD.pdf

Koedel, Cory, Kata Mihaly, and Jonah E. Rockoff, "Value-Added Modeling: A Review," *Economics of Education Review*, Vol. 47, August 2015, pp. 180–195.

Kraft, Matthew A., David Blazar, and Dylan Hogan, "The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence," *Review of Educational Research*, February 22, 2018.

Kraft, Matthew A., and Allison F. Gilmour, "Can Principals Promote Teacher Development as Evaluators? A Case Study of Principals' Views and Experiences," *Educational Administration Quarterly*, Vol. 52, No. 5, 2016, pp. 711–753.

———, "Revisiting *The Widget Effect:* Teacher Evaluation Reforms and the Distribution of Teacher Effectiveness," *Educational Researcher*, Vol. 46, No. 5, 2017, pp. 234–249.

Ladd, Helen F., *Value-Added Modeling of Teacher Credentials: Policy Implications*, paper presented at the second annual research conference of the National Center for Analysis of Longitudinal Data in Education Research, "The Ins and Outs of Value-Added Measures in Education: What Research Says," Washington, D.C., November 21, 2008.

———, "Teachers' Perceptions of Their Working Conditions: How Predictive of Planned and Actual Teacher Movement?" *Educational Evaluation and Policy Analysis*, Vol. 33, No. 2, June 2011, pp. 235–261.

Lankford, Hamilton, Susanna Loeb, and James Wyckoff, "Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis," *Educational Evaluation and Policy Analysis*, Vol. 24, No. 1, 2002, pp. 37–62.

Larkin, Douglas, and Joseph O. Oluwole, *The Opportunity Costs of Teacher Evaluation: A Labor and Equity Analysis of the TEACHNJ Legislation*, New Jersey Educational Policy Forum, March 1, 2014. As of September 20, 2017: https://njedpolicy.files.wordpress.com/2014/03/douglarkinjosepholuwole-opportunitycostpolicybrief.pdf

Levin, Jessica, and Meredith Quinn, *Missed Opportunities: How We Keep High-Quality Teachers Out of Urban Classrooms*, Brooklyn, N.Y.: New Teacher Project, November 7, 2003. As of February 6, 2018: https://tntp.org/publications/view/missed-opportunities-how-we-keep-quality-teachers-out-of-urban-classrooms

Lockwood, J. R., and Daniel F. McCaffrey, "Correcting for Test Score Measurement Error in ANCOVA Models for Estimating Treatment Effects," *Journal of Educational and Behavioral Statistics*, Vol. 39, No. 1, 2014, pp. 22–52.

Lockwood, J. R., Daniel F. McCaffrey, Laura S. Hamilton, Brian Stecher, Vi-Nhuan Le, and José Felipe Martinez, "The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures," *Journal of Educational Measurement*, Vol. 44, No. 1, March 2007, pp. 47–67.

Loeb, Susanna, Demetra Kalogrides, and Tara Beteille, "Effective Schools: Teacher Hiring, Assignment, Development, and Retention," *Education Finance and Policy*, Vol. 7, No. 3, Summer 2012, pp. 268–304.

Loeb, Susanna, Luke C. Miller, and James Wyckoff, "Performance Screens for School Improvement: The Case of Tenure Reform in New York City," *Educational Researcher*, Vol. 44, No. 4, 2015, pp. 199–212.

Loeb, Susanna, James Soland, and Lindsay Fox, "Is a Good Teacher a Good Teacher for All? Comparing Value-Added of Teachers with Their English Learners and Non–English Learners," *Education Evaluation and Policy Analysis*, Vol. 36, No. 4, 2014, pp. 457–475.

Marsh, Julie A., Susan Bush-Mecenas, Katharine O. Strunk, Jane Arnold Lincove, and Alice Huguet, "Evaluating Teachers in the Big Easy: How Organizational Context Shapes Policy Responses in New Orleans," *Educational Evaluation and Policy Analysis*, Vol. 39, No. 4, 2017, pp. 539–570.

Martin, Jessica, *Best Practices in Minority Teacher Recruitment: A Literature Review*, Connecticut Regional Educational Service Centers Alliance, Minority Teacher Recruiting Initiative, November 2011. As of February 6, 2018: http://www.crecnow.info/RESCMTR/docs/national_best_practices_in_minority_teacher_recruitment.pdf

McCaffrey, Daniel F., J. R. Lockwood, Daniel Koretz, Thomas A. Louis, and Laura Hamilton, "Models for Value-Added Modeling of Teacher Effects," *Journal of Educational and Behavioral Statistics*, Vol. 29, No. 1, 2004, pp. 67–101.

McCann, Colleen, and Sasha Zuflacht, with Tierra Gilbert, "The Decade-Plus Teaching Career: How to Retain Effective Teachers Through Teacher Leadership," *Teach Plus*, February 6, 2015. As of February 6, 2018:
http://teachplus.org/news-events/publications/
decade-plus-teaching-career-how-retain-effective-teachers-through-teacher

McGuinn, Patrick, *Ringing the Bell for K–12 Teacher Tenure Reform*, Washington, D.C.: Center for American Progress, February 9, 2010. As of February 6, 2018:
https://www.americanprogress.org/issues/education-k-12/reports/2010/02/09/7314/
ringing-the-bell-for-k-12-teacher-tenure-reform/

Miller, Jackson, Matthew Finster, Richard Adrien, Bob Harmon, Julia Koppich, Amy Potemski, and Maureen Yoder, *The Evolution of the Teacher Incentive Fund Program*, Washington, D.C.: U.S. Department of Education, 2015. As of February 6, 2018:
https://www.tlpcommunity.org/resources/
evolution-teacher-incentive-fund-program

Miller, Raegen, and Robin Chait, *Teacher Turnover, Tenure Policies, and the Distribution of Teacher Quality: Can High-Poverty Schools Catch a Break?* Washington, D.C.: Center for American Progress, December 2, 2008. As of February 6, 2018:
https://www.americanprogress.org/issues/education-k-12/reports/2008/12/02/5344/
teacher-turnover-tenure-policies-and-the-distribution-of-teacher-quality/

Monk, David H., "Subject Area Preparation of Secondary Mathematics and Science Teachers and Student Achievement," *Economics of Education Review*, Vol. 13, No. 2, June 1994, pp. 125–145.

Mooney, John, "School Districts Comparison Shop for Teacher Evaluation Systems," *NJSpotlight*, October 17, 2012. As of December 19, 2017:
http://www.njspotlight.com/stories/12/10/17/
school-districts-comparison-shop-for-teacher-evaluation-systems/

Murnane, Richard J., and Jennifer L. Steele, "What Is the Problem? The Challenge of Providing Effective Teachers for All Children," *Future of Children*, Vol. 17, No. 1, Spring 2007, pp. 15–43.

Natale, Catherine Fisk, Lynn Gaddis, Katherine Bassett, and Katherine McKnight, *Teacher Career Advancement Initiative: Lessons Learned from Eight Case Studies*, Pearson Education and National Network of State Teachers of the Year, January 2016. As of February 6, 2018:
http://www.nnstoy.org/publications/
teacher-career-advancement-initiatives-lessons-learned-from-eight-case-studies/

National Association of Secondary School Principals and National Association of Elementary School Principals, *Supporting Principals in Implementing Teacher Evaluation Systems*, Alexandria, Va., undated. As of December 19, 2017:
http://www.naesp.org/sites/default/files/TeacherEvalBrief_Final.pdf

National Board for Professional Teaching Standards, *2015 State Rankings by Total Number of National Board Certified Teachers*, 2015.

National Center for Education Statistics, "2011–2012 SASS Tables," undated (a). As of February 16, 2018:
https://nces.ed.gov/surveys/sass/tables_list.asp#2012

———, "Common Core of Data: America's Public Schools," undated (b). As of February 9, 2018:
https://nces.ed.gov/ccd/

National Center for Teacher Residencies, "About: The Residency Model," undated. As of September 20, 2017:
https://nctresidencies.org/about/residency-model-teacher-mentor-programs/

NCES—*See* National Center for Education Statistics.

Network for Public Education, *Teachers Talk Back: Educators on the Impact of Teacher Evaluation*, 2016. As of September 20, 2017:
https://networkforpubliceducation.org/wp-content/uploads/2016/04/NPETeacherEvalReport.pdf

New Teacher Project, The, *The Irreplaceables: Understanding the Real Retention Crisis in America's Urban Schools*, Brooklyn, N.Y., c. 2012. As of September 20, 2017:
https://tntp.org/assets/documents/TNTP_Irreplaceables_2012.pdf

———, *Professional Development and Support: End-of-Year Stepback—Shelby County Schools*, July 2014.

Nuttall, Rebecca, "Grade Scale: PPS, Union at Odds over Teacher-Evaluation Standards," *Pittsburgh City Paper*, October 30, 2013. As of November 13, 2017:
https://www.pghcitypaper.com/pittsburgh/grade-scale-pps-union-at-odds-over-teacher-evaluation-standards/Content?oid=1705181

Odden, Allan, and Carolyn Kelley, *Paying Teachers for What They Know and Do: New and Smarter Compensation Strategies to Improve Schools*, 2nd ed., Thousand Oaks, Calif.: Corwin Press, 2002.

Odden, Allan, and Marc Wallace, *Rewarding Teacher Excellence: A Teacher Compensation Handbook for State and Local Policy Makers*, Madison, Wis.: Consortium for Policy Research in Education, Wisconsin Center for Education Research, University of Wisconsin, February 2007. As of February 6, 2018:
http://picusodden.com/wp-content/uploads/2013/09/tcomp-handbook-feb-28-07-final-3-05-07.pdf

Organisation for Economic Co-operation and Development, "How Much Time Do Teachers Spend Teaching?" *Education at a Glance 2014: OECD Indicators*, Indicator D4, Paris: OECD Publishing, 2014. As of February 6, 2018: http://www.oecd-ilibrary.org/education/education-at-a-glance-2014_eag-2014-en

Papay, John P., "Refocusing the Debate: Assessing the Purposes and Tools of Teacher Evaluation," *Harvard Educational Review*, Vol. 82, No. 1, Spring 2012, pp. 123–141.

Papay, John P., Eric S. Taylor, John H. Tyler, and Mary Laski, *Learning Job Skills from Colleagues at Work: Evidence from a Field Experiment Using Teacher Performance Data*, Cambridge, Mass.: National Bureau of Economic Research, Working Paper 21986, February 2016. As of February 6, 2018: http://www.nber.org/papers/w21986

Partnership of California Charter Management Organizations, *Proposal for the Intensive Partnership to Empower Effective Teachers*, submission to the Bill & Melinda Gates Foundation, July 31, 2009.

Partnerships to Uplift Communities, "Design Principles," undated (a). As of November 13, 2017: http://www.pucschools.org/about/design-principles/

———, "Schools," undated (b). As of November 13, 2017: http://www.pucschools.org/schools

Pazzanese, Christina, "Targeting Teacher Tenure," *Harvard Gazette*, August 18, 2014. As of February 6, 2018: http://news.harvard.edu/gazette/story/2014/08/targeting-teacher-tenure/

Pennsylvania General Assembly, Public School Code of 1949: Omnibus Amendments, Act of Pennsylvania Law 684-82, June 30, 2012. As of February 9, 2018: http://www.legis.state.pa.us/cfdocs/legis/li/uconsCheck.cfm?yr=2012&sessInd=0&act=82

Pham, Lam D., Tuan D. Nguyen, and Matthew G. Springer, *Teacher Merit Pay and Student Test Scores: A Meta-Analysis*, Nashville, Tenn.: Vanderbilt University, April 3, 2017. As of February 6, 2018: https://s3.amazonaws.com/vu-my/wp-content/uploads/sites/868/2013/02/05145950/Pham-Nguyen-Springer-2017.pdf

Pitts, David W., "Representative Bureaucracy, Ethnicity, and Public Schools: Examining the Link Between Representation and Performance," *Administration and Society*, Vol. 39, No. 4, 2007, pp. 497–526.

Pittsburgh Public Schools, *Career Ladder FAQs*, Pittsburgh, Pa., 2012.

Podolsky, Anne, Tara Kini, Joseph Bishop, and Linda Darling-Hammond, *Solving the Teacher Shortage: How to Attract and Retain Excellent Educators*, Palo Alto, Calif.: Learning Policy Institute, September 15, 2016. As of February 7, 2018: https://learningpolicyinstitute.org/product/solving-teacher-shortage

PPS—*See* Pittsburgh Public Schools.

Public Law 107-110, No Child Left Behind Act of 2001, January 8, 2002. As of February 20, 2017: https://www.gpo.gov/fdsys/pkg/PLAW-107publ110/content-detail.html

Public Schools of North Carolina, State Board of Education, Department of Public Instruction, *Report to the North Carolina General Assembly: 2014–2015 State of the Teaching Profession in North Carolina*, Report 70, November 15, 2015. As of September 20, 2017: http://www.dpi.state.nc.us/docs/educatoreffectiveness/surveys/leaving/2014-15turnoverreport.pdf

PUC—*See* Partnerships to Uplift Communities.

Raudenbush, Stephen W., and Anthony S. Bryk, *Hierarchical Linear Models: Applications and Data Analysis Methods*, Thousand Oaks, Calif.: Sage Publications, 2002.

Reardon, Sean F., and Stephen W. Raudenbush, "Assumptions of Value-Added Models for Estimating School Effects," *Education Finance and Policy*, Vol. 4, No. 4, Fall 2009, pp. 492–519.

Reform Support Network, "Targeting Growth: Using Student Learning Objectives as a Measure of Educator Effectiveness," undated. As of November 13, 2017: https://www2.ed.gov/about/inits/ed/implementation-support-unit/tech-assist/targeting-growth.pdf

Reinhorn, Stefanie K., Susan Moore Johnson, and Nicole S. Simon, "Investing in Development: Six High-Performing, High-Poverty Schools Implement the Massachusetts Teacher Evaluation Policy," *Educational Evaluation and Policy Analysis*, Vol. 39, No. 3, 2017, pp. 383–406.

Rice, Jennifer King, *The Impact of Teacher Experience: Examining the Evidence and Policy Implications*, Washington, D.C.: Urban Institute, Brief 11, August 2010. As of February 7, 2018: https://www.urban.org/research/publication/impact-teacher-experience

Rockoff, Jonah E., Douglas O. Staiger, Thomas J. Kane, and Eric S. Taylor, "Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools," *American Economic Review*, Vol. 102, No. 7, December 2012, pp. 3184–3213.

Ronfeldt, Matthew, Susanna Loeb, and James Wyckoff, "How Teacher Turnover Harms Student Achievement," *American Educational Research Journal*, Vol. 50, No. 1, 2013, pp. 4–36.

Rothstein, Jesse, "Teacher Quality Policy When Supply Matters," *American Economic Review*, Vol. 105, No. 1, January 2015, pp. 100–130.

Rowan, Brian, and Stephen Raudenbush, "Teacher Evaluation in American Schools," in Drew H. Gitomer and Courtney A. Bell, eds., *Handbook of Research on Teaching*, 5th ed., Washington, D.C.: American Educational Research Association, 2016, pp. 1159–1216.

Rowan, Brian, Stephen G. Schilling, Angeline Spain, Prem Bhandari, Daniel Berger, and John Graves, *Promoting High Quality Teacher Evaluations in Michigan: Lessons from a Pilot of Educator Effectiveness Tools*, Ann Arbor, Mich.: University of Michigan Institute for Social Research, December 2013. As of February 7, 2018:
http://www.mcede.org/system/resources/BAhbBlsHOgZmSSJXMjAxNC8wMi8wNC8yMC80NS8wNy80NzcvUHJvbW90aW5nX0hpZ2hfUXVhbGl0eV9UZWFjaGVyX0V2YWx1YXRpb25zX2luX01pY2hpZ2FuLnBkZgY6BkVU/Promoting-High-Quality-Teacher%20Evaluations-in-Michigan.pdf

Rubin, Joel, and Howard Blume, "Green Dot Charter Organization to Take over Locke High School," *Los Angeles Times*, September 12, 2007. As of February 25, 2018:
http://articles.latimes.com/2007/sep/12/local/me-lausd12

Sargrad, Scott, Samantha Batel, Melissa Lazarin, and Catherine Brown, *Investing in Educator Capacity: An Analysis of State Race to the Top Spending*, Washington, D.C.: Center for American Progress, December 2015. As of February 7, 2018:
https://www.americanprogress.org/issues/education-k-12/reports/2015/12/21/127748/investing-in-educator-capacity/

Sartain, Lauren, and Matthew P. Steinberg, "Teachers' Labor Market Responses to Performance Evaluation Reform: Experimental Evidence from Chicago Public Schools," *Journal of Human Resources*, Vol. 51, No. 3, August 1, 2016, pp. 615–655.

Sawchuk, Stephen, "What Will New Evaluation Systems Cost?" *Teacher Beat*, April 1, 2013. As of February 4, 2018:
http://blogs.edweek.org/edweek/teacherbeat/2013/04/What_will_new_evaluation.html

Schochet, Peter Z., and Hanley S. Chiang, *Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains*, Washington, D.C.: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, NCEE 2010-4004, July 2010. As of March 28, 2018:
https://ies.ed.gov/ncee/pubs/20104004/

SCS—*See* Shelby County Schools.

Sebastian, James, Erin M. Camburn, and James P. Spillane, "Portraits of Principal Practice: Time Allocation and School Principal Work," *Educational Administration Quarterly*, Vol. 54, No. 1, 2018.

Shakman, Karen, Jacqueline Zweig, Candice Bocala, Natalie Lacireno-Paquet, and Jessica Bailey, *Teacher Evaluation and Professional Learning: Lessons from Early Implementation in a Large Urban District*, Washington, D.C.: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands, REL 2016-115, March 2016.

Shelby County Schools, "iZone," undated. As of April 2, 2018:
http://www.scsk12.org/izone/

Smith, P. A., "The Relationship Between Principal Time Allocation and Work Effectiveness: Evidence Across City, Suburban, Town and Rural School Districts," *National Teacher Education Journal*, Vol. 6, No. 2, Summer 2013, pp. 27–36.

"Spending: Does California Skimp on Education?" *Ed100*, undated. As of November 14, 2017:
https://ed100.org/lessons/californiaskimps

Spillane, James P., Brian J. Reiser, and Todd Reimer, "Policy Implementation and Cognition: Reframing and Refocusing Implementation Research," *Review of Educational Research*, Vol. 72, No. 3, Autumn 2002, pp. 387–431.

Sporte, Susan E., W. David Stevens, Kaleen Healey, Jennie Jiang, and Holly Hart, *Teacher Evaluation in Practice: Implementing Chicago's REACH Students*, Chicago, Ill.: University of Chicago Consortium on Chicago School Research, research report, September 2013. As of September 20, 2017:
https://consortium.uchicago.edu/sites/default/files/publications/REACH%20Report_0.pdf

Springer, Matthew G., Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J. R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, and Brian M. Stecher, *Final Report: Experimental Evidence from the Project on Incentives in Teaching (POINT)*, Nashville, Tenn.: National Center on Performance Incentives, Project on Incentives in Teaching, September 30, 2012. As of February 7, 2018:
https://my.vanderbilt.edu/performanceincentives/files/2012/09/Full-Final-Report-Experimental-Evidence-from-the-Project-on-Incentives-in-Teaching-2012.pdf

Staiger, Douglas O., Robert Gordon, and Thomas J. Kane, *Identifying Effective Teachers Using Performance on the Job*, Washington, D.C.: Brookings Institution, April 1, 2006. As of February 7, 2018:
https://www.brookings.edu/research/identifying-effective-teachers-using-performance-on-the-job/

Staiger, Douglas O., and Jonah E. Rockoff, "Searching for Effective Teachers with Imperfect Information," *Journal of Economic Perspectives*, Vol. 24, No. 3, Summer 2010, pp. 97–118.

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, *Improving Teaching Effectiveness: Implementation—The Intensive Partnerships for Effective Teaching Through 2013–2014*, Santa Monica, Calif.: RAND Corporation, RR-1295-BMGF, 2016. As of February 7, 2018:
https://www.rand.org/pubs/research_reports/RR1295.html

Stecher, Brian, Mike Garet, Deborah Holtzman, and Laura Hamilton, "Implementing Measures of Teacher Effectiveness," *Phi Delta Kappan*, Vol. 94, No. 3, November 2012, pp. 39–43.

Steinberg, Matthew P., and Morgaen L. Donaldson, "The New Educational Accountability: Understanding the Landscape of Teacher Evaluation in the Post-NCLB Era," *Education Finance and Policy*, Vol. 11, No. 3, Summer 2016, pp. 340–359.

Steinberg, Matthew P., and Lauren Sartain, "Does Teacher Evaluation Improve School Performance? Experimental Evidence from Chicago's Excellence in Teaching Project," *Education Finance and Policy*, Vol. 10, No. 4, Fall 2015, pp. 1–38.

Stockard, Jean, and Michael Bryan Lehman, "Influences on the Satisfaction and Retention of 1st-Year Teachers: The Importance of Effective School Management," *Educational Administration Quarterly*, Vol. 40, No. 5, December 2004, pp. 742–771.

Strunk, Katharine O., Nathan Barrett, and Jane Arnold Lincove, *When Tenure Ends: The Short-Run Effects of the Elimination of Louisiana's Teacher Employment Protections on Teacher Exit and Retirement*, New Orleans, La.: Education Research Alliance for New Orleans, technical report, April 12, 2017. As of February 7, 2018:
https://educationresearchalliancenola.org/files/publications/041217-Strunk-Barrett-Lincove-When-Tenure-Ends.pdf

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, 2014, pp. 281–306.

Strunk, Katharine O., Tracey L. Weinsten, and Reino Makkonen, "Sorting Out the Signal: Do Multiple Measures of Teachers' Effectiveness Provide Consistent Information to Teachers and Principals?" *Education Policy Analysis Archives*, Vol. 22, No. 100, November 10, 2014.

Taylor, Eric S., and John H. Tyler, "The Effect of Evaluation on Teacher Performance," *American Economic Review*, Vol. 102, No. 7, December 2012, pp. 3628–3651.

Thomsen, Jennifer, "Teacher Performance Plays Growing Role in Employment Decisions: Teacher Tenure—Trends in State Laws," Education Commission of the States, policy report, May 22, 2014. As of February 7, 2018:
https://www.ecs.org/
teacher-performance-plays-growing-role-in-employment-decisions/

Thorn, Chris, "Carnegie Cost Calculator: A Tool for Exploring the Cost of Educator Evaluation Systems," *Carnegie Commons Blog*, September 25, 2013. As of December 19, 2017:
https://www.carnegiefoundation.org/blog/carnegie-cost-calculator-a-tool-for-exploring-the-cost-of-educator-evaluation-systems/

TNTP—*See* New Teacher Project.

Toch, Thomas, "Fixing Teacher Evaluation," *Educational Leadership*, Vol. 66, No. 2, October 2008, pp. 32–37. As of January 8, 2018:
http://www.ascd.org/publications/educational-leadership/oct08/vol66/num02/Fixing-Teacher-Evaluation.aspx

Toch, Thomas, and Robert Rothman, *Rush to Judgment: Teacher Evaluation in Public Education*, Washington, D.C.: Education Sector, January 29, 2008.

Tooms, Autumn, and Alicia Crowe, "Hiring Good Teachers: The Interview Process," *Principal*, November–December 2004. As of September 20, 2017:
https://www.naesp.org/sites/default/files/resources/2/Principal/2004/N-Dp50.pdf

Tulare County Office of Education, *2008 Best Practices in the Recruitment of Qualified Teachers: Findings of the California Recruitment Program*, Tulare, Calif.: California Teacher Recruitment Program, 2008. As of February 7, 2018:
http://www.teachincal.org/bestpractices.pdf

Walsh, Kate, Sam Lubell, and Elizabeth Ross, *Backing the Wrong Horse: The Story of One State's Ambitious but Disheartening Foray into Performance Pay*, Washington, D.C.: National Council on Teacher Quality, August 2017. As of February 7, 2018:
https://www.nctq.org/publications/Backing-the-Wrong-Horse:-The-Story-of-One-States-Ambitious-But-Disheartening-Foray-Into-Performance-Pay

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, and David Keeling, *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness*, Brooklyn, N.Y.: New Teacher Project, June 8, 2009. As of February 7, 2018:
https://tntp.org/publications/view/
the-widget-effect-failure-to-act-on-differences-in-teacher-effectiveness

Weiss, Carol H., "Knowledge Creep and Decision Accretion," *Knowledge: Creation, Diffusion, Utilization*, Vol. 1, No. 3, March 1980, pp. 381–404.

Wellington, Alison, Hanley Chiang, Kristin Hallgren, Cecilia Speroni, Mariesa Herrmann, and Paul Burkander, *Evaluation of the Teacher Incentive Fund: Implementation and Impacts of Pay-for-Performance After Three Years*, Jessup, Md.: National Center for Education Evaluation and Regional Assistance, NCEE 2016-4004, August 2016. As of February 7, 2018:
https://eric.ed.gov/?id=ED568150

Whitney, Myra, Tishsha Hopson, Lytania Black, and Charles New, *Teacher Effectiveness Initiative: Teacher Development and Support—Individualized Growth: Resource Book*, Memphis, Tenn.: Memphis City Schools, undated. As of March 28, 2018:
https://www.scribd.com/document/66591232/TEM-Aligned-PD-Resource-Guide

Winters, Marcus A., and Joshua M. Cowen, "Would a Value-Added System of Retention Improve the Distribution of Teacher Quality? A Simulation of Alternative Policies," *Journal of Policy Analysis and Management*, Vol. 32, No. 3, Summer 2013, pp. 634–654.

———, "Who Would Stay, Who Would Be Dismissed? An Empirical Consideration of Value-Added Teacher Retention Policies," *Educational Researcher*, Vol. 42, No. 6, 2014, pp. 330–337.

Yaffe, Deborah, "Find, Keep, Cultivate the Best Teachers: A Guide to Recruiting and Nurturing Successful Teachers," *District Administration*, February 2015, pp. 31–40. As of September 20, 2017:
https://www.districtadministration.com/article/find-keep-cultivate-best-teachers

York-Barr, Jennifer, and Karen Duke, "What Do We Know About Teacher Leadership? Findings from Two Decades of Scholarship," *Review of Educational Research*, Vol. 74, No. 3, Fall 2004, pp. 255–316.

Zamarro, Gema, John Engberg, Juan Esteban Saavedra, and Jennifer Steele, "Disentangling Disadvantage: Can We Distinguish Good Teaching from Classroom Composition?" *Journal of Research on Educational Effectiveness*, Vol. 8, No. 1, 2015, pp. 84–111.

The Intensive Partnerships for Effective Teaching initiative, designed and funded by the Bill & Melinda Gates Foundation, was a multiyear effort to dramatically improve student outcomes by increasing students' access to effective teaching. Participating sites adopted measures of teaching effectiveness (TE) that included both a teacher's contribution to growth in student achievement and his or her teaching practices assessed with a structured observation rubric. The TE measures were to be used to improve staffing actions, identify teaching weaknesses and overcome them through effectiveness-linked professional development, and employ compensation and career ladders as incentives to retain the most-effective teachers and have them support the growth of other teachers. The developers believed that these mechanisms would lead to more-effective teaching, greater access to effective teaching for low-income minority students, and greatly improved academic outcomes.

Beginning in 2009–2010, three school districts—Hillsborough County Public Schools in Florida; Memphis City Schools in Tennessee (which merged with Shelby County Schools during the initiative); and Pittsburgh Public Schools in Pennsylvania—and four charter management organizations—Alliance College-Ready Public Schools, Aspire Public Schools, Green Dot Public Schools, and Partnerships to Uplift Communities Schools—participated in the Intensive Partnerships initiative. RAND and the American Institutes for Research conducted a six-year evaluation of the initiative, documenting the policies and practices each site enacted and their effects on student outcomes. This is the final evaluation report.

$49.95

Download a free copy at
www.rand.org/t/RR2242

ISBN-10 1-9774-0079-5
ISBN-13 978-1-9774-0079-6
54995
9 781977 400796

RAND EDUCATION

RR-2242-BMGF