Solving the Achievement Gap

Overcoming the Structure of School Inequality

STUART S. YEH

Solving the Achievement Gap

Stuart S. Yeh

Solving the Achievement Gap

Overcoming the Structure of School Inequality

palgrave
macmillan

Stuart S. Yeh
University of Minnesota
Minneapolis, Minnesota, USA

ISBN 978-1-137-58765-7 (hardcover) ISBN 978-1-137-58767-1 (eBook)
ISBN 978-1-137-58766-4 (softcover)
DOI 10.1057/978-1-137-58767-1

Library of Congress Control Number: 2016956397

Cover illustration: © PhotoAlto / Alamy Stock Photo

Printed on acid-free paper

This Palgrave Macmillan imprint is published by Springer Nature
The registered company is Nature America Inc.
The registered company address is: 1 New York Plaza, New York, NY 10004, U.S.A.

Acknowledgments

I thank Lyn Corno and Teachers College Record for publishing articles that challenge conventional wisdom. I also thank Steve Glazerman, Dan Goldhaber, Edward Haertel, Cory Koedel, Susanna Loeb, Joseph Ritter, Michael Rodriguez, and Douglas Staiger for comments on earlier drafts of articles that are adapted here as book chapters.

Figure 5.1 is adapted from *Journal of Experimental Child Psychology* vol. 8, no. 1, Smith, Donald E. P., Dale Brethower, and Raymond Cabot. "Increasing Task Behavior in a Language Arts Program by Providing Reinforcement," page 48, copyright 1969, with permission from Elsevier.

Figure 5.2 is adapted from *Journal of Experimental Child Psychology* vol. 8, no. 1, Smith, Donald E. P., Dale Brethower, and Raymond Cabot. "Increasing Task Behavior in a Language Arts Program by Providing Reinforcement," page 58, copyright 1969, with permission from Elsevier.

Figure 5.5 and Figure 10.1 are adapted from *Journal of Education Finance* vol. 36, no. 1, Yeh, Stuart S. "The Cost-Effectiveness of 22 Approaches for Raising Student Achievement," page 63, copyright 2010, with permission from the *Journal of Education Finance*.

Table 5.1 is adapted from *Journal of Education Finance* vol. 36, no. 1, Yeh, Stuart S. "The Cost-Effectiveness of 22 Approaches for Raising Student Achievement," page 43, copyright 2010, with permission from the *Journal of Education Finance*.

Chapters Three, Four and Five are adapted from *Teachers College Record* vol. 117, no. 12, Yeh, Stuart S. "Two Models of Learning and Achievement: An Explanation for the Achievement Gap?" copyright 2015, with permission from *Teachers College Record*.

Chapter Six is adapted from *Teachers College Record,* Yeh, Stuart S. "Contradictions Resolved: An Analysis of Two Theories of the Achievement Gap," in press, with permission from *Teachers College Record*.

Chapter Seven is adapted from *Teachers College Record,* Yeh, Stuart S. "Race, Test Scores and Educational Outcomes," in press, with permission from *Teachers College Record*.

Chapter Eight is adapted from *Teachers College Record* vol. 115, no. 4, Yeh, Stuart S. “A Re-Analysis of the Effects of KIPP and the Harlem Promise Academies,” pages 1–20, copyright 2013, with permission from *Teachers College Record*.

Chapter Nine is adapted from *Teachers College Record* vol. 115, no. 12, Yeh, Stuart S. “A Re-Analysis of the Effects of Teacher Replacement Using Value-Added Modeling,” pages 1–35, copyright 2013, with permission from *Teachers College Record*.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Solving the Achievement Gap offers a novel view regarding the cause of the achievement gap between low-income minority students and their middle-class peers. The book suggests that the prevailing emphasis on socioeconomic factors, sociocultural influences, and teacher quality is misplaced. The cause of the gap may instead be traced to a flaw in the way schools are currently structured. The book suggests that misdiagnosis of the nature of the achievement gap has led to misguided solutions. The book outlines a new theory of the achievement gap and draws upon a range of research studies that support this view.

Solving the Achievement Gap suggests that the cause of the achievement gap is not differences in parenting styles, or the economic advantages of middle-class parents, or differences in the quality of teachers. Instead, schools present learning tasks and award grades in ways that inadvertently undermine the self-efficacy, engagement, and effort of low-performing students, causing demoralization and exacerbating differences in achievement that exist at kindergarten. This process systematically maintains and widens initial gaps in achievement that might otherwise be expected to disappear over the K-12 years. Thus, the book identifies a mechanism that is built into the existing structure of schooling and appears to perpetuate the achievement gap. No previous analysis has identified this mechanism.

A disproportionate number of black and Hispanic children enter the school system performing below their middle-class peers. They are subsequently tested regarding their knowledge of letters, letter–sound relationships, numbers, and ability to decode printed text. The children rapidly discover that they are being compared to each other and some children are being treated as “smarter” than others. These discouraging comparisons are reinforced through comparative grading, testing, and grouping practices throughout the students’ K-12 careers.

White and Asian children who enter kindergarten performing above-average are less affected than black and Hispanic children because white and Asian children tend to receive more “A” grades and high scores on classroom assessments. Except in rare cases, however, even the most proficient students

S.S. Yeh, *Solving the Achievement Gap*,
DOI 10.1057/978-1-137-58767-1_1

receive an occasional bad grade. As a consequence, almost all children become demoralized, but the degree and pace of demoralization are most severe for black and Hispanic children.

The book offers multiple pieces of evidence supporting the proposed model of the achievement gap. First, the proposed model of the gap builds upon voluminous research regarding conditions that elicit "learned helplessness"— conditions that demonstrably undermine learning and achievement and appear to characterize the conventional model of schooling. If schools are characterized by conditions that are known to elicit learned helplessness, this is evidence of a structural flaw that must be corrected in order to address the achievement gap.

Second, the proposed model of the gap builds upon voluminous research regarding the conditions that correct learned helplessness. The conditions that correct learned helplessness involve individualization of task difficulty and rapid performance feedback so that each student can receive high scores on daily math and reading assessments, feel a sense of accomplishment on a daily basis, see a direct relationship between effort and achievement, and enjoy learning. The proposed model of the gap points directly to a specific intervention for correcting learned helplessness, building self-efficacy, and improving engagement and achievement. Significantly, this intervention is more efficient than 22 alternative approaches for raising achievement and reducing the achievement gap. This is evidence supporting the validity of the proposed model of the achievement gap.

Third, the book offers evidence that differences that exist prior to the age when children enter the school system do not adequately explain the precipitous decline in the self-efficacy and engagement of low-achieving children after they enter the school system. Both Asian and white children suffer a similar decline in self-efficacy and engagement, and the self-efficacy and engagement of Asian children decline despite their superior academic achievement, relative to white children. This points to a generic factor, something inherent in the structure of the conventional school system—rather than a factor that is specific to black and Hispanic children—as the cause of the decline. This is consistent with the proposed model of the achievement gap but is not consistent with sociocultural and socioeconomic theories that emphasize differences in parental styles between white and minority parents.

Oppositional peer culture—the notion that black students might encounter peers who are hostile to blacks who adopt white ways of speaking and acting and demonstrate high-achievement—cannot explain the persistence of the achievement gap.[1] The evidence from nationally-representative datasets either does not support the oppositional peer culture explanation[2] or suggests that the phenomenon is limited to the most segregated schools in the nation.[3] Thus, oppositional culture can only explain the achievement gap at a limited number of schools.

Fourth, the book offers evidence that neither school nor teacher quality can explain the persistence of the achievement gap. While black and white students attend schools that differ with regard to the percentage of students eligible for free lunch, the degree of gang problems in school, the amount of loitering in front of the school by nonstudents, and the amount of litter around the schools, these measures explain only a small fraction of the variation in student outcomes.[4] Furthermore, the inclusion of these covariates does not prevent the black-white gap from widening as children progress through the kindergarten and first-grade years.[5] By the end of first grade, black students have lost the equivalent of almost three months of schooling relative to whites.[6] In addition, both Hispanic and Asian students experience worse schools than whites, but neither of those groups loses ground over time.[7] Finally, the black-white gap grows through third grade even between black and white students attending the same school.[8] All of the ground lost by black students between first grade and third grade is within, rather than across, schools.[9] Black students lose substantial ground relative to white students within the same school and even within the same classrooms.[10] Controlling for school-fixed effects or teacher-fixed effects in models of achievement does little to explain the divergent trajectories of black and white students between kindergarten and third grade.[11] This pattern of results is inconsistent with the hypothesis that differences in school or teacher quality explain the black-white achievement gap. The pattern is, instead, consistent with the hypothesis that there is something about the interaction between black students and schools that interferes with the learning process.[12]

Asian and white children, as well as black and Hispanic children, suffer a dual decline in self-efficacy and engagement. This points to a generic factor, something inherent in the structure of the conventional school system, rather than differences in the quality of schools and teachers experienced by black and Hispanic children compared to Asian and white children. Furthermore, almost all children—including high-achieving children—suffer a dual decline in self-efficacy and engagement. Again, this points to a generic factor, something inherent in the structure of the conventional school system, rather than differences in the quality of schools and teachers experienced by low-achieving children. This is consistent with the proposed model of the achievement gap but is not consistent with theories that emphasize differences in school and teacher quality.

Fifth, the book offers evidence that "no-excuses" charter schools are not the answer to the achievement gap. Such schools can only maintain their superior performance by attracting a disproportionate fraction of the most-dedicated teachers in the nation. The book offers an analysis indicating that the superior performance of these schools would disappear if an attempt was made to scale-up these charter school models nationwide. This suggests that improvements in school quality are not the answer to the achievement gap.

Sixth, the book offers evidence that the use of value-added statistical methods to identify and replace low-performing teachers is not the answer to the achievement gap. Three research studies demonstrate that a basic assumption underlying these methods is not valid and, thus, teacher rankings based on these methods are not valid for the purpose of identifying and replacing low-performing teachers. Furthermore, the book offers evidence that teacher rankings based on these methods are not sufficiently reliable—and therefore not valid—for the purpose of identifying and replacing low-performing teachers. This suggests that the strategy of replacing low-performing teachers is not the answer to the achievement gap.

Together, this evidence supports the proposed model of the achievement gap and contradicts existing theories that the gap may be traced to sociocultural and socioeconomic differences that exist prior to the age when children enter the school system or theories that the gap may be traced to differences in the quality of the schools and teachers experienced by minority students.

The proposed model is supported because no other model is consistent with all of these research findings: basic research regarding the conditions that cause learned helplessness, research suggesting that these conditions characterize the conventional model of schooling, research indicating how learned helplessness may be corrected, research indicating that an intervention based on this knowledge is more efficient than 22 other strategies for raising student achievement, research contradicting existing theories that the gap may be traced to sociocultural and socioeconomic differences that exist prior to the age when children enter the school system, research contradicting the oppositional culture explanation, research contradicting theories that the gap may be traced to differences in the quality of the schools and teachers experienced by minority students, research contradicting the hypothesis that "no-excuses" charter schools are the answer to the achievement gap, and research contradicting the hypothesis that the use of value-added statistical methods to identify and replace low-performing teachers are the answer to the achievement gap. The validity of the proposed model is indicated by its consistency with all of these results.

Chapter Two, "Hypotheses," defines the student achievement gap and introduces hypotheses about the nature of the achievement gap, drawing upon research by Betty Hart, Todd Risley, and Nobel-laureate James Heckman implicating the parenting practices of low-income parents. Chapter Two contrasts this conventional view with the views of stratification economist William Darity, who challenges the notion that differences in cultural practices are the primary source of inequality. His view suggests a need to investigate the role of structural factors that contribute to observed disparities in outcomes.

Chapter Three, "A Fresh View," reviews—and rejects—hypotheses that early parenting behaviors, summer setbacks, oppositional peer culture, poverty, discrimination, or differences in school or teacher quality explain the persistence of the student achievement gap throughout the K-12 years. Chapter Three suggests, instead, that conventional school environments are inadvertently structured in ways that trigger learned helplessness, demoralization, and reductions of self-efficacy, effort, and potential achievement that are especially severe among low-income black and Hispanic children who enter kindergarten performing below the level of their white and Asian peers. This explanation for the achievement gap is consistent with much of what is currently known about the nature of schooling and factors that cause learned helplessness, low-self-efficacy, passive, disengaged behavior, and depressed achievement.

Chapter Four, "Evidence from Three National Studies," draws upon evidence from three studies employing nationally-representative samples of students and demonstrates that math self-efficacy and engagement decline at an accelerating rate across race and ethnicity as children progress from grade three to grade eight, but the decline is sharpest for low-income black and Hispanic students, suggesting that the same factor causing the decline in self-efficacy and engagement is a major factor contributing to or causing the persistence of the student achievement gap throughout the K-12 years. The pattern is not easily explained by theories that the achievement gap is due to the quality of schools and teachers experienced by black and Hispanic students. High-scoring students, including those who presumably attend high-quality schools and are taught by high-quality teachers, experience a similar decline in self-efficacy.

Chapter Five, "A New Model of Learning," explains how the conventional school system is structured in ways that perpetuate initial gaps in achievement that exist upon entry in kindergarten. Learned helplessness, passive, disengaged behavior, and depressed achievement may be expected when task difficulty is fixed, students are graded in relation to each other, and, as a consequence, low-performing students routinely have the discouraging experience of receiving low grades. Significantly, this condition is reversible. Individualized task difficulty, in combination with frequent, objective performance feedback, serves to reverse the condition of learned helplessness, promoting engagement and achievement, suggesting that a lack of individualized task difficulty and performance feedback is a major factor contributing to or causing the persistence of the achievement gap throughout the K-12 years.

Chapter Six, "Contradictions Resolved," explains a contradiction. Previous research using value-added statistical models suggested that the contribution of teachers to student achievement is large. However, teacher rankings based on value-added estimates of performance are highly unreliable measures of future performance. Furthermore, those estimates predict not only gains in student performance during the period of instruction, but they also predict student performance *prior* to the point when a teacher has any contact with her students! Chapter Six explains this illogical result by suggesting that value-added models omit controls for self-efficacy, causing the influence of this factor to contaminate value-added calculations of teacher contributions to student achievement. Chapter Six explains why this would account for previous findings that the contribution of teachers to student achievement is large.

Chapter Seven, "Consequences for Minorities," investigates the significance of the finding that student achievement can be raised efficiently throughout the K-12 years by individualizing task difficulty and supplying rapid performance feedback. Chapter Seven draws upon data from the National Education Longitudinal Study (NELS), a nationally-representative sample of eighth-grade students who were followed for 12 years, to investigate the predicted results for racial minorities of implementing a system that raises student achievement over the K-12 years by one standard deviation. The results indicate very large predicted impacts on key educational outcomes. The results predict that the number of black and Hispanic students who attain baccalaureate degrees would double.

Chapter Eight, "No-Excuses Charter Schools," investigates and rejects the hypothesis that so-called "no-excuses" charter schools represent a promising strategy for addressing the achievement gap. Chapter Eight reanalyzes key studies regarding two promising charter school models—the Knowledge is Power Program (KIPP) and the Harlem Children's Zone Promise Academies. The reanalysis indicates that gains would fall to zero once these programs are implemented in every school across the nation. The analysis suggests why existing impact results are likely due to artifacts stemming from high teacher attrition and hoarding of a disproportionate share of the nation's limited pool of highly-dedicated teachers, rather than gains that could be sustained when the programs are scaled up and implemented nationwide.

Chapter Nine, "Better Teachers," investigates and rejects the hypothesis that replacing poor teachers with good teachers represents a promising strategy for addressing the achievement gap. Chapter Nine reanalyzes a key study suggesting that the use of value-added statistical methods to identify and replace low-performing teachers would greatly increase student achievement and earnings. The reanalysis indicates that value-added modeling lacks sufficient reliability and validity for the purpose of hiring and firing teachers. Once gains are averaged over all students, they would be very small. Furthermore, it appears that any gains would fade away very quickly. Significantly, the approach is neither cost-effective nor does it meet the test of a benefit-cost analysis.

Chapter Ten, "22 Strategies," evaluates the major policy interventions that have been proposed to address the achievement gap. The key finding is that a technology-based intervention that individualizes task difficulty, provides rapid performance feedback, and is based on the theory of the achievement gap described in this book is dramatically more cost-effective than any of the other 22 interventions for which data are available. This result offers key evidence supporting the proposed theory of the achievement gap. The efficiency of this approach implies that the underlying theory is correct because it is unlikely that an incorrect theory could produce an efficient solution.

Chapter Eleven, "Solving the Achievement Gap," revisits James Heckman's argument and evidence that the achievement gap may be traced to the parenting practices of low-income parents. This chapter reconciles Heckman's arguments with the evidence presented in this book that the persistence of the gap is due to school-related factors—that is, the harmful psychological effects that occur when students are graded, tested, and compared to their same-age classmates—rather than parenting practices. The evidence that Heckman relies upon is consistent with the thesis that early differences in personality influence achievement at entry into kindergarten but this initial difference is maintained, perpetuated, and magnified after school entry because grading and testing practices systematically demoralize and depress low-performing children over the entire 13-year K-12 period.

In the concluding section of this book I raise a broad question: why has the perspective offered in this book not received more attention? The differences that are observed when children enter school in kindergarten and the apparent differences in the quality of schools experienced by black and white children make it easy to conclude that culture, family advantages, and school quality are sufficient to explain the achievement gap. It appears that there is no need to reexamine the evidence.

The evidence presented in this book, however, points to a structural factor that is embedded in the conventional model of schooling, rather than parenting style or the quality of individual schools. Only a structural factor could exert a systematic influence across race, ethnicity, and level of achievement. This factor must be something that is hidden in plain view because it would not be possible to hide anything this powerful. It is hidden only in the sense that we do not recognize it.

Fixed task difficulty and the use of grades to rank, compare, and categorize students as either above- or below-average characterize the conventional model of schooling and are known to depress engagement and achievement. These features are invisible to the casual observer because they are structural features of every school. It is difficult for most observers to imagine schools without these characteristics.

This book seeks to demonstrate that these characteristics are not benign. Instead, they have extremely powerful unintended negative effects on children. Recognition of this central fact is the key to unraveling the mystery of the achievement gap, understanding the source of the problem, and formulating effective public policies to address it.

CHAPTER 2

Hypotheses

While popular discussions of the student achievement gap focus on the gap in performance between black students and white students, or the gap in performance between Hispanic students and white students, it would be more accurate to characterize the gap as a difference in the average performance of students raised in low-income families compared to students raised in families with higher incomes. The income achievement gap, defined as the average achievement difference between a child from a family at the 90th percentile of the family income distribution and a child from a family at the 10th percentile, is nearly twice as large as the black-white achievement gap.[1] The income achievement gap has been growing for at least 50 years and is roughly 30 to 40 percent larger among children born in 2001 than among those born 25 years earlier.[2] Black and Hispanic students from high-income families perform well, while white students from low-income families perform poorly. On average, however, black and Hispanic students tend to be raised in families with lower mean incomes and less wealth, compared to white students, and are disproportionately affected by gaps in performance linked to income and wealth. For this reason, I continue to frame the discussion of the achievement gap in terms of the gap in performance between black students and white students, and the gap in performance between Hispanic students and white students.

Various hypotheses have been proposed to explain the existence of the achievement gap. Black and Hispanic family incomes and wealth are lower, on average, compared to the income and wealth of white households. These differences in income and wealth are associated with differences in the educational opportunities that are available to black and Hispanic children, versus white children, and these differences in opportunities are associated with differences in achievement that exist when children enter kindergarten.

Betty Hart and Todd Risley, however, found that the quality and quantity of the linguistic interactions between parents and their children during the period before the children entered kindergarten, rather than income, predicted achievement outcomes. Children whose parents provided a rich linguistic environment

S.S. Yeh, *Solving the Achievement Gap*,
DOI 10.1057/978-1-137-58767-1_2

were more advanced linguistically and intellectually when they started school, and performed better in school, than children whose parents did not provide the same rich linguistic environment. Hart and Risley found that during the preschool years, children raised by parents who were receiving welfare and were disproportionately black, heard—and were exposed to—less than one-third of the quantity of words that were heard by children raised by parents who were working in professional occupations and were disproportionately white.[3] Hart and Risley projected that, by age four, children in families receiving welfare were exposed to 30 million fewer words than children raised in professional families.[4] The children who performed best at ages nine and ten were those who heard the most words, were given the most feedback, received the most positive feedback, and received the most complete answers to their questions during their preschool years.[5]

Other studies demonstrated that low-income minority children live in homes with fewer books, have fewer books available in the school and classroom library, and live farther from public libraries than do children raised by middle- and upper-income families.[6] There is evidence that parents of low-income minority children read fewer books to their children, compared to the parents of middle-class white children, and evidence that these differences influence vocabulary and reading development.[7] The child-directed speech of upper-middle-class mothers is more contingent on their children's speech, and less directive, compared to the child-directed speech of working-class mothers.[8] Middle-class parents are more likely to teach their preschool children to blend phonemes and sound out words, compared to low-income parents.[9] Middle-class children ages three to six are more likely to be able to recognize letters, count to 20, write their names, and read or pretend to read, compared to low-income children.[10] These findings led many researchers to conclude that race-related sociocultural factors and parenting style are mainly responsible for the achievement gap.[11]

The analysis of nationally-representative samples of students reported in Chapter Six indicates that race and ethnicity are also associated with significant differences in the educational preparation of children with regard to mathematics, measured at entry into kindergarten. Using standardized measures of mathematics achievement, black children enter kindergarten performing 0.24 standard deviations (SD) below the performance of white children. Hispanic children enter kindergarten performing 0.42 SD below white children. American Indian children enter kindergarten performing 0.19 SD below white children. Children from Hawaii and the Pacific islands enter kindergarten performing 0.46 SD below white children. These differences seem to implicate race-related sociocultural and family influences, rather than schools, as the main cause of the achievement gap.

In addition to differences in cognitive skills, research conducted by Nobel laureate James Heckman suggests that the parenting practices of low-income parents contribute to differences in noncognitive skills that exist when their children enter the school system and persist into adulthood. Based on studies conducted over the past four decades, Heckman found that individuals with higher levels of noncognitive skills, defined as strength of motivation, an ability to act on long-term plans, and the socio-emotional regulation needed to work with others, are more likely to complete high school and college.[12] Deficits in noncognitive skills cause high school equivalency diploma recipients to drop out of high school and underperform later in life.[13] They may lack the abilities to think ahead, to persist in tasks, or to adapt to their environments.[14] Conscientiousness, defined as the tendency to be organized, responsible, and hardworking, predicts years of schooling with the same strength as measures of intelligence.[15] Conscientiousness predicts college grades to the same degree that SAT scores do, and personality measures predict performance on achievement tests and, to a lesser degree, performance on intelligence tests.[16] Personality traits strongly predict four-year college graduation rates at all deciles of the personality distribution.[17]

Heckman and his colleagues presume that these noncognitive traits are established prior to the point when children enter the school system, resulting in differences in achievement that exist upon entry and persist throughout each child's academic career: "The gaps in cognitive achievement by level of maternal education that we observe at age eighteen—powerful predictors of who goes to college and who does not—are mostly present at age six, when children enter school."[18] Heckman concluded: "Schooling—unequal as it is in America—plays only a minor role in alleviating or creating test score gaps."[19] Instead, "Personality traits predict and cause outcomes."[20] Heckman drew upon impact evaluations of the Perry preschool program to argue that the results demonstrate how personality traits can be changed in ways that produce beneficial lifetime outcomes.[21] Based on these results, Heckman recommended a strategy of investing in high-quality early childhood programs, such as the Perry and Abecedarian preschool programs, which explicitly teach children the traits and behaviors of perseverance, cooperation with other children, and formulating and acting on plans. Economist Samuel Bowles reviewed Heckman's analysis and spelled out the implied conclusion: "Heckman makes a strong case that government should provide substitutes to fill the gaps when poor parents—too preoccupied and stressed out making ends meet or getting adequate medical care, or for any other reason—do not do all that is needed for their kids."[22]

Thus, the hypothesis that the cause of the achievement gap may be traced to deficiencies in cognitive and noncognitive skills that are caused by poor parenting is now well-established. However, economist William Darity argues that it is premature to blame cultural practices for the racial inequality that characterizes the United States with regard to key outcomes, including educational achievement:

> The conventional wisdom embodied in public policy discourse, popular media representations, and much social science research about intergroup disparity has it that group-based deficits in personal responsibility and cultural practices are explanatory. Indeed, the politically acceptable trope of cultural determinism progressively has displaced the impolite trope of biological determinism…Economists of all stripes have been far from immune from this tendency; they include an array as diverse as George Akerlof and Rachel Kranton, Glenn Loury, David Austen-Smith and Roland Fryer, Tom Sowell, Walter Williams, and Barry Chiswick. Incantations of group dysfunction now substitute for incantations of genetic inferiority (although I fear that the latter magic has not vanished altogether).[23]

Darity argues, along with other stratification economists, that there is a need for a closer examination of structural processes that generate inequality:

> Stratification economics examines the structural and intentional processes generating hierarchy and, correspondingly, income and wealth inequality between ascriptively distinguished groups. For the stratification economist, claims about the defectiveness of a group with outcast/caste status are an ideological mask that absolves the social system and privileged groups from criticism for their role in perpetuating the condition of the dispossessed.[24]

Darity is especially critical of Abigail and Stephan Thernstrom, authors of *No Excuses: Closing the Racial Gap in Learning*, which posits that the racial gap in academic performance between white and Asian, versus Hispanic and black students, may be traced to differences in culture and parenting styles that are best addressed through "no excuses" charter schools such as the Knowledge is Power Program (KIPP) and Harlem Children's Zone Promise Academies that teach, and demand that students learn, not only academic skills but the disciplined, respectful, hardworking attitudes and behaviors that are adaptive and rewarded in American society.

It is difficult to deny that race and culture play an important role in explaining differences in educational preparation that exist at entry in kindergarten. And it is difficult to deny the Thernstroms' basic point that disciplined, respectful, hardworking attitudes and behavior are important and must be taught. At the same time, Darity's point is that there is a need to go beyond conventional wisdom to identify ways in which American society is either intentionally or inadvertently structured to perpetuate inequality.

If the source of inequality is deemed to be cultural, as Heckman's studies seem to suggest, then it may be convenient to stop searching for structural forces that perpetuate inequality. It may be tempting to point a finger at culture and resign ourselves to the task of teaching adaptive attitudes and behaviors to children whose cultural experiences do not adequately prepare them to be successful in American society. However, if the root cause of inequality is structural, then the only way real progress will be made is if the root structural causes are identified and addressed. That is the topic of this book.

CHAPTER 3

A Fresh View

The Coleman report was the first national study to document the existence of substantial differences in educational achievement between black and white students at every grade level—differences that increased as students progressed from first through twelfth grade.[1] Almost all research on the topic concludes that the black-white achievement gap grows during the school years, particularly in elementary school.[2] A clue to the mystery is a finding that was originally reported in the Coleman study and has subsequently been widely replicated. The Coleman study found that the strongest predictor of student achievement for black and Hispanic students was a student's perceived control over his or her environment.[3] For black and Hispanic students, this factor was stronger than any other school or background variable, including parental education.[4]

Subsequent studies focused specifically on the influence of a student's perceived control over his or her academic performance.[5] A closely related construct is academic self-efficacy: the degree to which an individual believes that he or she can accomplish academic tasks such as reading and math. Perceptions of self-efficacy appear to have a strong influence on effort and persistence with difficult tasks, or after experiences of failure.[6] If students have a very low level of self-efficacy for a given task, they are unlikely to attack the task with much enthusiasm or persistence.[7] Conversely, students with high levels of self-efficacy tend to display a mastery orientation.[8] They seek challenge, persist in the face of obstacles, enjoy exerting effort in the pursuit of task mastery, and seem undaunted by challenges.[9]

The relationship between academic self-efficacy and achievement is reciprocal: gains with respect to one factor lead to gains in the other, in a virtuous, self-reinforcing cycle.[10] The main mechanism for building self-efficacy in a particular domain appears to be experiencing repeated success on tasks in that domain[11] while ensuring that the difficulty of the tasks is appropriate.[12] Conversely, losses in one factor lead to losses in the other, in a self-reinforcing negative downward spiral.[13] Repeated failure leads to lowered self-efficacy.[14] When setbacks occur, intrinsic motivation declines.[15] This suggests a psychological explanation for

S.S. Yeh, *Solving the Achievement Gap*,
DOI 10.1057/978-1-137-58767-1_3

low minority student achievement: low-performing minority students who feel that they have lost control over their academic performances become discouraged and disengaged, resulting in a downward spiral.

Theories

One possibility is that early differences in parenting or other sociocultural and socioeconomic factors contribute to initial differences in academic achievement that are magnified over time as low-performing students become more and more discouraged. For example, Hart and Risley found that preschool children raised by parents on welfare heard, and were exposed to, less than one-third of the quantity of words heard by children raised by parents working in professional occupations.[16] Thus, one hypothesis is that black children may be raised in environments that are less conducive to high educational attainment.

Differences in early parenting behaviors explain approximately one-third of the black-white gap in mathematics skills at the beginning of kindergarten that remains after taking conventional family background characteristics into account.[17] However, the impact of early maternal sensitivity on mathematics skills declines as students advance from grade to grade—differences in early parenting account for approximately one-quarter of the black-white mathematics skills gap at the end of third grade that remains after taking conventional family background variables into account.[18] Furthermore, after controlling for an array of variables that capture a child's environment, the residual gap increases more than the raw gap (without controls) as black and white children advance from grade to grade.[19] In addition, black students lose ground across virtually all types of skills, not merely skills requiring mastery of abstract concepts or higher-order thinking.[20] This pattern of results is consistent with the hypothesis that home environment contributes to early differences that exist at entry into kindergarten. The pattern suggests that the influence of home environment declines as students advance from grade to grade. The pattern suggests that unmeasured factors gain influence after students enter the school system. The pattern is inconsistent with the hypothesis that home environment is the primary factor driving the persistence of the achievement gap throughout the K-12 years.

Thus, while some researchers continue to focus on evidence that the bulk of the gap arises prior to school entry, implicating differences in parenting styles or socioeconomic and sociocultural factors that influence the family environment surrounding low-achieving minority students, rather than school-related factors, Jencks and Phillips argue:

> This conclusion strikes us as premature. If stability were the natural order of things and did not need to be explained, the fact that blacks typically score at about the same percentile in first and twelfth grades would be evidence that schools did not contribute to the gap. In human development, however, change is at least as common as stability. That means we have to explain stability as well as change.[21]

The persistence of the achievement gap over the K-12 years is a phenomenon that requires explanation. From a theoretical perspective, one might expect the importance of parenting style and socioeconomic and sociocultural factors to decline with age since the proportion of a school-age child's life spent in school steadily increases with age. The empirical evidence suggests that the influence of parenting style does, in fact, decline with age—but the gap continues to persist. The persistence of the gap in the face of strenuous interventions by teachers throughout the academic careers of low-achieving students suggests the need for an explanation that identifies a ubiquitous factor (or factors) that consistently reinforce downward spirals throughout the K-12 years.[22] To assert that the differences that exist when children start kindergarten govern their entire school careers begs the question of why those differences persist. Failing to identify a mechanism that would maintain the gap in achievement, the differences-at-kindergarten explanation is unsatisfactory.

A second hypothesis is that black students lose more ground over the summer than white students as a consequence of worse home and neighborhood environments.[23] However, analysis of data from a nationally-representative sample of over 20,000 children lends little support to the hypothesis that differential summer setbacks explain the lost ground of black students.[24]

A third hypothesis seeking to explain the persistence of the achievement gap involves oppositional peer culture—the notion that black students might encounter peers who are hostile to blacks who adopt white ways of speaking and acting and demonstrate high-achievement.[25] However, the evidence from nationally-representative datasets either does not support the oppositional peer culture explanation[26] or suggests that the phenomenon is limited to the most segregated schools in the nation.[27] Thus, oppositional culture can only explain the persistence of the achievement gap at a limited number of schools.

A fourth hypothesis is that poverty or factors related to poverty explain the persistence of the achievement gap. However, an experimental test does not support the assertion that addressing poverty would fix low student achievement. The *Moving to Opportunity for Fair Housing Demonstration Program* randomly-assigned 4,600 low-income families to a program that permitted the families to relocate to middle-class neighborhoods.[28] However, the children in these families did not demonstrate improved academic achievement.[29] This contradicts the hypothesis that fixing poverty would fix low achievement.

A fifth hypothesis is that white teachers have lower expectations for black children or otherwise discriminate against them in the classroom.[30] However, evidence from a nationally-representative dataset does not support this hypothesis. The black-white test score gap widens to a greater extent between fall of the kindergarten year and spring of first grade for children whose kindergarten and/or first-grade teacher is black, compared to children whose kindergarten and first-grade teachers are not black.[31] This result is the exact opposite of the result that would be predicted if white teachers discriminate against black students.

A sixth hypothesis is that differences in the quality of schools attended by black and white students account for the persistence of the achievement gap.[32] However, while black and white students attend schools that differ with regard to the percentage of students eligible for free lunch, the degree of gang problems in school, the amount of loitering in front of the school by nonstudents, and the amount of litter around the schools, these measures explain only a small fraction of the variation in student outcomes.[33] Furthermore, the inclusion of these covariates does not prevent the black-white gap from widening over the kindergarten and first-grade years.[34] By the end of first grade, black students have lost the equivalent of almost three months of schooling relative to whites.[35] In addition, both Hispanic and Asian students experience worse schools than whites, but neither of those groups loses ground over time.[36] Finally, the black-white gap grows through third grade even between black and white students attending the same school.[37] All of the ground lost by black students between first grade and third grade is within, rather than across, schools.[38] Black students lose substantial ground relative to white students within the same school and even within the same classrooms.[39] Controlling for school-fixed effects or teacher-fixed effects in models of achievement does little to explain the divergent trajectories of black and white students between kindergarten and third grade.[40] This pattern of results is inconsistent with the hypothesis that differences in school or teacher quality explain the black-white achievement gap. The pattern is, instead, consistent with the hypothesis that there is something about the interaction between black students and schools that interferes with the learning process.[41]

A New Model

The purpose of this book is to propose a new model of the achievement gap. This model suggests that there is something about the interaction between disadvantaged minority students and schools that interferes with the learning process. The model incorporates the known reciprocal relationship between student self-efficacy and student achievement but goes beyond this by identifying two ubiquitous features of conventional school environments that appear to trigger and reinforce this reciprocal relationship throughout the academic careers of low-achieving students. The contribution of the book is the identification of this triggering/reinforcement mechanism. The significance of this contribution is that it potentially explains how downward spirals may be triggered and reinforced in the routine daily operation of a typical school for many—if not most—low-achieving students. The book seeks to demonstrate that this explanation is consistent with much of what is currently known about the nature of schooling and factors that cause learned helplessness and passive, disengaged behavior—factors that may be presumed to trigger low self-efficacy and a downward spiral in achievement for students who arrive at kindergarten performing below the level of their more advantaged peers.

This explanation is consistent with the observed persistence of the achievement gap over the K-12 years. It is consistent with the observation that nationally-representative datasets do not support the hypotheses of oppositional culture, summer setbacks, or teacher discrimination. It is consistent with the evidence that the achievement of observationally equivalent black and white children diverges after entry into school. It is consistent with the observation that available measures of school quality explain only a small fraction of the variation in student outcomes. It is consistent with the evidence that the black-white gap grows through third grade even between black and white students attending the same school. Finally, it is consistent with the evidence that black students lose substantial ground relative to white students not only within the same schools but also within the same classrooms.

This book reports original analyses of nationally-representative datasets and reviews empirical literature including studies suggesting that an intervention based upon the proposed model of the achievement gap is more efficient in raising achievement than 22 other approaches. This array of results is consistent with the proposed model of the gap. To the extent that no other theory adequately explains this pattern of results, the results may be interpreted as evidence supporting the proposed diagnosis of the achievement gap.

Learned Helplessness

Students who are immersed in a task environment that fails to produce feelings of success may become discouraged and their level of effort and achievement may be depressed.[42] The phenomenon of "learned helplessness" is well-known and has been studied in laboratory experiments. The original experiment involved dogs; when subjected to electric shocks that could not be avoided, the dogs became passive and did not attempt to avoid similar shocks even after the conditions were altered so that the dogs could avoid the shocks.[43]

Research regarding learned helplessness has been extended to humans.[44] When humans are placed in situations where they have little control over aversive stimuli and their efforts to avoid these situations do not produce success, they may become discouraged and passive. Examples range from the experiences of individuals who were imprisoned at Auschwitz, the experiences of assembly-line workers, and individuals who experience chronic pain and illness.[45]

Similarly, students who are regularly presented with tasks that are too difficult may be unable to avoid the painful experience of low grades.[46] Students who are regularly presented with tasks that are too easy may find that they are unable to avoid the painful experience of boredom.[47] While both groups of students are subjected to conditions that foster passivity and depress achievement, the effects of low grades for low-performing students are likely to be more painful and severe, leading to a larger decrease in engagement and performance over time.[48] This would explain the increase in the achievement gap as low-performing students advance from grade to grade.[49]

When adolescents cannot perform the same work as their peers and cannot succeed they quickly become resentful and profoundly discouraged.[50] Four nationally-representative surveys, including a longitudinal survey of approximately 30,000 students, suggest that students who become disengaged and drop out of school do so because they do not have much success in school and do not like it.[51] Grade retention and low grades are associated with disengagement and tend to predict dropping out of high school.[52]

Children's engagement in school declines during the elementary school years.[53] Disaffection climbs during the transition to middle school, between the fifth and the sixth grades.[54] National studies indicate that students become more disengaged from school as they advance from elementary to middle to high school.[55] A 2014 Gallup survey of over 800,000 students indicates that student engagement declines precipitously between grade five and grade twelve.[56] Consequently, 40 percent of all high school students may become chronically disengaged.[57] In sum, children's intrinsic motivation for schooling deteriorates steadily beginning in kindergarten and continuing until the end of high school.[58] It is difficult to escape the conclusion that for many students schooling tends to reduce, rather than increase, interest in learning.[59]

A Loss of Control

Studies suggest that this pattern of disengagement and low achievement may be explained by the universal practice of comparing students through classroom tests and letter grade report cards, causing low-performing children to experience anxiety, demoralization, and a loss of control.[60] Normative grading that emphasizes social comparison and competition produces undesirable consequences for most students, including reduction of intrinsic motivation.[61] Norm-referenced evaluation tends to undermine the self-efficacy, motivation, and learning of students who regularly score near the bottom of their classes.[62] Students with low self-efficacy with regard to academic tasks exhibit anxiety, lose concentration and focus, tend to avoid challenge, and display low persistence and helpless behavior when facing difficult tasks.[63] Over time, their effort decreases, they experience cognitive demobilization and inhibited generation of ideas, they initiate fewer responses, and they become discouraged, despondent, and passive.[64] They are severely hampered in the acquisition and display of cognitive skills when facing obstacles.[65] Think-aloud protocols indicate that helpless children focus on their failures and spend little time searching for ways to overcome failure and solve problems.[66] Children who receive low test scores and report card grades experience high levels of anxiety that adversely influence performance.[67] The magnitude of this effect increases steadily across the K-12 school years.[68]

Children's perceived control exerts a strong effect on their academic achievement.[69] In turn, children's actual school performances influence their sense of control.[70] Children with profiles reflecting high control are more likely to maintain concentration, attention and effort, persist in the face of failure, and remain optimistic and interested in academic activities.[71] They are more likely to select challenging tasks, set high and concrete goals, and form well-structured sequential plans.[72] They initiate action, exert effort, and persist in the face of obstacles.[73] They focus on the task at hand and do not ruminate about the causes of failure.[74] They generate and monitor strategies flexibly and test alternative hypotheses systematically.[75] They are able to maintain access to their entire cognitive and problem-solving repertoires.[76] Even if their efforts are unsuccessful, they are able to maintain an optimistic outlook.[77]

In contrast, children who perceive school outcomes as due to factors beyond their control and who doubt their own capacities are more likely to give up in the face of challenges and become discouraged, anxious, and passive.[78] Children who earn lower grades are more likely to doubt not only their capacity to exert effort, but also their own ability.[79] Children who doubt their control and ability are more likely to show a pattern of increasing disaffection.[80]

Longitudinal studies indicate that children's early academic performances predict their five-year trajectories of beliefs about their perceived control over academic outcomes, which influences academic engagement, which feeds back and influences academic performance in a reciprocal relationship.[81] This reciprocal relationship between academic performance and control beliefs extends into the upper grades as well. Using panel data from a nationally-representative sample of 8,802 American students, Ross and Broh found that academic achievement in eighth grade predicted an internal locus of control in tenth grade, which in turn predicted academic achievement in twelfth grade.[82] By far the strongest predictor of academic achievement in twelfth grade is earlier academic achievement. A large-scale study involving a random sample[83] as well as meta-analytic studies[84] confirm that children's beliefs about their degree of control over academic outcomes are strong predictors of academic achievement.

The picture that emerges is that children who start school academically and socially behind their peers and are subsequently presented with tasks geared to the average student in their cohort tend to have less success, begin to doubt their own abilities, are less engaged, less persistent, and exhibit lower performance.[85] By the seventh grade, a record of cumulative failure makes it nearly impossible to maintain a sense of control.[86] These children are especially likely to become anxious, angry, passive, apathetic, and depressed.[87] Conversely, children who start school academically and socially ahead of their peers and whose early efforts are met with academic success tend to maintain confidence in their academic abilities and engagement in school, further enhancing their academic skills and achievement.[88] These syndromes—one negative, the other positive—provide insight into a process that may perpetuate early differences in academic performance and may explain the persistence of the academic achievement gap between children from poor families and their more advantaged peers throughout the K-12 years.[89]

CHAPTER 4

Evidence from Three National Studies

The Early Childhood Longitudinal Study of the Kindergarten Class of 1998–1999 (ECLS-K) followed a nationally-representative cohort of 21,260 children from kindergarten into middle school.[1] The data suggest a pattern of results that is difficult to explain except in terms of the loss of control hypothesis outlined in the previous chapter. Student achievement in math is higher for students with high self-efficacy in math (respondent agreement with "I am good at math"), compared to students with low self-efficacy, from grade three to grade eight (Figures 4.1, 4.2, 4.3, and 4.4).

These results are consistent with the hypothesis that initial differences in achievement lead to differences in self-efficacy. However, math self-efficacy

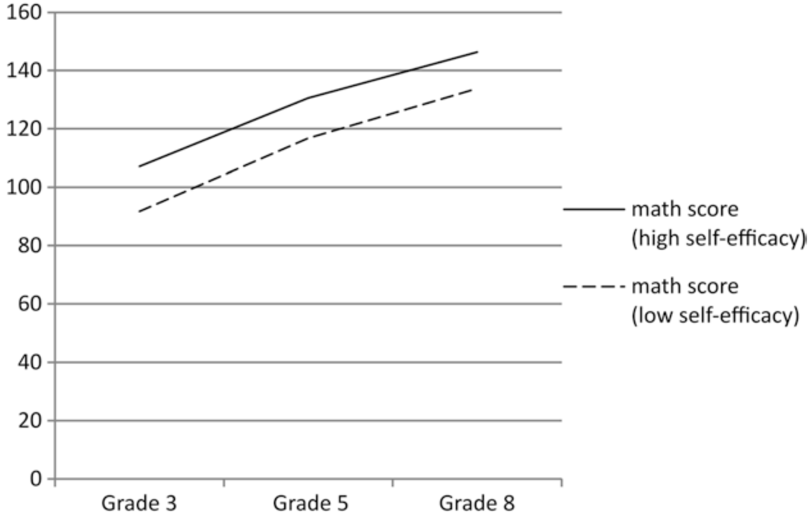

Figure 4.1 Math scores are higher for white students with high self-efficacy

S.S. Yeh, *Solving the Achievement Gap*,
DOI 10.1057/978-1-137-58767-1_4

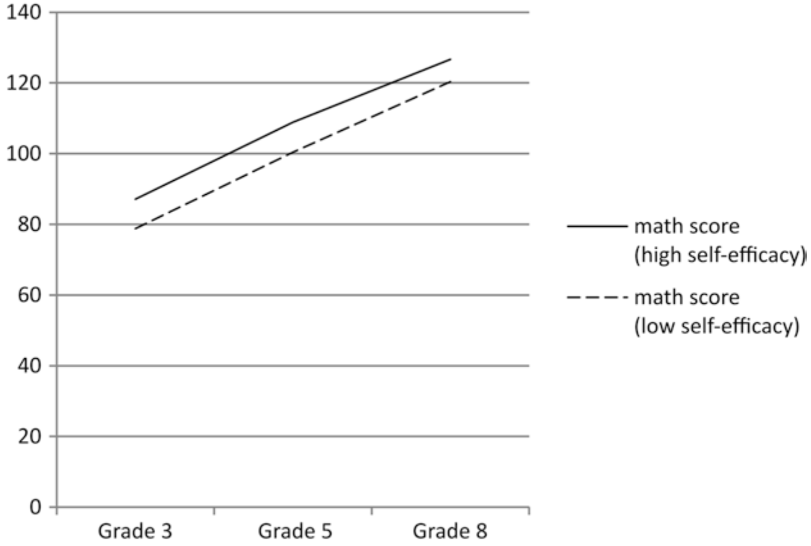

Figure 4.2 Math scores are higher for black students with high self-efficacy

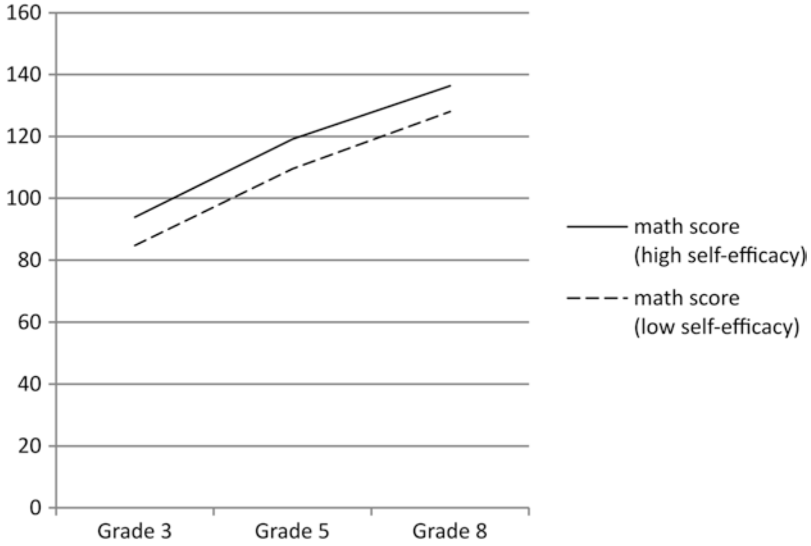

Figure 4.3 Math scores are higher for Hispanic students with high self-efficacy

declines at an accelerating rate as students advance from grade three to grade eight, even though math achievement increases (Figures 4.5, 4.6, and 4.7). The breadth of the decline in self-efficacy is remarkable. Math self-efficacy declines for almost all students (Figure 4.8).

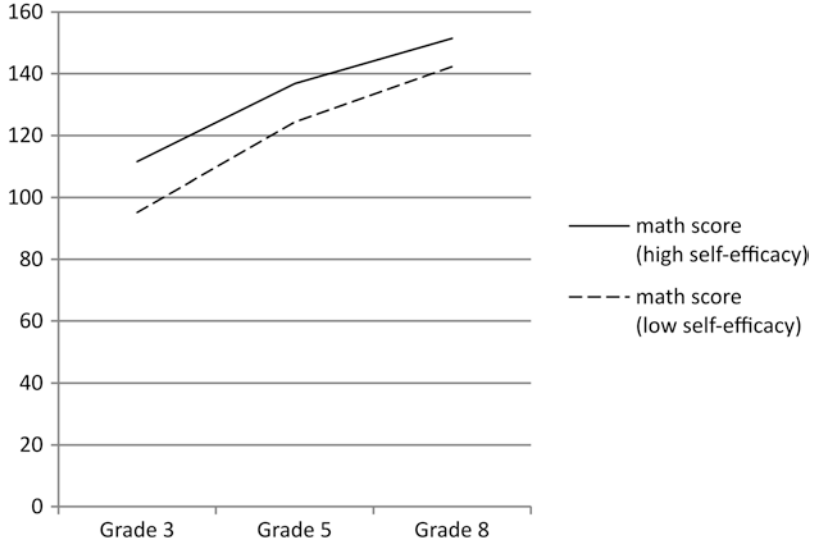

Figure 4.4 Math scores are higher for Asian students with high self-efficacy

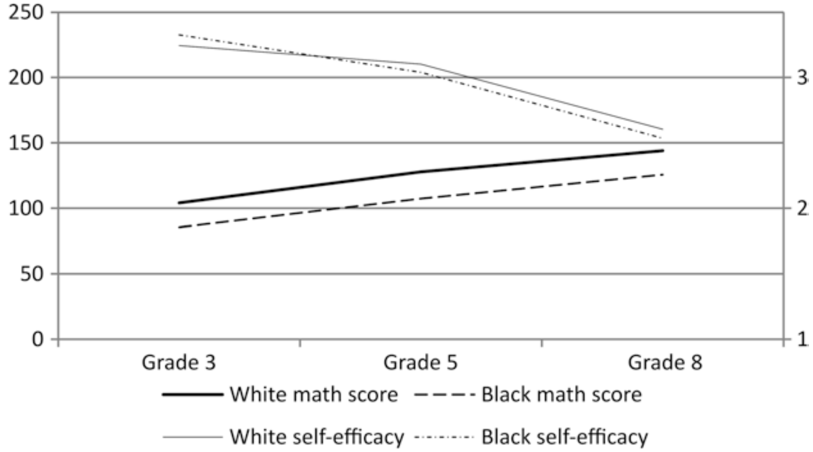

Figure 4.5 Self-efficacy of black students declines at an accelerating rate despite rising math scores

Erosion occurs even among the top 0.5 percent of all students. This is consistent with the hypothesis that otherwise high-achieving students may be discouraged by receiving an occasional bad grade. The decline in self-efficacy accelerates between grade five and grade eight and is sharpest for the lowest-performing students. The pattern is consistent with the hypothesis that exposure to low letter-grade report cards erodes self-efficacy and erosion of

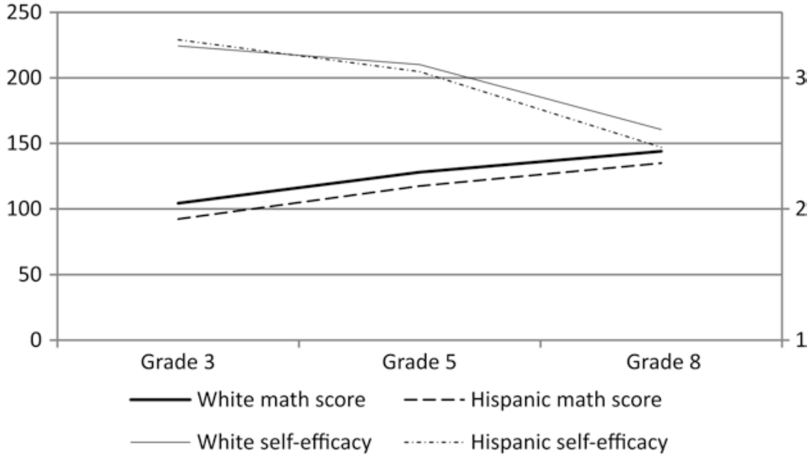

Figure 4.6 Self-efficacy of Hispanic students declines at an accelerating rate despite rising math scores

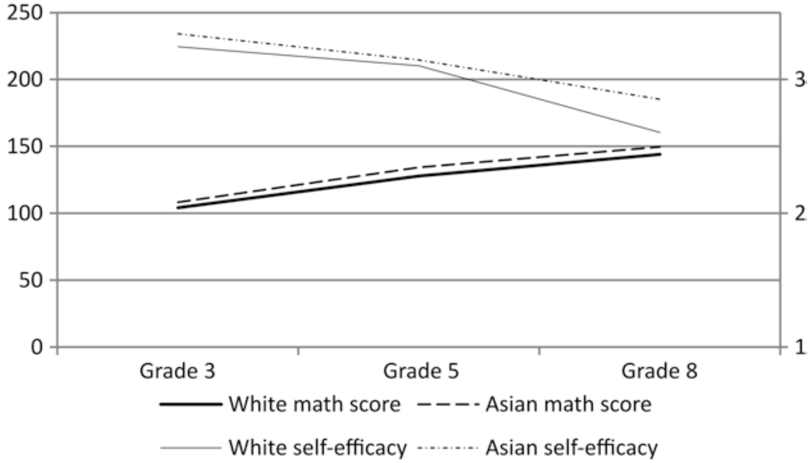

Figure 4.7 Self-efficacy of Asian students declines at an accelerating rate despite rising math scores

self-efficacy translates into depressed achievement, especially for low-achieving students. In contrast, while initial differences in achievement might explain a downward spiral for low-achieving students, initial differences cannot explain the downward spiral in self-efficacy for high-achieving students.

The pattern is not easily explained by theories that the achievement gap is due to the quality of schools and teachers experienced by black and Hispanic students. High-scoring students, including those who presumably attend high-quality schools and are taught by high-quality teachers, experience a similar

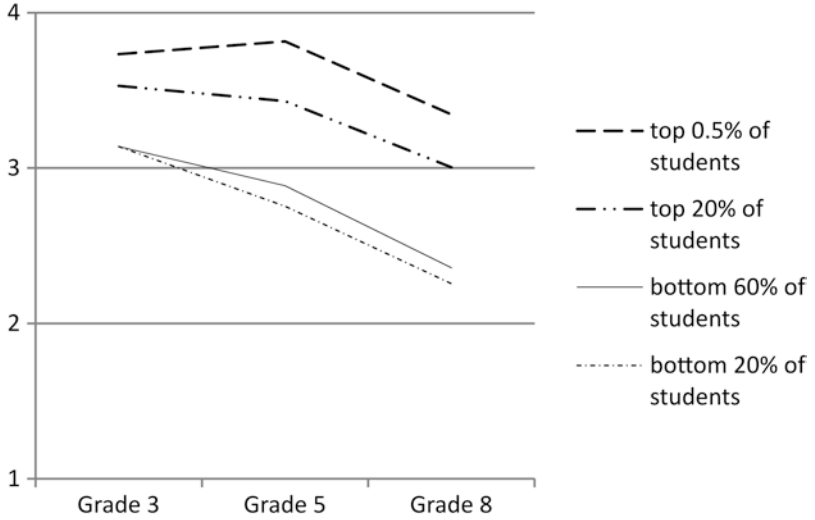

Figure 4.8 Self-efficacy declines at an accelerating rate for almost all students

decline in self-efficacy. The pattern is consistent with the hypothesis that fixed task difficulty, plus the universal practice of grading, ranking, and comparing students to their same-age classmates—exacerbated by the introduction of letter-grade report cards in middle school—serves to discourage any student who does not record straight-A letter grades. This hypothesis suggests that students who previously believed that they were very capable are forced to confront evidence that they are not as capable as they previously believed, resulting in a decline in self-efficacy and a loss of perceived control as they advance from grade three to grade eight.

One might argue that the decline in self-efficacy simply reflects more accurate self-assessments and, furthermore, greater accuracy is desirable because students might be motivated to work harder. However, enjoyment of math declines in parallel with the decline in self-efficacy as students advance from grade three to grade eight (Figure 4.9).

This indicates that the interaction between students and the school system has an increasingly negative effect on both student engagement and self-efficacy as students advance to grade eight. Furthermore, the dual decline in both engagement and self-efficacy at an accelerating rate for all four ethnic groups—white, black, Hispanic, and Asian—suggests that the phenomenon is not limited to minority students and, therefore, is not limited to students who may experience below-average schools and teachers. Regardless of race, ethnicity, and quality of schooling, students become increasingly disengaged and demoralized as they advance to grade eight, and the disengagement and demoralization accelerates.

The cause of the loss of control experienced by black and Hispanic students occurs after these students enter school (Figures 4.5 and 4.6). Until grade

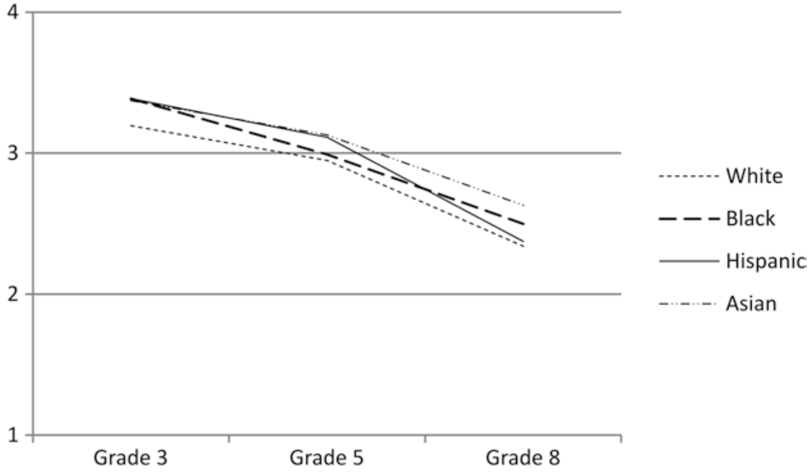

Figure 4.9 Enjoyment of math declines in parallel with the decline in self-efficacy

three, it appears that their levels of self-efficacy exceed the comparable levels of white student self-efficacy, despite lower average levels of achievement. This suggests that the influence of black and Hispanic parents (or related sociocultural factors) is positive, and appears to equip black and Hispanic children with higher—not lower—levels of self-efficacy, compared to white children, prior to school entry. However, some factor related to the interaction of black and Hispanic students with the school system causes their levels of self-efficacy to fall below the comparable levels of white student self-efficacy by grade five, and this interaction causes the decline in self-efficacy to accelerate by grade eight. This pattern of results, in a nationally-representative sample of students, suggests that the cause of the decline in self-efficacy may be traced to school-related factors, rather than differences in parenting or other sociocultural or socioeconomic differences that exist prior to the age when students enter the school system.

The self-efficacy of low-income children (family income below the poverty level) drops below the self-efficacy of middle-income children (family income above the poverty level) after grade three (Figure 4.10). The self-efficacy of low-SES children (who fall in the bottom quintile of the SES distribution) drops below the self-efficacy of high-SES children (those who fall in the top four quintiles of the SES distribution) after grade three (Figure 4.11). This suggests that the cause of the decline in self-efficacy experienced by low-income and low-SES children occurs after these children enter school. Until grade three, it appears that their levels of self-efficacy exceed the self-efficacy levels of middle-income and high-SES children, despite lower average levels of achievement. This suggests that the influence of low-income and low-SES parents (or related sociocultural factors) is positive, and appears to equip low-income and low-SES children with higher—not lower—levels of self-efficacy, compared to middle-income

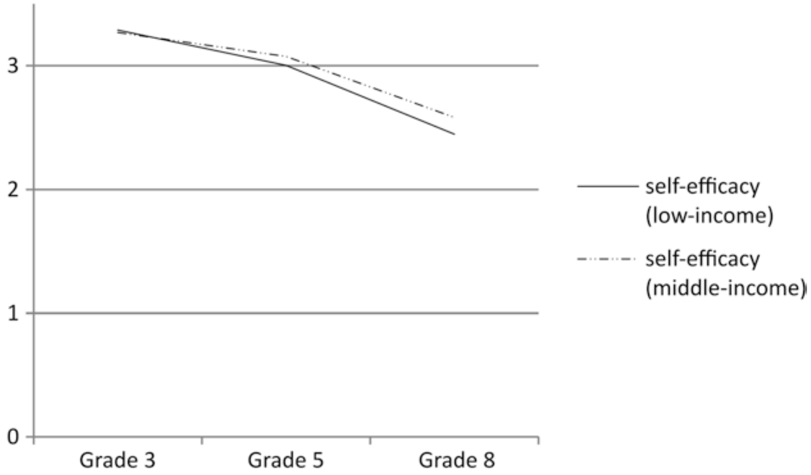

Figure 4.10 Self-efficacy of low-income children falls below self-efficacy of middle-income children after grade 3

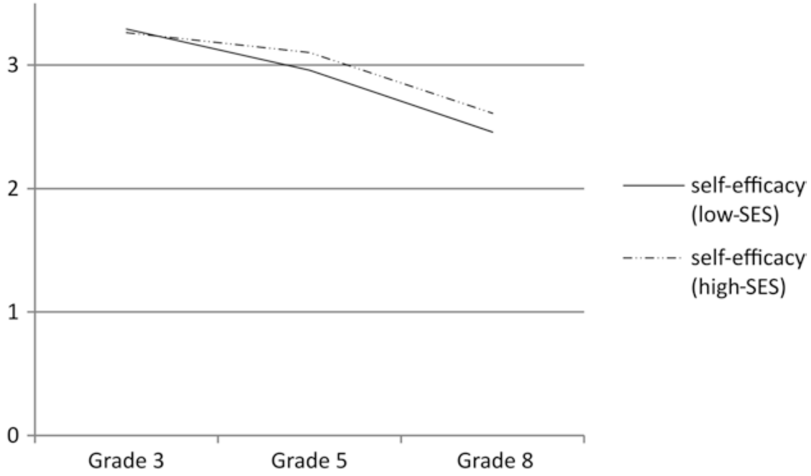

Figure 4.11 Self-efficacy of low-SES children falls below self-efficacy of high-SES children after grade 3

and high-SES children, prior to school entry. However, some factor related to the interaction of low-income and low-SES children with the school system causes their levels of self-efficacy to fall below the comparable levels of middle-income and high-SES student self-efficacy by grade five, and this interaction causes the decline in self-efficacy to accelerate by grade eight. This pattern of results, in a nationally-representative sample of students, suggests that the cause of the decline in self-efficacy may be traced to school-related factors, rather than differences in parenting or other sociocultural or socioeconomic differences that exist prior to the age when children enter the school system.

Figure 4.7 suggests that the cause of the decline in self-efficacy is not due to racial differences in student achievement, or the resentful demoralization that might occur because black and Hispanic students resent the gap in achievement relative to white students, or the quality of schools or teachers experienced by black and Hispanic students. Figure 4.7 indicates that white and Asian students suffer a similar decline in self-efficacy and, furthermore, the self-efficacy of Asian students declines despite their superior academic performance relative to whites. If, on average, white and Asian students attend higher-quality schools and are taught by higher-quality teachers, it is unclear why their self-efficacy declines, and why it declines at an accelerating rate. Figure 4.9 indicates not only that white students suffer a similar decline in engagement compared to black, Hispanic, and Asian students, but the level of white engagement remains below the levels of engagement for the other groups at every grade from grade three to grade eight. If, on average, white students attend higher-quality schools and are taught by higher-quality teachers, it is unclear why their level of engagement is lower. These results are difficult to reconcile with racial theories of resentful demoralization or theories that the achievement gap is due to the quality of schools and teachers experienced by black and Hispanic students, but are consistent with the hypothesis that almost all students experience a loss of control and engagement as they are subjected to fixed task difficulty and the universal practice of grading, ranking, and comparing them to their same-age classmates. A parallel analysis with regard to reading achievement, reading self-efficacy, and reading engagement (not reported here) suggests that this conclusion is not limited to math achievement, math self-efficacy, and math engagement.

Consistent with this evidence, the income achievement gap, defined as the average difference in achievement between a child from a family at the 90th percentile of the family income distribution and a child from a family at the 10th percentile, has been growing within the white, black, and Hispanic student populations separately, as well as within the overall student population as a whole.[2] This suggests that the cause of the achievement gap is a structural factor that is independent of race. If the cause of the gap was specific to black and Hispanic families, then fluctuations in the level of this causal factor would affect black and Hispanic student populations but not the white student population. However, it appears that the factor driving the achievement gap is affecting white, as well as black and Hispanic, student populations.

A second empirical finding points to a structural cause that is related to income, rather than race. For cohorts of students born in the 1940s, 1950s, and 1960s, the black-white achievement gap was substantially larger than the income achievement gap.[3] However, for cohorts born in the 1970s and afterwards, the opposite is true.[4] This pattern points to a structural factor related to family income, rather than race. The pattern is consistent with the hypothesis that children from low-income families enter kindergarten performing below their more-advantaged peers and are then systematically demoralized by existing grading and testing practices. This occurs across race and ethnicity but disproportionately affects black and Hispanic students because they are disproportionately raised in families with low income.

The growth in the income achievement gap is not easily explained by rising income inequality.[5] If rising income inequality was the cause, the sharper growth between 1967 and 1987 of the 50/10 income ratio (i.e., the ratio of family income received by families at the 50th percentile of the income distribution compared to income received by families at the 10th percentile), compared to the growth of the 90/50 income ratio, should be reflected in sharper growth for cohorts of students born between 1967 and 1987 of the 50/10 income achievement gap compared to the 90/50 income achievement gap.[6] However, the 90/50 income achievement gap grew faster than the 50/10 income achievement gap during the 1970s and 1980s, the opposite of what would be predicted on the basis of the rates of growth of the 90/50 and 50/10 income ratios.[7] This pattern is inconsistent with the hypothesis that rising income inequality is driving growth in the income achievement gap.

Nor is growth in the income achievement gap easily explained by the increasing correlation between income and parental education. The increasing correlation between income and parental education should have caused education to explain more of the income achievement gap over time. However, parental education explains less of the gap over time.[8]

Instead, the steady growth in the income achievement gap parallels—and is perhaps best explained by—the steady increase in emphasis on grades and testing over the past several decades, the rise of test-based accountability systems in education, and the increasing importance of test scores in defining educational success. Anxiety among students regarding report card grades and test scores has increased and is especially intense among students raised in middle-to-high-income families.[9] Students raised in middle-to-high-income households are increasingly susceptible to even a single low grade or disappointing test score. Students may be easily demoralized and depressed, affecting their performance. This is consistent with the observation that growth in the 90/10 income achievement gap has been driven by growth in the 90/50 income achievement gap.[10] The increased emphasis on grades and testing would be expected to exacerbate the effects of low grades and low test scores on student morale and increase the disparity in the morale of high-performing students raised in high-income families, compared to middle-performing students raised in middle-income families. Students from middle-income families who live in the same communities as high-income families would be expected to enter kindergarten performing below the level of classmates raised in high-income households, would be expected to receive lower grades and test scores compared to their high-income classmates, would be expected to suffer demoralization relative to their high-income classmates throughout their K-12 school careers, and would be expected to underperform their high-income classmates throughout their K-12 school careers.

Consistent with the hypothesis that children from low-income families enter kindergarten performing below their more-advantaged peers and are then systematically demoralized by existing grading and testing practices, an analysis by Roland Fryer and Steven Levitt concluded that the gap in achievement between black and white children that exists upon entry into kindergarten disappears after controlling for SES and a small number of other covariates.[11] The gap widens from that point forward.[12] This phenomenon is readily explained if black children are typically assigned to classes that include a majority of white children whose family income is typically higher than the family income of the black children. The raw data (without controls for SES) indicate that white children enter kindergarten performing at higher levels than black children. The black children are graded in relation to their classmates and presumably receive lower grades, on average, than their classmates, causing demoralization and depressed levels of achievement. The hypothesis is that the achievement of observationally equivalent black and white children with identical levels of SES diverges because the black children experience classes where the black level of SES is lower than the white level of SES. Observationally-equivalent black and white children with identical levels of SES have divergent experiences because those children do not experience the same classroom. The hypothesis is that divergence would not occur if black children experienced classes, over their entire school careers, where the black level of SES was the same as the white level of SES. Fryer and Levitt indicate that socioeconomic factors explain 85 percent of the black-white math gap and all of the black-white reading gap when children enter the school system in kindergarten but, in third grade, the same socioeconomic factors account for only 57 percent of the black-white math gap and 63 percent of the black-white reading gap.[13] Using the same Early Childhood Longitudinal Study data but a different version of the test score metric, an analysis by Murnane and his colleagues indicates that socioeconomic differences between black and white students account for the entire black-white math and reading gap in kindergarten, but only one-third of the math gap and 14 percent of the reading gap in third grade.[14] This pattern suggests that socioeconomic factors explain, in large part, the black-white differences in cognitive skills at the start of formal schooling, but do not account for the growth of the black-white gap as children progress through elementary school.[15] Sean Reardon and his colleagues concluded, "the black-white gap appears to widen during the school years—particularly in early elementary school—in ways that are not explained by socioeconomic family background characteristics, a pattern that suggests that schooling appears to contribute to the growth of the gaps."[16]

The pattern is consistent with the loss-of-control hypothesis. Socioeconomic factors contribute to differences in skills that exist when children begin kindergarten. These differences are then perpetuated throughout the K-12 years by grading and testing practices that demoralize low-performing students. In this view, factors other than socioeconomic differences become important after school entry. One factor—grading and testing practices—perpetuates initial differences in achievement. Other factors, such as differences among parents and teachers with regard to the degree to which they provide emotional support and encouragement to children, serve to buffer lucky children or undermine the unlucky ones. By third grade, these other influences begin to wash out the effects of socioeconomic differences that exist at entry into kindergarten. This may explain the decline in the influence of socioeconomic differences as children advance from grade to grade.

As children advance from grade to grade, the unexplained portion of the black-white gap in test scores grows larger and larger.[17] Measures of socioeconomic differences are increasingly inadequate to explain the gap. This pattern is inconsistent with theories that socioeconomic factors dominate and drive the persistence of the black-white gap throughout the K-12 years. The pattern is, instead, consistent with the hypothesis that the driving factor is embedded in the existing structure of schooling. The pattern is consistent with the hypothesis that grading and testing practices are at fault, and consistent with the hypothesis that the driving factor is the influence of those practices on student self-efficacy. It is significant that self-efficacy is typically not measured and not controlled in conventional models of student achievement. This may explain why the unexplained portion of the black-white gap in test scores grows larger and larger as children advance from grade to grade.

The regression analyses reported in Tables 4.1 and 4.2 include a measure of self-efficacy, unlike conventional models of achievement. The analyses indicate that self-efficacy predicted, and was predicted by, student achievement. Contemporaneous measures of self-efficacy predicted achievement in math (Models 2, 4, and 7) and reading (Models 12, 14, and 17). Lagged measures of self-efficacy predicted achievement in math (Models 5 and 8) and reading (Models 15 and 18), suggesting a causal influence of self-efficacy on achievement as much as two to three years later. Lagged test scores predicted self-efficacy in math (Models 9 and 10) and reading (Models 19 and 20), suggesting a causal influence of achievement on self-efficacy. Partial correlations (a measure of effect size) indicate sizable reciprocal effects of self-efficacy on math and reading achievement, as well as effects of math and reading achievement on self-efficacy, relative to the effects of race (see Appendix A).

Consistent with the hypothesis that low-performing students suffer downward spirals, Figure 4.5 indicates that the decline in black student self-efficacy is sharper than the decline in white student self-efficacy. Figure 4.6 indicates that the decline in Hispanic student self-efficacy is sharper than the decline in white student self-efficacy. Figure 4.7 indicates that the decline in white student self-efficacy is sharper than the decline in Asian student self-efficacy. Figure 4.8 indicates that the decline in self-efficacy is sharpest for low-performing students and Figure 4.9 indicates that the decline in engagement is sharpest for black and Hispanic students. Consistent with the hypothesis of downward spirals, differences in educational achievement widen as both black and white students progress from kindergarten through twelfth grade.[18]

Table 4.1 Models Predicting Math Achievement in Grades 3, 5, and 8 and Math Self-Efficacy in Grades 5 and 8

	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*	*(9)*	*(10)*
Outcome	*Gr3 Math*	*Gr3 Math*	*Gr5 Math*	*Gr5 Math*	*Gr5 Math*	*Gr8 Math*	*Gr8 Math*	*Gr8 Math*	*Gr5 Self-Efficacy*	*Gr8 Self-Efficacy*
Predictors										
AmIndian	–17.71**	–17.90**	–14.50*	–14.20*	–14.79**	–14.66**	–13.55**	–14.40*	0.15	0.02
	(5.56)	(5.37)	(5.68)	(6.20)	(5.50)	(5.53)	(5.77)	(5.97)	(0.14)	(0.15)
PacIsland	–12.76**	–12.03**	–7.11	–3.99	–6.58	–6.96	–5.15	–4.43	–0.26	–0.02
	(4.53)	(4.54)	(4.28)	(4.40)	(4.42)	(4.50)	(4.87)	(4.74)	(0.14)	(0.14)
black	–19.52***	–19.78***	–21.26***	–20.80***	–21.53***	–19.18***	–18.37***	–18.67***	0.15**	0.20**
	(1.29)	(1.33)	(1.54)	(1.68)	(1.59)	(1.50)	(1.44)	(1.60)	(0.05)	(0.06)
Hispanic	–12.47***	–12.75***	–10.79***	–10.40***	–11.11***	–9.25***	–8.24***	–8.94***	0.09*	0.00
	(1.25)	(1.26)	(1.34)	(1.27)	(1.33)	(1.10)	(1.05)	(1.06)	(0.03)	(0.05)
Asian	4.28	3.73	6.71**	6.15**	6.19**	5.63**	4.03*	5.37**	0.01	0.16*
	(2.18)	(2.15)	(2.05)	(1.83)	(2.01)	(1.92)	(1.64)	(1.74)	(0.06)	(0.07)
sex	–4.00***	–2.66**	–4.13***	–2.45**	–3.06**	–1.51	0.18	0.34	–0.17***	–0.13**
	(0.88)	(0.82)	(0.91)	(0.83)	(0.91)	(0.85)	(0.84)	(0.79)	(0.03)	(0.04)
math self-efficacy		5.02***		7.79***			6.38***			
		(0.45)		(0.55)			(0.41)			
math self-efficacy (lagged)					4.25***			6.44***		
					(0.50)			(0.52)		
math score (lagged)									0.01***	0.01***
									(0.00)	(0.00)
observations	8357	8352	8366	8364	8342	8354	8297	8337	8340	8299
degrees of freedom	443	443	443	443	443	443	443	443	443	443

Notes: "Gr3" is Grade 3; "Gr5" is Grade 5; "Gr8" is Grade 8; "Math" is math test scores; lagged values are prior wave values (grade 3 predictors for grade 5 outcomes and grade 5 predictors for grade 8 outcomes)

*p<.05. **p<.01. ***p<.001. Linearized standard errors in parentheses

Table 4.2 Models Predicting Reading Achievement in Grades 3, 5, and 8 and Reading Self-Efficacy in Grades 5 and 8

	(11)	*(12)*	*(13)*	*(14)*	*(15)*	*(16)*	*(17)*	*(18)*	*(19)*	*(20)*
Outcome	*Gr3 Read*	*Gr3 Read*	*Gr5 Read*	*Gr5 Read*	*Gr5 Read*	*Gr8 Read*	*Gr8 Read*	*Gr8 Read*	*Gr5 Self-Efficacy*	*Gr8 Self-Efficacy*
Predictors										
AmIndian	–23.66***	–24.47***	–19.50*	–18.75**	–20.04**	–18.08**	–15.15*	–17.40**	0.19	–0.19
	(5.59)	(5.47)	(7.50)	(7.46)	(7.32)	(6.44)	(6.04)	(6.44)	(0.11)	(0.10)
PacIsland	–9.74	–8.86	–7.07	–4.35	–6.47	–11.27	–6.43	–8.80	–0.20	–0.51**
	(6.88)	(5.88)	(6.55)	5.16)	(5.80)	(6.33)	(6.85)	(5.01)	(0.14)	(0.19)
black	–18.47***	–18.96***	–19.83***	–20.01***	–20.30***	–22.63***	–21.61***	–22.89***	0.21**	0.11
	(1.77)	(1.82)	(1.62)	(1.83)	(1.70)	(2.14)	(2.05)	(2.37)	(0.08)	(0.07)
Hispanic	–17.40***	–17.05***	–14.94***	–13.03***	–14.74***	–15.98***	–14.21***	–14.23***	–0.03	–0.09
	(1.35)	(1.33)	(1.30)	(1.25)	(1.27)	(1.54)	(1.44)	(1.47)	(0.04)	(0.05)
Asian	1.79	1.50	2.33	2.55	2.05	4.29*	4.01*	4.58**	–0.03	0.02
	(1.97)	(1.95)	(1.93)	(1.82)	(1.93)	(1.75)	(1.78)	(1.72)	(0.05)	(0.07)
sex	5.68***	5.06***	4.35***	3.53***	4.33***	6.98***	4.86***	5.66***	0.11**	0.24***
	(1.03)	(0.97)	(0.93)	(0.90)	(0.90)	(1.08)	(1.06)	(0.99)	(0.03)	(0.04)
reading self-efficacy		7.91***		9.12***			7.07***			
		(0.56)		(0.52)			(0.54)			
reading self-efficacy (lagged)					6.61***			8.45***		
					(0.52)			(0.56)		
reading score (lagged)									0.01***	0.01***
									(0.00)	(0.00)
observations	8311	8306	8359	8356	8335	8298	8207	8282	8295	8260
degrees of freedom	443	443	443	443	443	443	443	443	443	443

Notes: "Gr3" is Grade 3; "Gr5" is Grade 5; "Gr8" is Grade 8; "Read" is reading test scores; lagged values are prior wave values (grade 3 predictors for grade 5 outcomes and grade 5 predictors for grade 8 outcomes)

**p*<.05. ***p*<.01. ****p*<.001. Linearized standard errors in parentheses

In summary, the pattern of results is consistent with the loss of control hypothesis but is inconsistent with hypotheses that the decline in self-efficacy and engagement is caused by: a.) initial differences in student achievement, b.) differences in parenting or other sociocultural or socioeconomic differences that exist prior to the age when students enter the school system, c.) racial differences in achievement or resentful demoralization, or d.) differences in the quality of schools and teachers experienced by black and Hispanic students, compared to white and Asian students. The results indicate that the decline in self-efficacy is extraordinarily broad, affects all ethnic groups, and is accompanied by a parallel decline in student engagement. This is troubling and suggests that any solution to the achievement gap must address the demoralization that is occurring among almost all American students.

Grades and Demoralization

Regression analysis was used to test the hypothesized relationships between engagement, self-efficacy, and effort and achievement as measured by grade point average (GPA) using data from the National Education Longitudinal Study (NELS), a nationally-representative dataset involving 27,394 individuals who were surveyed as eighth-grade students in 1988, tenth-grade students in 1990, and twelfth-grade students in 1992 (see Table 4.3). The level of engagement in math courses, measured in grade eight, predicted GPA in grade nine math courses. The level of engagement in English courses, measured in grade eight, predicted GPA in grade nine English courses. The level of self-efficacy in math courses, measured in grade ten, predicted GPA in grade eleven and grade twelve math courses, and the effect size substantially exceeded the effect size for race. The level of self-efficacy in English courses, measured in grade ten, predicted GPA in grade eleven and grade twelve English courses, and the effect size substantially exceeded the effect size for race. GPA in grade nine math courses predicted self-efficacy in math courses, measured in grade ten, and the effect size was three times as large as the effect size for race. GPA in grade nine English courses predicted self-efficacy in English courses, measured in grade ten, and the effect size was seven times as large as the effect size for race. Partial correlations (a measure of effect size) are reported in Appendix B.

The breadth of the decline in self-efficacy and engagement from grade three to grade eight and the relationships among engagement, self-efficacy, effort and achievement extending from grade three to grade twelve are consistent with the loss of control hypothesis and consistent with the hypothesis that ubiquitous factors associated with the conventional model of schooling inadvertently demoralize and undermine the performance of all students, serving to reinforce, perpetuate, and magnify differences in achievement that exist at the time students enter the school system. In this view, demoralization affects all students but low-performing students are especially demoralized because they start at a lower level of achievement and receive signals throughout their academic careers (through letter grades) that their performances are substandard.

Table 4.3 Models Predicting GPA in Grades 9, 11, and 12 and Self-Efficacy in Grade 10

	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*
Outcome	*Grade 9 Math GPA*	*Grade 9 English GPA*	*Grade 11 Math GPA*	*Grade 11 English GPA*	*Grade 12 Math GPA*	*Grade 12 English GPA*	*Grade 10 Math Self-Efficacy*	*Grade 10 English Self-Efficacy*
Predictors								
sex	0.15***	0.30***	0.20***	0.32***	0.23***	0.35***	–0.38***	0.23**
	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.03)	(0.08)	(0.07)
Asian	0.18*	0.20*	0.03	0.19*	–0.04	0.14*	–0.00	0.07
	(0.07)	(0.09)	(0.10)	(0.10)	(0.07)	(0.06)	(0.17)	(0.15)
Hispanic	–0.31***	–0.29***	–0.26***	–0.34***	–0.29***	–0.34***	0.02	–0.06
	(0.06)	(0.05)	(0.05)	(0.06)	(0.07)	(0.06)	(0.11)	(0.10)
black	–0.46***	–0.55***	–0.41***	–0.47***	–0.43***	–0.43***	0.71***	0.38*
	(0.10)	(0.09)	(0.09)	(0.09)	(0.09)	(0.07)	(0.15)	(0.15)
AmIndian	–0.67***	–0.60***	–0.38**	–0.61***	–0.41	–0.37*	0.67	0.14
	(0.09)	(0.10)	(0.11)	(0.13)	(0.23)	(0.17)	(0.35)	(0.13)
Gr8 math engagement	0.10***							
	(0.02)							
Gr8 English engagement		0.08***						
		(0.02)						
Gr10 math self-efficacy			0.17***		0.14***			
			(0.01)		(0.01)			
Gr10 English self-efficacy				0.12***		0.11***		
				(0.01)		(0.01)		
Gr9 math GPA							0.62***	
							(0.04)	
Gr9 English GPA								0.51***
								(0.04)
observations	12222	12588	10904	12409	8959	12207	11936	12278
degrees of freedom	968	968	974	977	966	977	968	967

Notes: “GPA” is grade point average; “Gr8” is Grade 8; “Gr9” is Grade 9; “Gr10” is Grade 10

*p<.05. **p<.01. ***p<.001. Linearized standard errors in parentheses

The breadth and consistency of the decline in self-efficacy and engagement across race, ethnicity, and levels of student achievement are not easily explained by theories that the achievement of minority students is primarily depressed by the quality of their schools or teachers.

One might argue that both theories are correct: the conventional school system is structured in a way that erodes self-efficacy, engagement and achievement and, in addition, minority students attend schools that are lower in quality and are taught by teachers who are less effective, compared to more-advantaged students. However, the studies reviewed above suggest why school or teacher quality is a weak explanation for differences in student achievement. Furthermore, the evidence reviewed here suggests that it is not necessary to implicate school or teacher quality—fixed task difficulty and the existing system of letter-grade comparisons, in combination with initial differences in achievement, are sufficient to explain the observed persistence of the achievement gap from kindergarten through twelfth grade and offer a more parsimonious explanation.

Grades and Graduation

The hypothesis is that the conventional school system is inadvertently structured in a way that fosters disengagement, thereby reducing effort, which depresses achievement and grades, causing demoralization, which further reduces engagement and achievement. The previous sections presented evidence that demoralization and disengagement are occurring across ethnicity, across levels of achievement, and across grades from grade three to grade eight. While achievement as measured by test scores increases from grade three to grade eight, the hypothesis is that many—if not most—students are receiving what they view as low letter grades, and this depresses effort and future achievement. For high-achieving students, a letter grade of "B" may signal failure; for low-achieving students, letter grades of "C," "D," and "F" signal failure. Students who receive low grades are likely to feel that they have lost control of their grades. Likewise, low grades signal a loss of control (presumably, if students were able to control their grades they would earn high grades). Thus, the hypothesis is that grades are an important intermediate indicator of the underlying dysfunction in the conventional system of schooling.

One way to investigate this hypothesis is to examine the relationship between letter grades and the risk of dropping out or not graduating on time. For students who drop out, this outcome is arguably the end result of the process of demoralization and disengagement that begins in grade three. While it might seem obvious that low grades are associated with high rates of dropout, the hypothesis is that letter grades are not merely an indicator of current performance, but have a powerful psychological effect on students that drives future engagement, effort and achievement, and perpetuates the achievement gap. The hypothesis is that this is a causal mechanism that continually reinforces downward spirals in students regardless of race and ethnicity, dominating other influences, including preexisting differences in achievement when students enter school that are attributable to differences in parenting styles or socioeconomic and sociocultural differences. If letter grades are a better predictor of dropping out than ethnicity—and if this indicates the strength of a causal relationship running from grades to dropping out—the implication is that letter grades are not merely an indicator of performance but may be a causal factor that perpetuates the achievement gap.

Alex Bowers collected the entire teacher-assigned, subject-specific longitudinal grade histories, covering grades one through twelve, for the high school graduation class of 2006 in two Midwestern United States school districts.[19] The effects of multiple variables on a student's probability of dropout were estimated employing survival and discrete time hazard analysis and logistic regression with the person-period dataset. When letter grades were excluded from the statistical model, the risk of dropout was strongly influenced by ethnicity, consistent with the existence of the black-white achievement gap and consistent with Figures 4.5, 4.6, and 4.7: ethnicity → achievement/drop out.

However, when letter grades were introduced into the statistical model, the coefficient for ethnicity became insignificant, while the coefficient for letter grades was strongly significant. *This implies that all of the effects of ethnicity were explained—and presumably transmitted—by letter grades:*

$$\text{ethnicity} \rightarrow \text{letter grades} \rightarrow \text{achievement / dropout}$$

This supports the interpretation that initial racial differences in achievement when students enter school are reinforced and maintained through a process whereby low-achieving students become disengaged, their effort wanes, their grades suffer, they become demoralized by low grades, and their risk of dropout increases. This is consistent with the proposed mechanism explaining the persistence of the achievement gap but is not consistent with theories that focus on preexisting differences in parental styles and socioeconomic and sociocultural differences prior to school entry. Bowers' results indicate that disadvantaged minority students who achieve high letter grades demonstrate comparable levels of persistence and graduate at the same rates as white students.[20] This is consistent with evidence that disadvantaged minority students who achieve high letter grades perform comparably to their more-advantaged peers with regard to postsecondary outcomes.[21] These students are able to overcome any adverse differences in parental styles and socioeconomic and sociocultural influences. If differences in parenting styles and socioeconomic and sociocultural differences were decisive, then ethnicity should remain statistically significant despite the introduction of letter grades into the statistical model. White students should demonstrate better persistence and lower rates of dropout than black and Hispanic students even when holding letter grades constant. Bowers' results indicate that this is not the case.[22]

The magnitude of the letter-grade effect was large. For every one unit increase in letter grade, students were over six times more likely to graduate.[23] Eighty-six percent of students who received low grades did not graduate on time.[24] Bowers' results indicate that the risk of dropout begins in middle school, consistent with the results reported above indicating a broad, accelerating decline in student self-efficacy and engagement across all ethnic groups from grade three to grade eight. The median grade level for dropout for students with low grades was at grade eight.[25] The risk of dropout peaked in grade eleven.[26]

I replicated Bowers' results using data from the Education Longitudinal Study of 2002, a nationally-representative sample of 15,362 high school sophomores who were followed for a period extending two years beyond their expected high school graduation date.[27] Consistent with the existence of the black-white achievement gap and consistent with Figures 4.5, 4.6, and 4.7, the risk of dropout was strongly influenced by ethnicity and gender when letter grades were excluded from the statistical model. When letter grades were introduced into the statistical model, the coefficient was strongly significant; however, the coefficients for ethnicity and gender became insignificant. This implies that the effects of ethnicity and gender were explained, transmitted through, and dominated by letter grades. Significantly, a black male was less likely to drop out than a white or Asian female at every level of GPA (Table 4.4). This implies that the adverse effects associated with race can be eliminated by raising letter grades.

The letter-grade effect was strong. Over 80 percent of students who received low grades (GPA below 1.00) dropped out or experienced a dropout episode. Students who received low grades were almost twice as likely as students who received high grades (GPA above 3.51) to report that they left school because they did not like school. Consistent with the hypothesis that letter grades influence student engagement and self-efficacy, students who received high grades were more likely to agree that reading and math are fun, agree that they can do an excellent job on English and math assignments, and agree that they can avoid bad grades through their own efforts. Consistent with the hypothesis that exposure to the conventional system of schooling reduces student engagement, the Education Longitudinal Study data indicated that by the spring of their senior year, fewer than 22.5 percent of all students reported that they enjoyed school "a great deal."

Table 4.4 High School Dropout Rate, by GPA

GPA	*Black Male*	*White Female*	*Asian Female*
0.00-1.00	78.79	82.00	82.68
1.01-1.50	52.60	57.65	58.78
1.51-2.00	24.90	28.91	29.87
2.01-2.50	9.01	10.83	11.29
2.51-3.00	2.87	3.50	3.66
3.01-3.50	0.88	1.07	1.12
3.51-4.00	0.26	0.32	0.34

In summary, the picture that emerges is that student self-efficacy and engagement decline broadly starting in grade three. The decline accelerates from grade three to grade eight, depressing achievement and steadily increasing the risk of dropout starting in middle school. This picture is consistent with the hypothesis that the conventional school system is inadvertently structured in a way that demoralizes students. Significantly, all of the effects of ethnicity and the race-related achievement gap are explained by letter grades, consistent with the hypothesis that the system of letter grades inadvertently serves to transmit, perpetuate, and exacerbate initial ethnic differences in achievement that exist at entry into school, and may do so through powerful, unintended, negative psychological effects on children. Instead of being neutral indicators of performance, it appears that letter grades contribute to the erosion of students' self-efficacy and are akin to an extremely powerful pharmaceutical that has the unfortunate effect of demoralizing many children.

Bowers' results and the results from the Education Longitudinal Study predict that minority students who achieve high letter grades would graduate at the same rates as white students who achieve high letter grades, with few dropouts. If dropping out is the end result of the erosion of self-efficacy and engagement that begins by grade three, elimination of dropout necessarily implies that the broad erosion of self-efficacy and engagement must be reversed. The results reported here imply that this can only occur if all students achieve high letter grades. Thus, the question becomes: how can all students achieve high letter grades?

CHAPTER 5

A New Model of Learning

If a mismatch between task difficulty and individual student ability undermines engagement and achievement, then matching task difficulty to student ability should promote engagement and achievement. Numerous studies have controlled for individual differences in achievement motivation, experimentally manipulated task difficulty, and found that persistence improves when task difficulty is individualized.[1] Task difficulty influences task involvement and feelings of competence, which influence intrinsic motivation.[2] This evidence suggests that helpless, passive, disengaged student behavior may be addressed by individualizing task difficulty and performance expectations so that each student experiences a schedule of success and failure that establishes a strong contingency between effort and success.

For example, Kennelly, Dietz, and Benson randomly assigned matched students aged nine to eleven who were previously identified as "helpless" to three treatment groups.[3] Each group was assigned to solve arithmetic problems but the difficulty of the problems was varied across the groups, resulting in differential success rates: 100 percent, 76.9 percent, and 46.2 percent, respectively. The group achieving success 76.9 percent of the time demonstrated significantly better persistence in the face of failure.[4] Persistence, defined as unwillingness to stop performing a task, is central to the construct of motivation.[5] The optimal schedule involved individualization of task difficulty for each student on a daily basis so that success was achieved on 77 percent of trials but students experienced failure on a sufficient number of trials to experience a modest challenge to their competence. This schedule was an effective treatment for learned helplessness. However, since students may normally be expected to advance on a daily basis, individualization of task difficulty and performance expectations may be required on a daily basis.

It is significant that most schools are organized so that each student is presented with tasks that are appropriate for an average student at each age level, even if individual students are substantially above or below grade level.[6] As a consequence, the experience of schooling for students who are above grade

S.S. Yeh, *Solving the Achievement Gap*,
DOI 10.1057/978-1-137-58767-1_5

level involves tedious repetition of material that fails to challenge them, and the experience of schooling for students who are below grade level involves constant reminders that they are not up to par.[7] This lack of individualization potentially undermines engagement and achievement in a negative cycle that feeds on itself, as low engagement reduces achievement, which reduces engagement still further.[8] The effect is likely to be magnified for low-performing students and this may explain why the achievement gap grows as low-achieving students advance from grade to grade.

Rapid Performance Feedback

In addition to task difficulty, a second factor that influences engagement and achievement is the way that performance feedback is delivered. Currently, students are regularly tested, compared, and categorized as “above” or “below” grade level.[9] Students are subjected to the universal practice of grading, ranking, and comparing them to their same-age classmates through classroom assessments and standardized tests. In addition, the introduction of letter-grade report cards in middle school provides pointed signals when student performance is below the standard that is rewarded with an “A” letter grade.

The intent is to spur students to work harder when their performance falls below standards. However, from a psychological perspective, even very accomplished students may become discouraged by a letter grade of “B.” Less-accomplished students who receive “C”s, “D”s, and “F”s may become profoundly discouraged. The result is that almost all students may suffer a reduction in their self-efficacy and engagement in academic tasks as they are exposed to existing grading and ranking procedures, and the reduction may be expected to accelerate with increased exposure. This hypothesis is consistent with the observed decline in self-efficacy and engagement for almost all students from grade three to grade eight (Figures 4.8 and 4.9).

Individualization of task difficulty gives each student opportunities to be successful but students may continue to be discouraged if they continue to be graded with respect to the performance of other students. Thus, a second path through which learned helplessness and passive, disengaged behavior may be reversed is by providing rapid, objective performance feedback with regard to tasks whose difficulty is individualized for each student. Objective, positive performance feedback with regard to tasks that are at an appropriate level of difficulty tends to increase perceived competence and intrinsic motivation: as students’ perceptions of their abilities in an academic subject increase, not only do they try harder but they also enjoy the subject more.[10] Subjects who received positive performance feedback later displayed higher levels of intrinsic motivation compared with subjects who performed at the same rate of success, but did not receive the performance feedback.[11]

Positive effects of feedback on student engagement and achievement have been demonstrated in numerous studies dating back to the 1960s. For example, Smith, Brethower, and Cabot found that having students chart their progress

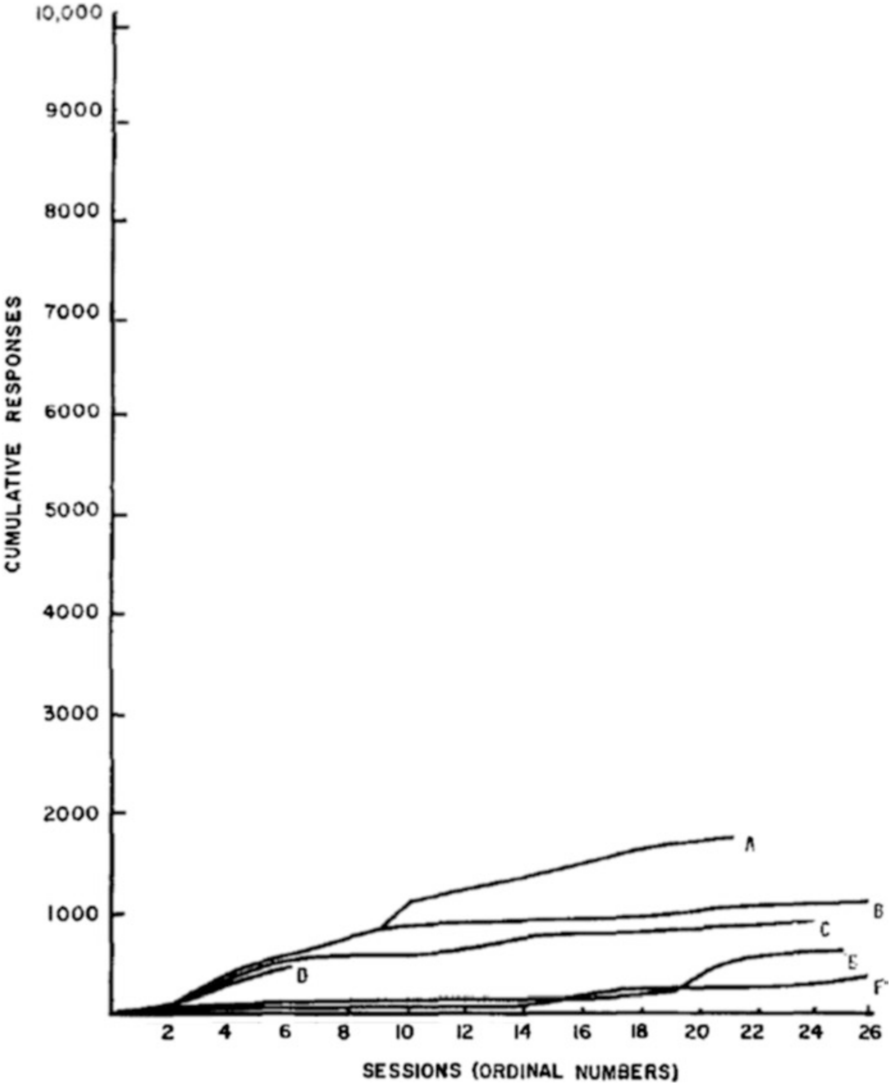

Figure 5.1 Language arts task output without performance feedback

significantly improved motivation and output.[12] Figure 5.1 shows student output, without feedback, on language arts tasks. Figure 5.2 shows that student output accelerated dramatically with feedback.

In a second study, Robinson, DePascale, and Roberts randomly assigned fifth- and sixth-grade students to two groups.[13] Both groups of students worked on identical sets of math problems in the same classroom at the same time with the same teacher. In the first session, neither group received feedback. In the second session, Group One received feedback, while Group Two

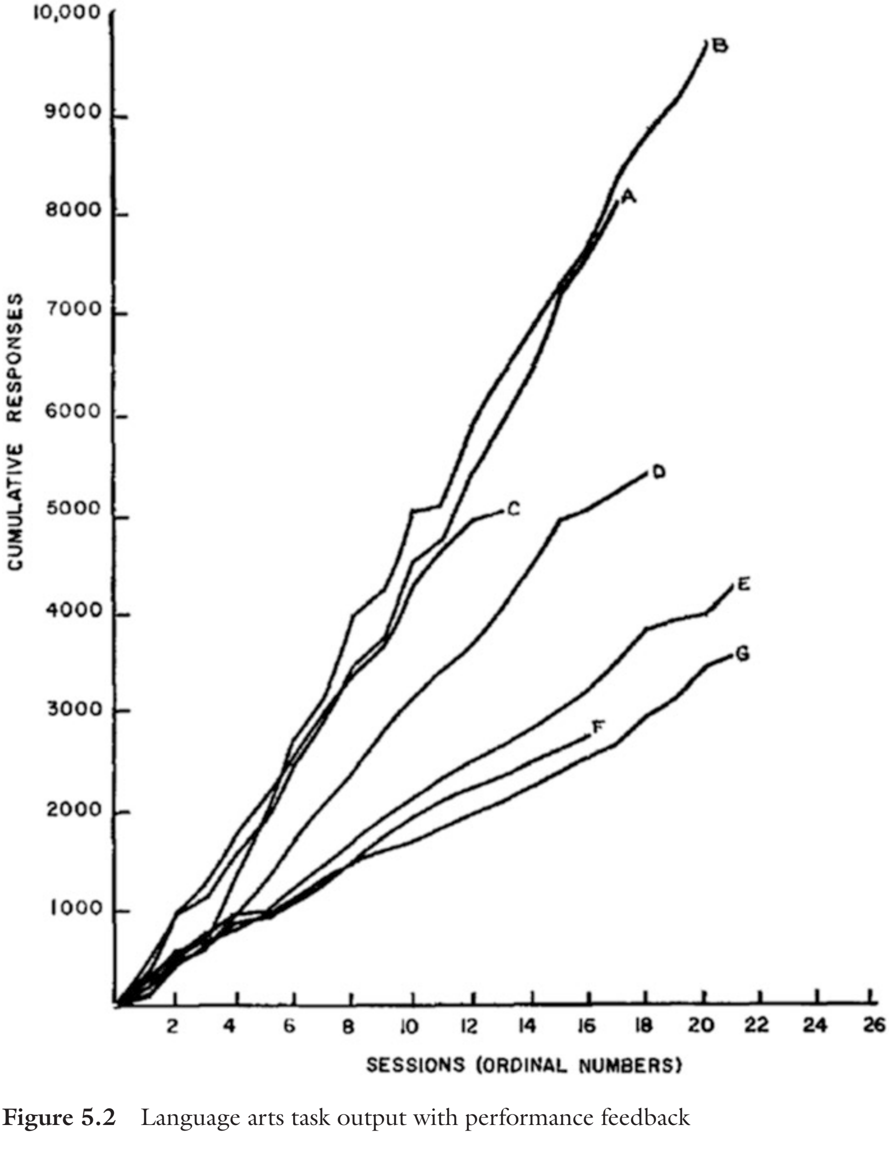

Figure 5.2 Language arts task output with performance feedback

did not. In the third session, both groups received feedback. In the fourth session, neither group received feedback. The results showed that whenever a group received feedback, students in that group completed more problems with greater accuracy, compared to the baseline condition. Whenever feedback was withdrawn, the completion and accuracy rates dropped. The design of this study virtually rules out any explanation other than the conclusion that feedback caused improved student engagement and achievement. It is difficult to attribute the results of this experiment to individual differences in student characteristics, teacher characteristics, classrooms, or schools. The research design controlled for those differences.

Several meta-analyses and reviews have been conducted regarding the effect of feedback on student achievement. A review of research summarized the results of previous meta-analyses regarding feedback and found an average effect size of 0.79 SD.[14] The meta-analyses included studies that experimentally compared the achievement of students who were frequently tested with a group of similar students who received the same curriculum but were not frequently tested. The results suggest that feedback is most effective when it is nonjudgmental, involving frequent testing (two- to five-times per week), and presented immediately after a test. Under these conditions, the meta-analyses and reviews of feedback interventions suggest that the effect size for testing feedback is no lower than 0.7 SD,[15] equivalent to raising the achievement of an average nation such as the United States to the level of the top five nations.[16] When teachers were required to follow rules about using the assessment information to change instruction for students, the average effect size exceeded 0.9 SD, and when students were reinforced with material tokens, in addition to the frequent testing, the average effect size increased even further, exceeding 1.1 SD.[17] Emotionally neutral (i.e., testing) feedback that is void of praise or criticism "is likely to yield impressive gains in performance, possibly exceeding 1 SD."[18] Thus, effective feedback includes feedback about how well a task is being accomplished or performed, such as distinguishing correct from incorrect answers.[19] This type of feedback is most common and is often called corrective feedback.[20] By itself, corrective feedback can be powerful. From various meta-analyses, Lysakowski and Walberg reported an effect size of 1.13, Walberg reported 0.82, and Tenenbaum and Goldring reported 0.74, all of which are substantial effects.[21]

The Conventional Model of Schooling

To summarize, the salient features of the conventional model of schooling are: 1.) a failure to individualize task difficulty and performance expectations so that each student experiences a modest challenge to his or her competence on a daily basis, and 2.) a failure to provide frequent, individualized, objective performance feedback. The salient pattern is a lack of student engagement and a gap in achievement that grows as students advance from grade to grade. These features of schooling, this pattern of disengagement, and the growth in the achievement gap as students move through the primary and secondary grades appear to be consistent with much of what is known about the conditions that can produce learned helplessness, passivity, lack of engagement, and low self-efficacy and the expected consequences of those conditions. Students who

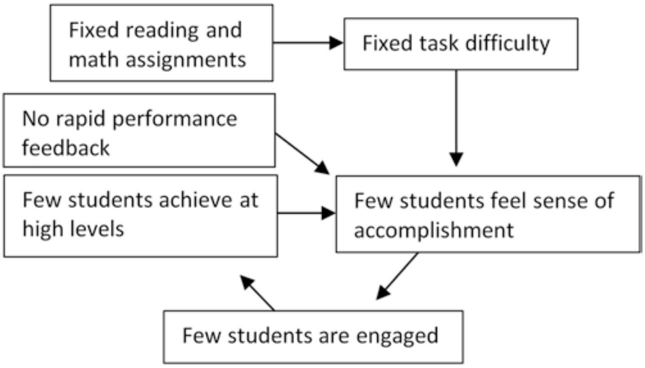
Fixed reading and math assignments
Fixed task difficulty
No rapid performance feedback
Few students achieve at high levels
Few students feel sense of accomplishment
Few students are engaged

Figure 5.3 Conventional model of schooling

exhibit low engagement and achievement also exhibit low self-efficacy, feelings of failure, and anxious, resentful, angry, discouraged, passive, apathetic, and depressed behavior.[22] These symptoms are consistent with the type of learned helplessness that may result when students do not have success in school.

Figure 5.3 diagrams the conventional model of schooling and its apparent consequences. In many instances, the same book (such as *Adventures of Huckleberry Finn*) may be assigned to the entire class. In other instances, students may be allowed to select books from the school library but the books in the library are not labeled and organized in a way that makes it easy for students to select books that are at an appropriate reading difficulty level. As a consequence, poor readers may frequently have the discouraging experience of selecting a book that is too difficult. Furthermore, there may be no system to provide rapid, objective performance feedback to students. Teachers may offer encouraging comments but students may sense that the feedback is not objective.

With regard to math, entire classes of students may be assigned the same set of math problems. Half of the students in the class may find the problems to be too difficult, while the other half may find the problems to be too easy. As a consequence, half of the students may feel discouraged, while the other half feel bored. There may be no system to provide rapid, objective performance feedback on a daily basis that would signal when progress has been achieved. In the conventional model, few students feel a sense of accomplishment. Few are excited, engaged, and enjoy learning. The achievement of both high- and low-performing students is undermined, but the impact is magnified for low-performing students because the impact of low grades is greater than the impact of boredom. In this model, small differences in achievement that exist when children enter school in kindergarten are magnified over time because the experience of schooling for low-performing children is relentlessly negative, undermining their sense of control, self-confidence, engagement, and effort.

RPF Model of Learning

The proposed model of the achievement gap is supported by the studies reviewed above suggesting that individualized task difficulty, in combination with frequent, objective performance feedback, serves to reverse the condition of learned helplessness, promoting engagement and achievement. If the model was incorrect, it seems unlikely that individualized task difficulty and frequent performance feedback would be effective.

To clarify this model of the achievement gap, the conventional model of schooling may be contrasted with what may be labeled a "rapid performance feedback" (RPF) model of instruction, learning, and achievement. The RPF model builds upon the studies reviewed above regarding the importance of perceived control over academic outcomes and implies that the achievement gap may be traced to dysfunction in the conventional educational system that inadvertently undermines the perceived control, engagement, and achievement of low-performing students.

The RPF model of instruction hypothesizes that an effective learning environment is individualized and structured so that each student is presented with tasks that are modestly challenging but not excessively difficult, where there is a high probability of success on a daily basis and a high probability that each student will receive positive performance feedback on a daily basis (Figure 5.4).

The model is illustrated by the *Reading Assessment* and *Math Assessment* programs (collectively labeled "Rapid Assessment" programs).[23] Students who use *Reading Assessment* select books to read from the school library according to each student's reading level. After reading a book, a student sits at the classroom computer to take a brief comprehension quiz that is specifically tailored to his or her book. Students who use *Math Assessment* receive sets of math problems that are tailored to each student's math level. After completing the day's set of math problems, the answers are electronically-scored with the help

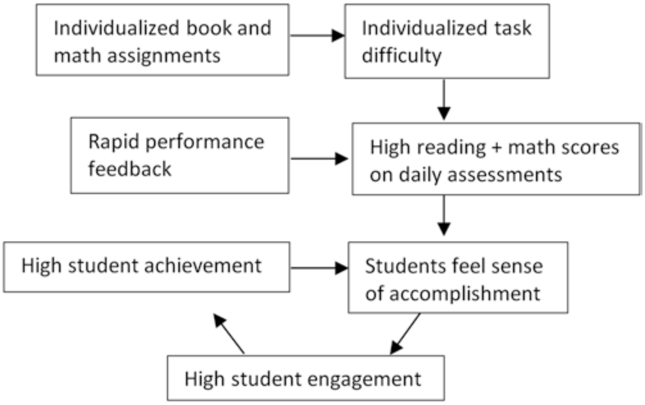
Individualized book and math assignments
Individualized task difficulty
Rapid performance feedback
High reading + math scores on daily assessments
High student achievement
Students feel sense of accomplishment
High student engagement

Figure 5.4 Rapid performance feedback model

of a mark scan device attached to the classroom computer. The RPF model hypothesizes that each student would be able to receive high comprehension scores on end-of-book quizzes, and would be able to complete his or her daily math problems with high accuracy scores, because the books and math problems are individually-assigned so that task difficulty is aligned with each student's current level on a daily basis.

This hypothesis was supported by two large studies. A national study involving 2,202 students examined student accuracy on *Math Assessment* practice items and end-of-unit test items: average accuracy on practice items was 80 percent, while average accuracy on end-of-unit test items was 87 percent.[24] A randomized study involving 1,665 Memphis students found that 80 percent of their teachers reported that their students averaged between 85 and 92 percent correct on end-of-book *Reading Assessment* comprehension quizzes.[25] These results suggest that students exposed to *Math Assessment* and *Reading Assessment* regularly received objective reports signaling mastery of reading and math.

The RPF model suggests that a learning environment where each child regularly receives objective, positive feedback signaling that he or she is advancing on a daily basis results in improved engagement. Consistent with this hypothesis, teachers who employ the RPF model report that students feel a sense of accomplishment, confidence, mastery and control, and are highly motivated in this environment because the students feel that they are successfully completing tasks that are modestly challenging but not overwhelming.[26] Teachers report that students enjoy school work, enjoy learning, and exert more effort, which increases their achievement and improves their self-esteem and engagement even further, in a virtuous cycle.[27]

Impact Studies of *Reading Assessment* and *Math Assessment*

If the model of the achievement gap that has been presented is incorrect it is unlikely that programs based on that model and designed to address the key factors identified by that model would be effective. Therefore, the proposed model of the achievement gap may be tested by evaluating evidence regarding the effectiveness of the two programs that are based on the model and designed to address the key factors identified by that model. *Reading Assessment* and *Math Assessment* are programs that individualize task difficulty for each student and provide rapid performance feedback. Two randomized experiments evaluated the effectiveness of the *Reading Assessment* rapid performance feedback program with regard to reading achievement.[28] The first experiment, involving 1,665 Memphis students (a district where 71 percent of all students are eligible for free/reduced price lunch), found an average effect size of 0.270 SD per grade in grades K through 6 over a nine-month school year.[29] The second experiment involved 978 students (89.9 percent African

American and 83 percent eligible for free/reduced price lunch), employed hierarchical linear modeling (HLM), and found an average effect size of 0.175 SD per grade in grades three through six over a nine-month school year.[30] In both studies, the effect sizes were achieved with an unusually disadvantaged population of students.

A randomized study of the *Math Assessment* rapid performance feedback program, involving 1,880 students in grades two through eight in 80 classrooms and seven states, found an effect size of 0.324 SD on math achievement over a seven-month period after controlling for treatment integrity.[31] A national, peer-reviewed quasi-experimental evaluation of *Math Assessment*, involving 2,202 students in grades three through ten in 125 classrooms in 24 states, found that students in the treatment group gained an average of 0.392 SD per grade over one semester (18 weeks), compared to students not receiving *Math Assessment* (at pretest the scores of treatment and comparison students were not significantly different).[32]

League Table Efficiency

A systematic league table comparison of the leading interventions for raising student achievement offers a test that compares competing theories about the factors that limit student achievement. For example, if the limiting factor is sociocultural differences in parenting practices, then an intervention such as high-quality preschool should provide a stronger boost than RPF, which merely individualizes task difficulty and provides performance feedback. Similarly, if the limiting factor is a lack of accountability, then an intervention that strengthens accountability should provide a stronger boost. If the limiting factor is a lack of choice and competition, then interventions such as voucher programs and charter schools that improve choice and competition should provide a stronger boost. If the limiting factor is low teacher quality, then interventions that boost teacher quality should provide a stronger boost to achievement. If the proposed model of the achievement gap is incorrect it would be unlikely that programs based on that model and designed to address the key factors identified by that model would top the league table.

It is necessary to ensure that the analysis is systematic and employs an appropriate outcome measure. League tables are typically constructed using student achievement effect size as the outcome measure but this can be misleading because it does not incorporate information about the social resources required to achieve a given effect size. It is preferable to use effectiveness-cost ratios, where each ratio is defined as the effect size for a given intervention divided by the cost per student to achieve that effect size.

To ensure that the analysis is systematic, it is important to annualize the effect sizes, include all social costs, and employ a single consistent set of assumptions

Table 5.1 Comparison of Effect Sizes, Costs, and Effectiveness-Cost Ratios for Interventions to Raise Student Achievement

	Effect Size (SD)		*Cost*	*Effectiveness-Cost Ratio*	
	Reading	*Math*		*Reading*	*Math*
Rapid Assessment	0.270		$9.45	0.028571	
(high estimates)		0.392	$18.89		0.020752
Rapid Assessment	0.175		$9.45	0.018519	
(low estimates)		0.324	$18.89		0.017152
Comprehensive school reform	0.510	0.510	$217.83	0.002341	0.002341
Cross-age tutoring	0.480	0.970	$555.61	0.000864	0.001746
Computer-assisted instruction	0.230	0.120	$238.12	0.000966	0.000504
Longer school day	0.070	0.030	$159.87	0.000438	0.000188
Teacher education	0.220	0.220	$702.62	0.000313	0.000313
Teacher experience	0.180	0.180	$702.62	0.000256	0.000256
Teacher salary	0.160	0.160	$702.62	0.000228	0.000228
Summer school	0.190	0.190	$1,515.00	0.000125	0.000125
Rigorous math classes	--	0.200	$1,911.17	--	0.000105
Value-added teacher assessment	0.057	0.057	$624.72	0.000091	0.000091
Class Size Reduction					
Nye et al. (2001)	0.104	0.090	$1,379.28	0.000075	0.000065
Finn et al. (2001)	0.120	0.129	$1,379.28	0.000087	0.000094
10 percent increase in spending	0.083	0.083	$1,118.83	0.000075	0.000075
Full-day kindergarten	0.181	0.181	$2,611.00	0.000069	0.000069
Head Start	0.324	0.165	$9,000.00	0.000036	0.000018
High standards exit exam	0.051	−0.062	$2,025.97	0.000025	−0.000031
NBPTS teacher certification	0.002	0.004	$326.53	0.000006	0.000012
Higher licensure test scores	0.004	0.015	$894.10	0.000004	0.000017
Perry Preschool	0.150	0.155	$12,147.03	0.000012	0.000013
Abecedarian Preschool	0.150	0.054	$10,188.09	0.000015	0.000005
Additional school year	0.150	0.150	$14,271.76	0.000011	0.000011
Voucher programs	0.032	0.081	$9,646.01	0.000003	0.000008
Charter schools	0.009	0.001	$8,086.30	0.000001	0.000000

Note: Table 5.1 is adapted from Yeh, S. S. (2010). The cost-effectiveness of 22 approaches for raising student achievement. *Journal of Education Finance, 36*(1), 38–75

to generate the estimates. This analysis has been completed for 22 alternative approaches for raising student achievement (Table 5.1). The effectiveness-cost ratios may be translated into a bar chart that graphically illustrates the large differences in the ratios (Figure 5.5).

The magnitude of the differences is primarily a function of the large differences in the cost to implement each intervention. RPF, embodied in the Rapid Assessment programs, is primarily implemented using computer software, the cost of which may be spread over hundreds of students per building and amortized over multiple years, resulting in a low annual cost per student. Alternative approaches for raising student achievement typically involve much larger annual expenditures which reduce the corresponding effectiveness-cost ratios.

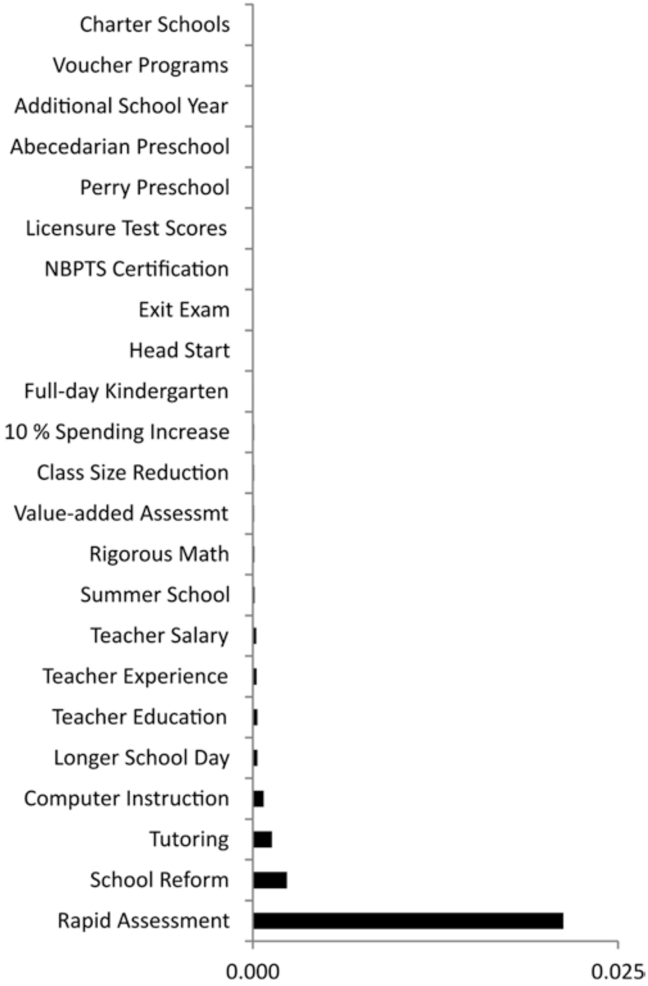

Figure 5.5 League table efficiency ratios for 23 interventions to raise achievement

Figure 5.5 indicates that RPF is more efficient than 22 alternative approaches for raising student achievement. In contrast to comparisons that may indicate relatively small differences in student achievement effect sizes, Figure 5.5 indicates the existence of very large differences in effectiveness-cost ratios.

A Misplaced Focus

The league table results suggest that the single most important factor limiting student achievement is the lack of a system for individualizing task difficulty and providing rapid performance feedback, rather than sociocultural factors, or a lack of accountability, or a lack of choice and competition, or low teacher quality. This result points away from currently-accepted views about the source of the achievement gap and effective treatments for low achievement. The proposed model of the gap draws upon and integrates a diverse set of literature that has generally been overlooked by researchers seeking to explain the gap: studies regarding a student's perceived control over academic outcomes, studies regarding learned helplessness, studies suggesting that individualization of task difficulty and rapid performance feedback are effective in reversing learned helplessness, studies regarding the RPF (Rapid Assessment) programs and the cost-effectiveness studies underlying Table 5.1 and Figure 5.5.

The primary alternatives to RPF do not seek to address the dysfunctional task structure and lack of performance feedback that is embedded in the conventional model of schooling. Thus, the alternatives do not seek to address structural factors that create a discouraging experience of schooling for students who are below-average and a tedious experience of schooling for students who are above-average. From this perspective, the league table results may not be surprising. The alternative interventions may address important issues but the league table results suggest that those issues are less significant than the rigid task structure and lack of performance feedback that characterize the conventional model of schooling.

The hypothesis that rigid task structure and lack of performance feedback are the key factors driving the persistence of the achievement gap does not imply that no other factors are important. The influence of factors other than task structure and degree of performance feedback may explain, for example, why economically-disadvantaged Asian students tend to outperform economically-disadvantaged black and Hispanic students; parental style and the cultural emphasis on education may play a role. The injection of competitive pressure through charter schools and voucher programs might improve outcomes. Similarly, an increase in accountability might have a positive effect. However, the league table results indicate that it may be more efficient to invest society's scarce resources into rapid assessment systems that individualize task difficulty and provide rapid performance feedback to students and teachers.

The league table results suggest, contrary to popular views, that teacher quality is not the most important factor limiting student achievement. How can these results be reconciled with previous research suggesting that the contribution of teachers to student achievement is large and value-added estimates of teacher contributions predict their students' measured achievement?[33]

First, it appears that many researchers have not recognized the instability of teacher rankings based on value-added estimates of performance. Rankings are unreliable whether they are based on a single year's ranking,[34] two or three years of rankings,[35] or four or five years of rankings.[36] In short, value-added methods do not permit the reliable identification of high- and low-performing teachers.

Second, the estimate of a teacher's contribution predicts the prior performances of his or her students.[37] Since it is impossible for a teacher to cause the prior performance of his or her students, this result implies there is nonrandom selection of students into teacher classrooms that is not controlled through the inclusion of time-invariant student characteristics.[38] Therefore, the central assumption underlying value-added models appears to be invalid.[39]

These results suggest that researchers who advocate the use of value-added modeling to identify high- or low-quality teachers have not fully recognized the limitations of value-added methods. The instability in teacher rankings suggests that it is misleading to assert that a student who has a high-quality teacher for three years in a row would greatly benefit. The reason is that a high-quality teacher this year is not likely to remain a high-quality teacher next year, if it is indeed the case that teacher rankings are highly unstable. Value-added methods cannot be relied upon to identify teachers who cause their students to achieve at high levels.

The analysis offered in this chapter suggests a different, potentially more profitable, way of understanding and addressing the achievement gap. This view suggests why the current focus on teacher quality may be misplaced and why another approach may be more productive. According to this view, teachers have been laboring in a system that is inadvertently structured in a way that undermines the sense of control, self-efficacy, engagement, and achievement of low-achieving students. This view suggests why it may be more productive to equip all teachers with the technology and support required to individualize task difficulty and provide rapid performance feedback, rather than to rely upon highly unreliable methods of identifying and replacing a relatively small number of teachers who may be low-performing this year but are equally likely to be higher-performing teachers next year.[40]

CHAPTER 6

Contradictions Resolved

Chapter Six compares two alternative models that seek to explain the persistence of the gap in academic achievement between minority students and their white peers. One model explains the persistence of the gap based on data suggesting that minority students tend to experience lower-quality schools and teachers, compared to their white peers. The second model explains the persistence of the gap as a psychological phenomenon, as described in Chapter Three. According to this model, minority students tend to become demoralized because they enter kindergarten performing below their same-age classmates and from that point forward receive comments, classroom test scores, grades, and other cues that trigger and reinforce negative self-images, undermining effort and achievement throughout their school careers.

Chapter Six advances an unorthodox argument: after parsing the available evidence regarding the reliability and validity of value-added teacher rankings, and after reviewing evidence that National Board for Professional Teaching Standards (NBPTS) teacher certification is a reliable measure of teacher quality but a weak predictor of gains in student performance, there is reason to question the prevailing view that the contribution of teachers to student performance is the largest factor influencing student achievement. Chapter Six begins by reviewing studies regarding the use of value-added modeling (VAM) to assess teacher and school quality. As described in Chapter Five, VAM has received attention as a promising approach for judging quality. The apparent relationship between value-added teacher rankings and gains in student performance provides the foundation for the argument that the contribution of teachers to student performance is the largest factor influencing student achievement. However, several studies raise questions about the reliability and validity of VAM. Chapter Six reviews this literature and offers a theoretical explanation for these contradictory results. This explanation is linked to the model of the achievement gap that explains the persistence of the achievement gap as a psychological phenomenon, rather than a problem of teacher or school quality. The two models are then tested

S.S. Yeh, *Solving the Achievement Gap*,
DOI 10.1057/978-1-137-58767-1_6

and compared using path analysis with three longitudinal datasets involving nationally-representative samples of schools and students.

Teacher Quality

A current theory regarding the persistence of the achievement gap is that minority students tend to experience lower-quality schools and teachers, compared to their white peers. In the United States, sizable race-related wealth inequalities persist. Black individuals generally have a smaller stock of accumulated wealth to bestow upon their offspring, thereby perpetuating the black-white gap in net worth.[1] Low wealth limits the prices of the homes that black families can afford. This translates into differences in the perceived quality of the neighborhoods and associated schools experienced by blacks versus whites. Whites and individuals with more education and income are disproportionately likely to reside in neighborhoods on the "high-quality" side of school boundaries, where homes command school-related premiums.[2] Blacks and individuals with less education and income are disproportionately likely to live in neighborhoods on the "low-quality" side of school boundaries, where home values are depressed by perceptions of low school quality.[3] Whites live in neighborhoods where public school test scores are higher, on average, by 15 percent, compared to neighborhoods where the average black family lives.[4] In sum, there is evidence of racial segregation that translates into differences in the quality of schools experienced by blacks versus whites.

Significantly, studies suggest that the contribution of teachers to student achievement is large and value-added estimates of teacher contributions predict their students' measured achievement.[5] If it is true that minority students experience lower-quality teachers and schools, and if teachers significantly influence student performance, this might explain the gap in achievement versus white peers.

Flawed Measures

While value-added modeling of student achievement is increasingly being adopted by school districts across the nation, growing evidence suggests that statistical estimates of teacher contributions are flawed. Teacher rankings based on value-added estimates of performance are unreliable measures of future performance. Six studies have investigated the predictive power of teacher rankings based on value-added measures.[6] In each study, teachers were ranked from high to low during a base period. In some studies, only one year of data was used to create the ranking.[7] In some studies, two or three years of data were used.[8] In one study, four to five years of ranking data were used.[9] In some studies, the focus was on the top and bottom quartiles of teachers. In other studies, the focus was on the top and bottom quintiles. However, in every study, the rankings were unreliable in predicting future performance. In all but one instance (top quartile teachers in the Aaronson et al. study), over half of the teachers ranked in the top and bottom quartiles (or quintiles) during the base period did not remain in those categories during the subsequent year.[10] While the value-added measure predicted that teachers in the top and bottom quartiles (or quintiles) would remain in those categories, over half of the teachers shifted out of those categories during the subsequent year.[11] In short, value-added methods do not permit the reliable identification of high- and low-performing teachers.

A second problem is that the estimate of a teacher's value-added contribution predicts the prior performances of his or her students.[12] For the purpose of predicting their students' fourth-grade scores, the value-added scores for fifth-grade teachers are nearly as strong a predictor as the value-added scores for their students' fourth-grade teachers.[13] However, since it is impossible for a teacher to cause the prior performance of his or her students, this result implies there is nonrandom selection of students into teacher classrooms that is not controlled through the inclusion of time-invariant student characteristics.[14] Therefore, the central assumption underlying value-added modeling appears to be invalid.[15]

These results suggest that the use of value-added measures to rank teacher quality is not warranted. The instability in teacher rankings suggests that it is misleading to assert that a student who has a high-quality teacher for three years in a row would greatly benefit. The reason is that a high-quality teacher this year is not likely to remain a high-quality teacher next year, if it is indeed the case that teacher rankings are highly unstable. Value-added methods cannot be relied upon to identify teachers who cause their students to achieve at high levels.

In sum, the value-added modeling (VAM) methods that are employed to calculate the contribution of each teacher to student achievement depend on assumptions that appear to be invalid. The evidence that these methods are unreliable in predicting future teacher performance suggests that key variables influencing student performance have been omitted from the statistical models. The evidence that the estimate of a teacher's value-added contribution predicts the prior performances of his or her students indicates that the central assumption underlying value-added modeling is invalid.

Omitted Variables

McCaffrey et al. explain that VAM estimates the contribution of each teacher each year by calculating, across all students served by the teacher, the average gain in student performance from the previous year after controlling for all other variables that are included in the VAM model.[16] The teacher's contribution is calculated indirectly, as the residual gain that remains after the influences of all other variables that are included in the VAM model are controlled. A problem arises if key variables are inadvertently omitted from the VAM model. The influences of the omitted variables are inadvertently lumped together in the calculation of the residual gain and, thus, are inextricably conflated with, and indistinguishable from, the value of the teacher's contribution calculated through VAM. The procedure that is used to calculate the teacher's estimated contribution mixes the teacher's actual contribution and the influences of all other variables that were inadvertently omitted from the model:

> The persistent teacher effect is simply the portion of the estimated effect that is common across years. It is not necessarily equal to the teacher's true performance; estimated effects from VAMs might not equal true causal effects of teachers due to violations of the model assumptions,[17] and even persistent components of the estimated effects might include confounding factors that endure over time. For example, if the achievement model fails to properly capture all unobservables that are correlated with classroom assignments and the classroom average of the unobservables is stable across years, these omitted variables will be part of the persistent teacher effect. In an extreme case of confounding, suppose annual teacher effects were measured by classroom average test scores without any adjustment for student heterogeneity. These effects would likely demonstrate strong persistence within teacher over time due to the stability in the types of students assigned to teachers across years.[18]

Suppose, for example, that children who enter kindergarten performing below their same-age classmates consistently receive signals through grades, test scores, and teacher comments throughout their academic careers indicating that the children are performing below-average, and suppose that this steady diet of negative feedback depresses the children's levels of self-efficacy, engagement, effort and achievement throughout their school careers. Depressed levels of self-efficacy would be an example of the type of unmeasured, unobserved characteristic described by McCaffrey et al. whose effects would be conflated with the estimated contribution of each teacher and would cause each teacher's contribution to be incorrectly estimated when using value-added statistical modeling. Teachers who are assigned to teach classes with high proportions of low-income minority students would be likely to receive, year after year, students whose levels of self-efficacy and learning potential are depressed relative to the average student in the school district, causing value-added estimates of the teachers' contributions to be artificially depressed. Teachers who are assigned to teach classes with low proportions of low-income minority students would be likely to receive, year after year, students whose levels of self-efficacy and learning potential are elevated relative to the average student in the school district, causing value-added estimates of the teachers' contributions to be artificially elevated.

Artificial Effects

As a consequence, it may be expected that a portion of *all* teachers would, by chance, periodically receive entire classrooms filled with high-self-efficacy, high academic potential students that would artificially boost the value-added estimates of the contributions of those teachers. For any given teacher, this might happen to occur over a one-, two-, or three-year period and the effect would be reflected in any VAM research study lasting one, two, or three years. However, this process would be unrelated to each teacher's skill and ability. The identities of the teachers whose students exhibit these gains would continually change. However, these fortuitous alignments of entire classrooms filled with high-self-efficacy students would mislead researchers into thinking that some teachers make large contributions to student achievement.

To draw an analogy, strong winds whip up unusually large waves. It would be wrong to conclude that these waves are "high-quality" waves that intentionally direct their skills to achieve impressive effects. Similarly, we might expect that any given sample of teachers would include, by chance, a number of teachers who happened to receive an unusually high proportion of high-self-efficacy, high academic potential students for a one-, two-, or three-year period of time. Since the effect that is attributable to self-efficacy is not measured and not controlled in the VAM model, the value-added procedure for calculating each teacher's contribution is unable to disentangle the self-efficacy effect from each teacher's actual contribution. The two effects would be lumped together when calculating the residual gain and would be indistinguishable. This would explain not only why some teachers exhibit large estimated value-added contributions to student achievement over one-, two-, or three-year periods of time, but why those estimates are poor predictors of future performance. The specific mix of high- and low-self-efficacy students received by each teacher each year is akin to fluctuation in winds that vary beyond the control of particular waves. In this view, the mix of students received by each teacher each year, rather than deliberate actions by individual teachers, drives the value-added statistical estimates of each teacher's contribution. Consistent with this view, the only study that investigated the stability of value-added estimates of teacher performance over a 10-year period found that the estimated performance of individual teachers fluctuates significantly over time due to unobserved factors that are currently not captured through VAM.[19] This invalidates the assumption of stable teacher performance that is embedded in key studies regarding VAM.[20]

Inadequate Controls

The inclusion of race and poverty covariates fails to fully adjust the value-added statistical calculations for differences in levels of self-efficacy (and differences in levels of potential achievement associated with differences in levels of self-efficacy) because race and poverty covariates measure race and poverty, not self-efficacy. For example, some minority students possess high levels of self-efficacy, while other minority students possess low levels of self-efficacy. Some minority students possess high levels of potential achievement, while other minority students possess lower levels of potential achievement. A value-added model might include race covariates, but this would not fully adjust the value-added calculation of a teacher's contribution to student achievement if all students taught by the teacher possess low levels of self-efficacy. A teacher may receive an entire class filled with low-self-efficacy black and Hispanic students, but the black and Hispanic covariate indicators underadjust for the effect of low self-efficacy because they adjust performance for a mixed group of high- and low-self-efficacy students rather than an entire group of low-self-efficacy students. Similarly, free-lunch eligibility status is an indicator that is used in value-added modeling to control for level of poverty, but it is an imprecise measure of self-efficacy. Many students who are not eligible for free-lunch programs may possess low levels of self-efficacy. A teacher may be assigned to teach an entire class of students who happen to have low levels of self-efficacy, but only some students may qualify for free lunch. The inclusion of free-lunch status as a covariate in the value-added statistical model fails to fully control for the depressed level of potential achievement that may be expected with entire classes of students who possess low levels of self-efficacy.

A similar issue arises if the focus is switched to school quality, measured by average school test scores. If large numbers of low-income minority children enter kindergarten performing below their same-age classmates and consistently receive signals through grades, test scores, and teacher comments throughout their academic careers indicating that the children are performing below-average, and if this steady diet of negative feedback depresses the children's levels of self-efficacy, engagement, effort, and potential achievement throughout their school careers, then it may be expected that certain schools in low-income neighborhoods would be filled every year with children whose levels of self-efficacy and potential achievement are depressed—not because of the quality of the schools, but instead because universal grading and testing practices systematically undermine the self-efficacy, effort, and potential achievement of every child who enters a school performing below grade level.

The use of race and free-lunch status as covariates in value-added statistical models would fail to fully adjust the models for the same reasons explained above. Both race and the indicator of free-lunch eligibility are flawed measures of self-efficacy. They are flawed measures of the degree to which potential achievement is depressed whenever children enter a school performing below grade level and are continuously subjected to grading and testing practices that systematically depress self-efficacy and potential achievement.

The use of prior student achievement as a covariate in value-added statistical models would be inadequate to fully adjust the models because the harm that is produced by existing grading and testing practices would be expected to accumulate throughout the career of each low-performing student at each school that he or she attends. The hypothesized effect would be perfectly confounded with attendance at every school and would be impossible to separate from the independent contribution of each school.

A Theoretical Explanation

The hypothesis that harmful effects from existing grading and testing practices persist and accumulate throughout the career of each low-performing student at each school that he or she attends provides a theoretical explanation for the finding that the estimate of a teacher's value-added contribution predicts the prior performances of his or her students.[21] If existing grading and testing practices exert a steady, cumulative effect throughout each student's career, and if teachers tend to be assigned the same types of students every year, certain unfortunate teachers will be systematically assigned students whose academic potentials have been undermined and continue to be undermined. The low performance of these students would necessarily be correlated with their prior performances. Since VAM omits a control for the level of self-efficacy and the VAM procedure lumps the influence of this factor together with the teacher's contribution into the calculation of the residual gain and then labels the entire residual gain as the teacher's contribution, what VAM labels as the teacher's contribution would be contaminated by the influence of the self-efficacy variable, creating the observed correlation between the residual gain score that is calculated for each student (i.e., what VAM labels the teacher's estimated contribution to student performance) and each student's prior performance. The evidence that contamination has occurred lies in the fact that the residual gain scores are correlated with the level of prior student performance. The residual gain scores must be contaminated by factors other than teacher contributions because teachers cannot cause the performance of their students during the period prior to the point where teachers receive the students.

To recapitulate, VAM teacher rankings predict future student performance. But they also predict past student performance, which implies that the correlation is not an indicator of causation but is instead an indicator of some third factor that influences student performance as well as the VAM rankings. It is difficult to identify a suitable factor that is a.) strongly correlated with student achievement and b.) not controlled in conventional value-added models of achievement. Self-efficacy is a likely candidate. It is strongly correlated with (and influences) student achievement,[22] yet it is not measured and controlled in conventional value-added models of achievement. Therefore, its influence is not controlled. Instead, any persistent influence attributable to self-efficacy is lumped together with each teacher's independent contribution to student achievement and the combined effect is attributed to the teacher. It is difficult to think of another factor that is strongly correlated with student achievement and is not controlled in conventional value-added models of achievement.

Checking the Assumptions

All of this suggests a need to reexamine the fundamental assumption that teachers exert strong influences on student performance. A strategy to check the premise that teachers exert strong influences on student performance is to investigate the best alternative measure of teacher quality that is independent of value-added statistical measures of teacher quality. If the best alternative measure of teacher quality is a strong predictor of student performance, this would support the theory that teachers make strong contributions to student performance. However, if the best available measures of teacher quality are weak predictors of student performance, this would undermine the theory that teachers make strong contributions to student performance.

Perhaps the best alternative measure of teacher quality is certification by the National Board for Professional Teaching Standards (NBPTS). NBPTS is an independent organization established in 1987 with the goal of advancing the quality of teaching and learning.[23] NBPTS developed professional standards for teaching, then contracted with the Educational Testing Service (and, later, Pearson Educational Measurement) to create a voluntary system to certify teachers who meet those standards.[24]

NBPTS certification is a lengthy, highly demanding process. Applicants for certification are required to submit a portfolio to NBPTS involving four entries.[25] Three are classroom based, where video recordings of teacher-student interaction and examples of student work serve as supporting documentation. A fourth entry relates to the candidate's accomplishments outside of the classroom—with families, the community, or colleagues—and how they impact student learning. Each entry requires some direct evidence of teaching or school counseling as well as a commentary describing, analyzing, and reflecting on this evidence. Following submission of the portfolio, candidates are tested on their content knowledge through six 30-minute exercises, specific to the candidate's chosen certificate area, at one of 300 NBPTS computer-based testing centers across the United States. Applicants are scored on a scale of 75 to 425, incorporating both the portfolio and the assessment center exercises, and they must earn a score of at least 275 to achieve certification.[26]

Drew Gitomer evaluated the interrater reliability of NBPTS ratings and found that there was agreement within one score point for approximately 90 percent of all ratings where two assessors performed the rating.[27] This indicates a high level of interrater reliability. Gitomer concluded that "the design features of the NBPTS system support a relatively reliable set of assessments."[28]

Seven large-scale studies investigated the impact of NBPTS certification and offer the necessary power to detect effects, if they exist, and either controlled for student- or school-fixed effects or used hierarchical linear modeling (HLM).[29] These studies provide the best available estimates of the signaling and human capital effects of NBPTS certification.

To summarize, there is a small signaling effect of NBPTS certification, and effects on human capital are either mixed or negative. The average signaling effect size across the seven key studies is 0.002 SD in reading and 0.004 SD in math.[30] This represents the average gain in student achievement of replacing an existing teacher with an NBPTS-certified teacher (the effect is diluted because some teachers in the general population are already teaching at the NBPTS level and would pass the NBPTS exam if they applied while others would fall below the NBPTS standard and would fail the NBPTS exam).

While other studies have investigated the relationship between NBPTS certification and student achievement, none involved a sample with more than 35 NBPTS-certified teachers, none controlled for student-fixed effects, and the only study that used HLM involved a small sample of 25 NBPTS-certified teachers and failed to find any impact on student achievement.[31] These studies are limited by key methodological weaknesses.

A third measure of teacher quality is whether a teacher meets the federal definition of a highly-qualified teacher: a person who has been awarded a minimum of a bachelor's degree from a four-year institution, is fully certificated or licensed by the state in which the teacher teaches, and demonstrates subject-matter competence in each core academic subject area taught by the teacher.[32] At the middle and high school level, teachers may demonstrate competency by showing that they possess a major in the subject area of instruction, credits equivalent to a major in the subject, an advanced level of state certification, a graduate degree, or have passed a state-developed test in the subject area or a test involving a high, objective, uniform state standard developed by a state for the purpose of establishing subject-matter competency. These credentials are only weakly associated with value-added estimates of teacher performance.[33] However, it is unclear whether the problem is weak reliability of the criterion measure or if credentials are indeed weak predictors of teacher contributions to student performance.

Measures of school quality based on value-added measures of school performance raise the issues described above with regard to the use of VAM to measure quality. What is needed is a measure that does not rely upon VAM. Measures that are independent of VAM include student or parent judgments of school quality. Significantly, De Jong and Westerhof found that the quality of aggregated student ratings of eighth-grade mathematics teacher quality is equal to the quality of data obtained from trained external observers.[34] With regard to the quality of data obtained from parents, a survey involving responses from 3,948 District of Columbia Public School parents yielded a strong test-retest reliability coefficient of .937, while the internal reliability of survey items for each section of the survey ranged from .69 to .90.[35] These results suggest that the reliability of data obtained from students and parents may be adequate for the purpose of rating teacher or school quality. Alternatively, an objective indicator that does not rely upon student or parent judgments of quality, such as reports of whether students or teachers have been physically attacked or whether students have been involved in fights, is likely to be correlated with the perceived quality of a school and may be used as a proxy indicator of quality. A reasonable approach might employ multiple measures to arrive at an overall judgment of the degree to which measures of teacher or school quality predict student achievement and explain the persistence of the achievement gap throughout the K-12 years.

Need for Research

Jesse Rothstein's results indicate that VAM omits one or more key variables and is a biased measure of teacher quality.[36] While NBPTS teacher certification is a reliable measure of teacher quality,[37] this measure is not a strong predictor of gains in student achievement, nor is there evidence that any other available measure of teacher or school quality is a strong predictor of gains in student achievement.

The study reported in this chapter employed other available measures of teacher and school quality to evaluate the hypothesis that teacher and school quality are strong influences on student achievement. While the reliability of these other measures is uncertain, a sensitivity analysis suggests that the results reported here are not sensitive to measurement reliability.[38] An alternative approach would be to suspend this type of evaluation until advances in technology produce reliable measures of teacher and school quality that are strong predictors of gains in student achievement. Pursuit of this alternative path, however, presumes that advances in technology will eventually overcome the limitations of current measures such as VAM, and presumes that teacher and school rankings based on these measures will, in the future, prove to be strong, reliable, measures of teacher and school quality. There are three difficulties with this approach.

First, federal funds, including Race-to-the-Top funds, are increasingly being directed toward state educational agencies that promise to implement measures such as VAM and promise to use these measures to make operational decisions regarding the identification and termination of low-performing teachers.[39] There is an urgent need for research to investigate the core assumption underlying this policy. Policymakers cannot and will not wait until ideal measures of teacher and school quality are perfected. Instead, they will continue to proceed with the implementation of policies based on the assumption that teacher and school quality are the primary influences on student achievement—unless and until researchers demonstrate otherwise. Researchers who insist upon methodological purity risk delaying the type of research studies that are urgently needed.

Second, the literature reviewed above suggests reasons to question the core assumption that teacher and school quality are the primary influences on student achievement. NBPTS teacher certification is a reliable measure of teacher quality, yet it is only weakly correlated with gains in student achievement. Other than VAM-based rankings, none of the available measures of teacher quality are strongly correlated with gains in student achievement. If it is true that teacher and school quality are not primary influences on student achievement, then no advance in technology will ever lead to the development of measures of teacher and school quality that are strongly correlated with gains in student achievement. It would only be possible to develop measures of teacher and school quality that are strongly correlated with gains in student achievement if there is, in fact, a strong relationship between teacher and school quality and student achievement. Substantial effort has been expended on extremely sophisticated value-added statistical modeling and the development of NBPTS certification procedures. The results are disappointing. It may be necessary to consider the possibility that the failure to identify strong measures of teacher and school quality is not due to the limits of current technology but instead reflects a need to pursue a different strategy that is based on an alternative view of factors influencing and maintaining the achievement gap.

Third, it is incumbent upon advocates of the view that teacher and school quality are the primary influences on student achievement to develop appropriate measures and demonstrate that those measures are strong, reliable, valid measures of teacher and school quality. If those measures are not developed and made available to other researchers, it would not be appropriate to fault those researchers for failure to use strong, reliable, valid measures of teacher and school quality.

Grading Practices

An alternative view is that the persistence of the achievement gap may be traced to the way that schools are currently structured. A factor that has been overlooked is the psychological impact on children of existing grading practices. When children enter the school system, they are graded and compared to their same-age classmates. This practice undermines children's self-efficacy, engagement, effort and achievement.[40] In particular, low-performing children are continually reminded that their performances are below par. The psychological impact is exacerbated by the introduction of letter grades in middle school. Even relatively high-performing children may be discouraged by an occasional bad grade as they realize they are not "straight-A" students. However, the impact on low-performing children is more severe and this may explain the persistence and growth of the achievement gap as children advance from grade to grade.

Children may receive low grades for several reasons. The tasks that are assigned may be too difficult. In addition, a child may receive a low grade even when demonstrating improvement if the performance of each child is scored in relation to other children, rather than in relation to each child's prior performance.

Lack of individualization has profound effects on children. For example, in one research study, matched children ages 9-11 who were identified as exhibiting poor engagement and performance were randomly-assigned to three groups.[41] Group One received easy math problems, Group Two received moderately-difficult problems, and Group Three received difficult problems. Group One completed all problems with 100 percent accuracy, Group Two achieved 76.9 percent accuracy, and Group Three achieved 46.2 percent accuracy. Significantly, Group Two exhibited the best level of persistence on a subsequent set of math problems. This study indicates the importance of individualizing task difficulty so that children experience a modest but not overwhelming challenge to their levels of competence.

Individualization of task difficulty permits low-performing children to achieve high accuracy scores on daily math assignments and high reading comprehension scores on reading comprehension tests. High scores presumably permit children to feel a sense of accomplishment. It appears that this promotes engagement, effort and achievement.[42]

In combination with individualized task difficulty, it may be important to provide rapid performance feedback on daily math and reading assignments.

Studies indicate that the most effective feedback is objective, involving daily testing, permitting children to see that they are making progress and promoting engagement, effort and achievement.[43]

While individualization of task difficulty and rapid performance feedback might be considered integral aspects of strong teaching, it is not unusual for teachers, including those typically categorized as strong teachers, to give the same set of math problems to all students in each class, and not to grade math homework until days later. It is not unusual for teachers categorized as strong teachers to employ letter-grade report cards and classroom assessments that compare students to each other and demoralize low-achieving students. NBPTS offers perhaps the most rigorous set of standards for distinguishing between strong and weak teachers. However, the NBPTS scoring process does not distinguish between teachers who do, and teachers who do not, employ letter-grade report cards and classroom assessments that compare students to each other. Nor does the NBPTS scoring process distinguish between teachers who do, and teachers who do not, give the same set of math problems to all students in each class, or delay grading math homework.

The challenge of individualizing task difficulty and providing rapid performance feedback on a daily basis for each student in a class of 25 students may be addressed through the use of technology. Evaluations of this technology indicate that it is more efficient than numerous alternative strategies for raising student achievement: voucher programs, charter schools, increased expenditure per pupil, stronger accountability for students and teachers, teacher certification by the National Board for Professional Teaching Standards, class-size reduction, comprehensive school reform, and the use of value-added statistical methods to identify and replace low-performing teachers.[44]

The evidence that the combination of individualized task difficulty and rapid performance feedback is more efficient than numerous alternative strategies for addressing the achievement gap suggests that the persistence of the gap may be understood as a lack of individualized task difficulty, a lack of rapid performance feedback, and a lack of attention to grading practices that inadvertently undermine children's self-efficacy, engagement, and achievement. In this view, differences in achievement that exist at kindergarten are perpetuated through a system of grading practices that undermine the self-efficacy of low-performing children in a way that maintains the gap in achievement through the end of high school.

This view received support from three randomized studies.[45] In each study, a randomized treatment group received a technology-based intervention that individualized task difficulty in either math or reading, in combination with rapid performance feedback. The results offer strong evidence that individualization of task difficulty, in combination with rapid performance feedback, raises student achievement. It appears that the intervention operates by improving student self-efficacy, engagement, and effort.

While randomized studies are preferred to regression analyses because the latter are correlational studies that, by nature, cannot establish causal relationships, it is impractical to perform randomized studies with nationally-representative samples of students. In addition, this type of study does not permit a direct test of hypothesized path effects with nationally-representative samples of students. Path modeling with nationally-representative samples of students would permit a direct test of the relative strength of factors that hypothetically mediate early and late academic achievement, explaining how differences in achievement that exist at kindergarten may be translated into differences in achievement that exist at the end of high school.

An advantage of using nationally-representative samples of students is that the results may be generalized to the entire population of American students. A disadvantage, however, is that researchers are limited to the indicators and measures for which data were collected. In particular, researchers who wish to investigate the influence of teacher or school quality are limited to the indicators and measures of teacher and school quality that were used in the available studies involving nationally-representative samples of students. These indicators include the federal measure of high-quality teachers, student judgments of the quality of their teachers, and parental judgments of the quality of the schools attended by their children. While some researchers may prefer the use of value-added measures of teacher quality or NBPTS certification as a signal of teacher quality, the studies reviewed above indicate that value-added measures are biased and NBPTS certification is only weakly correlated with gains in student achievement. Even if these measures were available with nationally-representative samples of students, it would not be sensible to substitute these measures for the measures included in the current study. While the federal measure of high-quality teachers, student judgments of the quality of their teachers, and parental judgments of the quality of the schools attended by their children may have limitations as indicators of teacher and school quality, there is no consensus regarding suitable indicators and there is no expectation that the problem of developing suitable indicators will be solved in the near future. Regardless, policymakers need information about the most promising strategies for improving student achievement. There is a need for studies that directly test and compare various theories about the nature of the achievement gap and compare promising ways of thinking about how to address the gap. This chapter reports the results of a study that employed path analysis with data from nationally-representative samples of students to investigate factors, using available measures, that are hypothesized to maintain and perpetuate initial differences in achievement that exist at kindergarten.

Methods

Path analysis was employed to compare two theories regarding the persistence of the achievement gap using data from three surveys sponsored by the National Center for Education Statistics (NCES). The Early Childhood Longitudinal Study of the Kindergarten Class of 2010-11 (ECLS-K:2011) is currently following a nationally-representative cohort of 18,170 children who attended

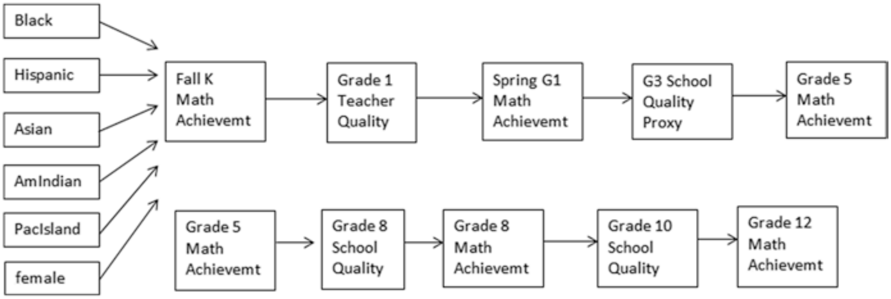

Figure 6.1 Model of achievement, teacher quality, and school quality

kindergarten during the 2010-2011 school year.[46] As of June, 2015, data from the kindergarten and first-grade years had been released. Data from this survey were used for path analyses covering the fall kindergarten through spring first-grade period of each student's academic career. The Early Childhood Longitudinal Study of the Kindergarten Class of 1998-1999 (ECLS-K) followed a nationally-representative cohort of 21,260 children from kindergarten into middle school.[47] Data from this survey were used for path analyses covering the spring first-grade through eighth-grade period of each student's academic career. The National Education Longitudinal Study (NELS) followed a nationally-representative cohort of 27,394 individuals who were surveyed as eighth-grade students in 1988, tenth-grade students in 1990, and twelfth-grade students in 1992.[48] Data from this survey were used for path analyses covering the eighth- through twelfth-grade period of each student's academic career. The data collected through each survey included data from student, parent, teacher, and school administrator questionnaires, standardized reading and math assessments, and administrative records.

The path diagram in Figure 6.1 suggests that differences in math achievement at entry at kindergarten are perpetuated and maintained by differences in the quality of the teachers and schools experienced by students. In this model, socioeconomic and sociocultural factors related to race as well as gender contribute to differences in achievement that exist upon entry at kindergarten. These differences in student achievement are presumed to be associated with race-related socioeconomic differences that influence residential location and are correlated with the quality of schools and teachers experienced by students. It is hypothesized that these differences in school and teacher quality magnify and perpetuate differences in academic achievement throughout the students' academic careers.

In this model, the expectation is that differences in school and teacher quality experienced by students are correlated with race and socioeconomic status, as well as achievement, throughout the students' academic careers, that is, low-achieving minority students from poor families are more likely than high-achieving white students from middle-income families to experience below-average schools and teachers. Therefore, indicators for race and socioeconomic status were intentionally omitted as covariates after the kindergarten time period because it would not be appropriate to include covariates that are highly correlated with school and teacher quality (i.e., in the presence of collinearity). The model was explicitly designed to investigate the *unconditional* influence of school and teacher quality on student achievement and, conversely, the *unconditional* influence of differences in student achievement that are presumably related to racial and socioeconomic differences influencing residential location and correlated with the quality of schools and teachers experienced by students.

Race and ethnicity data were collected from parent interviews. Math achievement was measured by standardized math assessment scores. Grade one teacher quality was measured by dichotomous teacher responses to the question, “This school year, do you qualify as a ‘highly qualified teacher (HQT)’ according to your state’s requirements?” Grade three school quality was measured by dichotomous school administrator responses to the question: “Have any of the following things happened during this school year at this school: Children or teachers being physically attacked or involved in fights?” Grade eight school quality was measured by level of parent agreement with the statement “(child)’s school is a good school.” Grade ten school quality was measured by level of student agreement with the statement “the teaching is good at this school.” All teacher and school quality variables were recoded so that high values indicated high quality and low values indicated low quality.

The decision to employ multiple measures of teacher and school quality throughout the path analyses was dictated by the availability of the measures in the ECLS-K and NELS datasets at each grade level where data were collected. It was not feasible to restrict the path analyses to a single type of school quality measure. An advantage, however, of using multiple measures is that this approach permits a judgment about whether the particular choice of measure affects conclusions about the influence of teacher and school quality on student achievement.

The path diagram in Figure 6.2 suggests that differences in math achievement upon entry at kindergarten are perpetuated and maintained by differences in the levels of self-efficacy experienced by students. In this model,

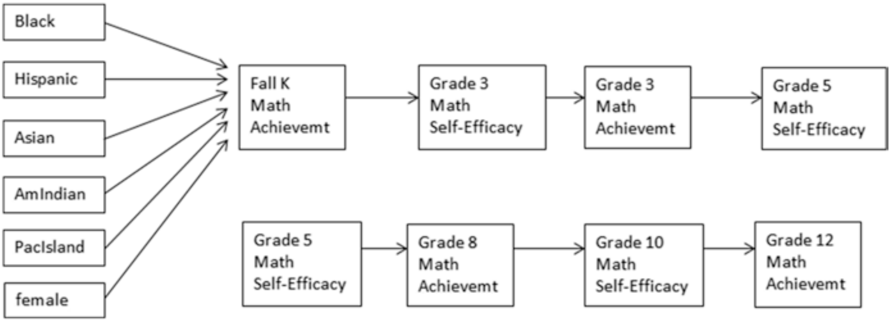

Figure 6.2 Model of achievement and self-efficacy

socioeconomic and sociocultural factors related to race as well as gender contribute to differences in achievement that exist upon entry at kindergarten. It is hypothesized that these differences are associated with differences in the comments, grades, test scores, and other cues received by students that magnify and perpetuate differences in self-efficacy and academic achievement throughout the students' academic careers.

In this model, the expectation is that differences in student self-efficacy across students are correlated with race and socioeconomic status, as well as achievement, throughout the students' academic careers, that is, low-achieving minority students from poor families are more likely than high-achieving white students from middle-income families to experience sharp, lengthy declines in self-efficacy over the K-12 years. Therefore, indicators for race and socioeconomic status were intentionally omitted as covariates after the kindergarten time period because it would not be appropriate to include covariates that are highly correlated with student self-efficacy (i.e., in the presence of collinearity).[49] The model was explicitly designed to investigate the *unconditional* influence of student self-efficacy on student achievement and, conversely, the *unconditional* influence of student achievement on student self-efficacy.[50] Appendix C addresses hypotheses that race, income, or socioeconomic status might explain why black, Hispanic, low-income, and low-SES children exhibit depressed levels of self-efficacy.

Race and ethnicity data were collected from parent interviews. Math achievement was measured by standardized math assessment scores. Math self-efficacy was measured by the level of student agreement with the statement "I am good at math" (grades three and five) or "mathematics is one of respondent's best subjects" (grade ten). These are the only items available in the ECLS-K and NELS datasets that are consistent with Albert Bandura's construct of perceived self-efficacy, defined (in this case) as expectations and convictions that a student can successfully execute the behavior required to solve math problems.[51] Other items measure the extent to which respondents "like" or "look forward to" or "enjoy work" in math. These measures relate to a student's affect, anticipation for the subject of math, and degree of positive emotions, rather than expectations and convictions about whether a student can successfully execute the required behavior. These constructs are different. A student might have positive feelings but lack a strong conviction that he or she can solve math problems. Or, a student might be confident about solving math problems, but not derive pleasure in pursuing them. Lumping inconsistent measures together would be conceptually problematic and would interfere with the interpretation of the results.

The path diagram in Figure 6.3 suggests that achievement in grade five influences later achievement (in grade eight) by influencing grades received by students in grade five, levels of academic interest and self-efficacy in grade five, and the level of effort exerted by students in grade five. Grades in grade five were measured by level of student agreement with the statement "I get good grades in all school subjects." This self-report measure reflects each student's

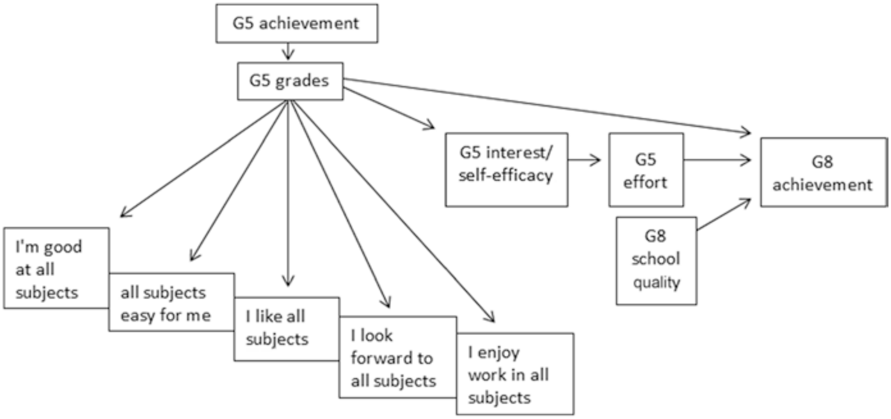

Figure 6.3 Grade 8 achievement model

own evaluation of his/her grades and is presumably a better measure for the purpose of evaluating the psychological impact of grades on each student's level of academic interest and self-efficacy. The academic interest and self-efficacy indicator is a composite indicator measuring level of student agreement with statements regarding academic interest ("I like all school subjects," "I enjoy work in all school subjects," "I look forward to all school subjects") and academic self-efficacy ("I am good at all school subjects," "work in all school subjects is easy for me"). This composite indicator was computed by the researcher as the mean of the items comprising the score and was only created if there were valid data on at least four of the five items. This indicator is similar but not identical to a composite indicator created by NCES and labeled "perceived interest/competence in all school subjects." Each student's effort in grade five was measured by combining the responses of the student's teacher to two queries ("How often does this child work to the best of her/his ability in reading?" and "How often does this child work to the best of her/his ability in math?") into a single composite measure of each child's level of effort. Each student's achievement in grade eight was measured by combining each student's standardized math and reading assessment scores. Grade eight school quality was measured by level of parent agreement with the statement "(child)'s school is a good school."

Path coefficients were derived from regressions of outcomes on predictors as indicated in Figure 6.1, Figure 6.2, and Figure 6.3. Dichotomous outcomes were modeled using logistic regression. Continuous measures were standardized to permit comparisons of effect magnitudes. For continuous predictors, each path coefficient corresponds to a one-standard deviation change in the corresponding predictor. For dichotomous (yes/no) predictors, each path coefficient corresponds to "yes" values of the predictors. The data, involving complex surveys and nonindependent observations, did not permit the use of log-likelihood, Akaike information criterion (AIC), or

Bayesian information criterion (BIC) statistics to compare the models in Figures 6.1 and 6.2.[52] However, the standardized path coefficients are comparable.

Results

The hypothesis that the contribution of teachers to student performance is the strongest factor influencing student achievement is not supported. A stronger factor is the degree to which students believe that they are proficient students. The path coefficients in Figures 6.4 and 6.5 indicate that self-efficacy is a stronger predictor of student achievement than school or teacher quality at every level of schooling.

The effect of the grade one measure of teacher quality on student achievement in spring of grade one is not significantly different than zero. The grade three measure of school quality, the grade eight measure of school quality, and the grade ten measure of school quality never exhibit path coefficients exceeding .18. The path coefficients bounce up and down and do not demonstrate a consistent pattern. In addition, the level of student achievement in fall kindergarten is not a significant predictor of grade one teacher quality. After kindergarten, the level of student achievement modestly predicts school quality. This supports the hypothesis that differences in achievement (presumably associated with socioeconomic status) are associated with residential decisions that influence the quality of the schools attended by students, but suggests that the effect is modest.

In contrast, the path coefficient for the grade three measure of self-efficacy on grade three achievement equals .17, the path coefficient for the grade five measure of self-efficacy on grade eight achievement equals .25, and the path coefficient for the grade ten measure of self-efficacy on grade twelve achievement equals .34, with all path coefficients significant at a .001 alpha level. The steady increase in the magnitude of the path coefficients as students advance from grade to grade, and the doubling of the path coefficient from grade three to grade twelve, are consistent with the hypothesis that grading and testing practices systematically erode student self-efficacy throughout the academic careers of low-achieving students, systematically eroding student achievement in a way that maintains, perpetuates, and widens the differential between high- and low-self-efficacy students. In addition, there appears to be a significant feedback effect that increases in magnitude as students advance from grade to grade, tripling in magnitude between kindergarten and grade five. The path coefficient for the fall kindergarten measure of student achievement on grade three math self-efficacy equals .09. The path coefficient for the grade three measure of student achievement on grade five self-efficacy equals .27. The path coefficient for the grade eight measure of student achievement on grade ten self-efficacy equals .28. Once again, the pattern of effects increases as students advance from grade to grade. All of these coefficients are significant at the .001 alpha level. This pattern of results is consistent with the hypothesis

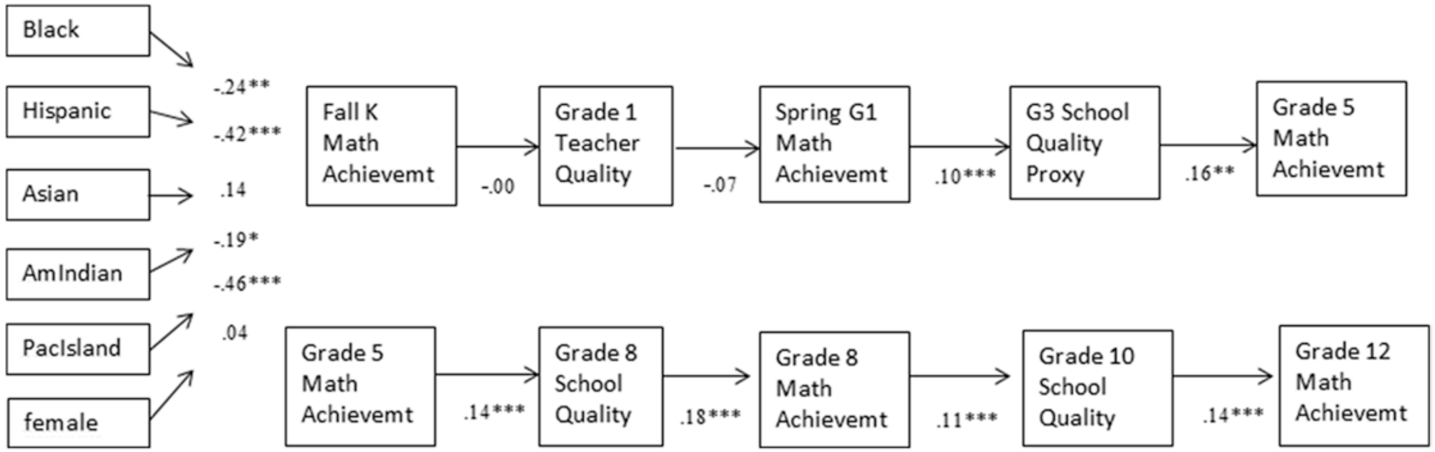

Figure 6.4 Model of achievement, teacher quality, and school quality with standardized path coefficients ($*p < .05$ $**p < .01$ $***p < .001$)

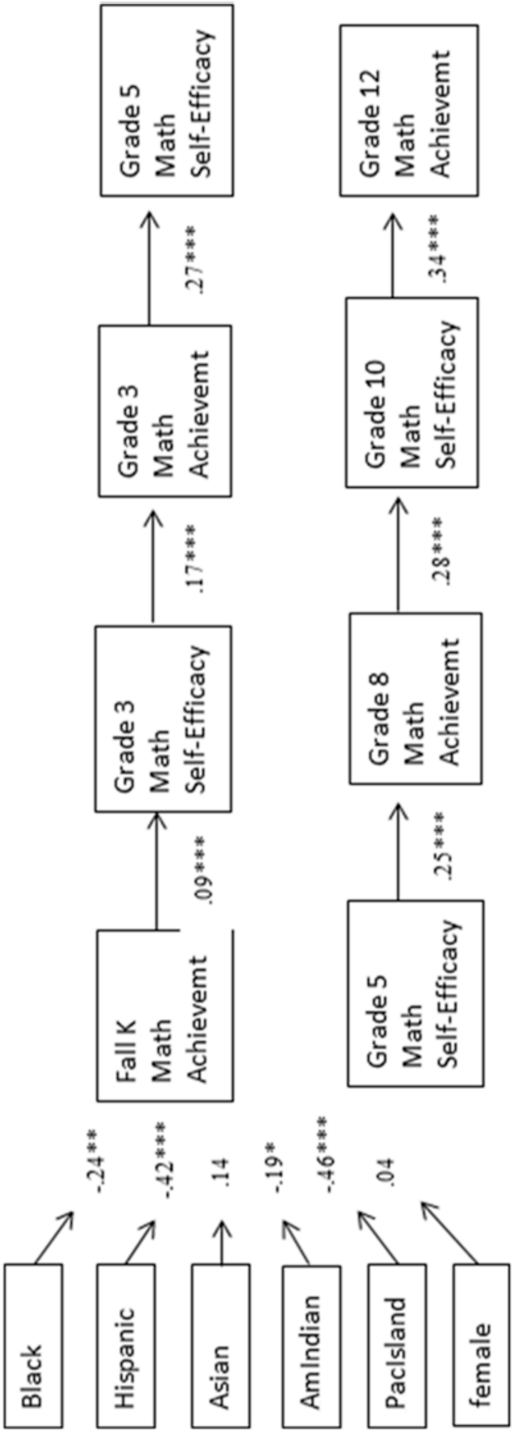

Figure 6.5 Model of achievement and self-efficacy with standardized path coefficients (*$p < .05$ **$p < .01$ ***$p < .001$)

that grading and testing practices exert a corrosive effect on student self-efficacy throughout the academic careers of low-achieving students and the effect is magnified because depressed achievement feeds back and further depresses self-efficacy in a negative downward spiral that strengthens in magnitude as students advance from grade to grade. These downward spirals may explain the persistence of the achievement gap despite the best efforts of teachers to address it.

The path coefficients in Figure 6.6 support this interpretation. Grades have a strong effect (path coefficient equal to .59) on the composite measure of academic interest and self-efficacy. The direction of influence may be inferred from previous research indicating that children enter kindergarten with relatively high levels of academic interest and self-efficacy but something about the interaction of children with the school system causes interest and self-efficacy to decline at an accelerating rate as children advance from grade to grade.[53] This suggests that sociocultural and family influences that exist prior to the point when children enter the school system equip children with relatively high levels of interest and self-efficacy that are then eroded after children enter the school system in kindergarten. This implies that the direction of causation runs from grades to interest/efficacy, not the reverse (otherwise, interest/efficacy would remain high, instead of declining after students enter kindergarten).

The path coefficient relating academic interest/efficacy to effort indicates that the level of student interest/efficacy is related to the level of effort exerted by each student. Presumably, the causal direction runs from interest/efficacy to effort, not the reverse (unless the exertion of effort causes interest and self-efficacy to increase). The components of the composite indicator of interest/efficacy suggest that students who receive high grades tend to like, enjoy, and look forward to, activities in all of their academic subject areas, feel confident about their abilities in those areas and, as a consequence, exert relatively high levels of effort that contribute to achievement three years later in grade eight. Conversely, it appears that students who receive low grades tend to dislike, do not enjoy, and do not look forward to, academic activities, do not feel confident about their academic abilities and, as a consequence, exert relatively low levels of effort that contribute to depressed achievement three years later in grade eight.

The strong association between grades and interest/efficacy, the association between interest/efficacy and effort, and the evidence suggesting that the causal direction runs from grades to interest/efficacy (and presumably from interest/efficacy to effort) suggest that the correlation between grades and effort is explained by the causal effect of grades operating through interest/efficacy on effort, not the reverse.

The direct effect of grades on achievement in grade eight is .20 SD. The indirect effect of grades on achievement in grade eight is .59 × .25 ×.31, or .05 SD. The total effect size is .20 plus .05, or .25 SD. This is substantially larger than the .14 SD effect size of grade eight school quality on achievement

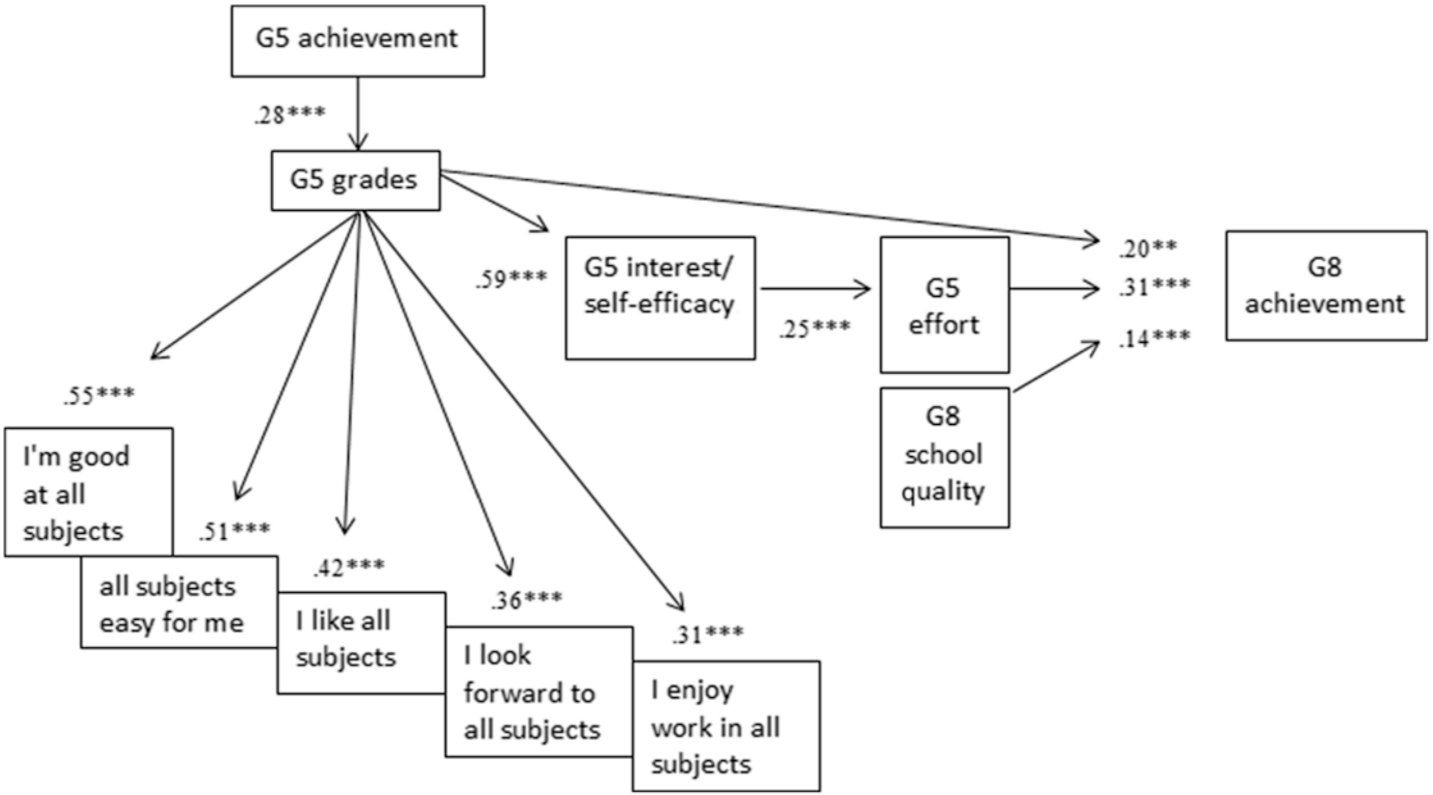

Figure 6.6 Grade 8 achievement model with standardized path coefficients (*p < .05 **p < .01 ***p < .001)

in grade eight. In addition, the largest immediate influence on achievement in grade eight is student effort in grade five. The effect size is .31 SD, more than twice the effect of grade eight school quality. This suggests that grades and student effort, measured in grade five and mediated by level of academic interest and self-efficacy in grade five, are much more important influences on achievement in grade eight than school quality measured in grade eight.

These results should not be surprising. The Coleman report was the first national study to document the existence of substantial differences in educational achievement between black and white students at every grade level.[54] The Coleman report demonstrated that these differences increased as students progressed from first through twelfth grade and demonstrated that the strongest predictor of student achievement for black and Hispanic students was a student's perceived control over his or her environment.[55] For black and Hispanic students, this factor was stronger than any other school or background variable, including parental education.[56] This suggests a psychological explanation for low minority student achievement instead of an explanation that focuses on school or teacher quality.

The results reported here support the hypothesis that students who enter kindergarten performing above their same-age classmates tend to receive grades, test scores, and teacher comments that reinforce student interest in academic activities and feelings of competence and self-efficacy with regard to academic activities throughout the K-12 years. This reinforces and promotes high levels of effort that tend to maintain high levels of achievement. Conversely, students who enter kindergarten performing below their same-age classmates tend to receive grades, test scores, and teacher comments that undermine student interest in academic activities and feelings of competence and self-efficacy with regard to academic activities. This undermines student effort in a way that tends to further depress achievement throughout the K-12 years.

Conclusion

In the United States, a disproportionate fraction of students who enter kindergarten performing above their same-age classmates happen to be white and Asian. A disproportionate fraction of students who enter kindergarten performing below their same-age classmates happen to be black and Hispanic. The results reported here suggest that the combination of this circumstance with the universal practice of grading, testing, and comparing students to their same-age classmates may be sufficient to explain the persistence of the achievement gap throughout the K-12 years.

What explains the persistence of the idea that low achievement is associated with low school quality? Schools where achievement is low are disproportionately characterized by classrooms filled with students who are disengaged, apathetic, and disruptive. Teachers may have difficulty commanding the attention of their pupils. Pupils and teachers may engage in testy exchanges. Discipline tends to be poor. Classroom management tends to be poor. Conversely, schools where achievement is high are characterized by classrooms filled with students who are engaged, on-task, and eager to learn. Teachers have no difficulty commanding the attention of their pupils. Interactions between pupils and teachers tend to be pleasant and cooperative. Discipline tends to be good. Classroom management tends to be good.

A casual observer would have no difficulty in categorizing the first set of schools as "bad" schools and the latter set of schools as "good" schools. Arguably, however, all of the characteristics of "good" and "bad" schools are predictable when "good" schools happen to be filled with students with high self-efficacy and high learning potential, and "bad" schools happen to be filled with students with low self-efficacy and low learning potential. Students with high self-efficacy are engaged, on-task, eager to learn, exhibit high learning potential and large gains in achievement. Students with low self-efficacy are disengaged, disruptive, apathetic, exhibit low learning potential and low gains in achievement. Teachers who teach in classrooms filled with high-self-efficacy students have no difficulty commanding the attention of their pupils, communicating pleasantly, maintaining discipline, managing their classrooms, and raising student achievement. Teachers who teach in classrooms filled with low-self-efficacy students have tremendous difficulty commanding the attention of their pupils, maintaining pleasant communication, maintaining discipline, managing their classrooms, and raising student achievement.

The analysis presented in this chapter suggests, however, that correlation has been mistaken for causation. The characteristics of "good" schools are associated with high student achievement, and the characteristics of "bad" schools are associated with low student achievement. But it appears that the relationship is not a causal relationship, where school or teacher quality causes differences in student achievement. Instead, the characteristics of "good" and "bad" schools, as well as the levels of achievement that characterize "good" and "bad" schools, may instead be traced to a third factor, namely, the nearly universal practice of grading, testing, and comparing students to their same-age classmates, causing demoralization among low-achieving students that triggers disengagement, reduction of effort, and reductions in learning potential. This translates into depressed achievement which further reduces self-efficacy, engagement, effort, and future achievement in a downward spiral that magnifies and perpetuates initial differences in achievement that exist at kindergarten. This process generates large numbers of low-performing students who fill schools in low-income urban areas. These schools then acquire reputations as bad schools.

The notion that teacher and school quality are the key influences on student achievement has been maintained by research suggesting that the contribution of teachers to student achievement is large and value-added estimates of teacher contributions predict their students' measured achievement.[57] Many researchers accept this evidence at face value.

However, the analysis presented here explains and resolves the puzzling contradictions involving VAM: the existence of large numbers of teachers who exhibit high value-added estimates of their contributions to student achievement; the poor predictive reliability of teacher rankings based on VAM; the failure of measures of race, poverty, and prior achievement to fully adjust and control for student heterogeneity when estimating the value-added contribution of each teacher; and the seemingly impossible finding that VAM estimates predict the achievement of each teacher's students prior to the point where the teacher receives those students. These contradictions are not easily explained in any other way. The capacity to explain these contradictions is powerful evidence supporting the proposed explanation of the achievement gap.

In October 2015, Jesse Rothstein analyzed and responded to Raj Chetty, John Friedman, and Jonah Rockoff's series of arguments and studies defending VAM.[58] Rothstein's analysis culminated five years of private and public communication and debate among the researchers in an attempt to pinpoint the source of differences in their assessments of the reliability and validity of using VAM for the purpose of estimating the contribution of individual teachers to student achievement. Rothstein concluded that none of the arguments and evidence presented by Chetty et al. alter the conclusion that VAM-based estimates of teacher quality are biased, unreliable, and invalid:

> My results are sufficient to re-open the question of whether high-value added elementary teachers have substantial causal effects on their students' long-run outcomes . . . [there is] no strong basis for conclusions about the long-run effects of high- vs. low-value added teachers, which in the most credible estimates are not distinguishable from zero.[59]

The most credible VAM-based estimates of the contribution of individual teachers to student achievement "are not distinguishable from zero" and any statement, based on VAM, that teachers make significant contributions to student achievement is open to question. Rothstein's analysis challenges not only the Chetty et al. analysis, but also the previous body of studies, based on VAM, suggesting that teachers make significant contributions to student achievement.[60] The conclusion that VAM is flawed raises serious questions about the assertion that teachers make significant contributions to student achievement.

What this suggests is a need for a fundamental reordering of current ideas about the key factors influencing student achievement and the gap in achievement. It suggests a need to rethink the best approaches for addressing the gap. Most importantly, it suggests a need to reconsider the idea that when student achievement is low, the cause is bad schools and bad teachers.

CHAPTER 7

Consequences for Minorities

Policymakers wish to understand the changes that must be made to the primary and secondary education system so that students are highly prepared, ready to perform at high levels, and likely to be successful when they apply to college. However, while previous studies have established that certain interventions are more effective and efficient than other interventions for the purpose of raising student achievement, this research has received little attention.[1] This may be attributed to a widespread perception that student achievement, as measured by standardized tests, may be useful for predicting freshman undergraduate letter grades (FGPA), but should not be overemphasized.[2] This perception may contribute to the problem of low achievement by inadvertently drawing attention away from the need to swiftly adopt the most efficient approaches for raising achievement in the elementary and secondary grades.

The purpose of the analysis reported in this chapter is to investigate the significance of strong educational preparation for all students—and especially for minority students—as measured by standardized test scores. While the importance of educational preparation may already seem well-established, previous studies that regressed baccalaureate attainment on SAT score and high school grade-point average (GPA) are not adequate for understanding the potential impact of interventions that raise student achievement in the K-12 years. These studies held GPA constant, whereas an intervention that raises student achievement may be expected to raise both test scores and GPA in tandem. What is needed is an analysis that aggregates the direct effect of an intervention that operates through test scores on postsecondary outcomes, plus the indirect effect of the intervention operating through GPA to improve postsecondary outcomes, to arrive at the total effect. Thus, the purpose of the analysis reported in this chapter is to estimate the total effect of a hypothetical intervention that raises student achievement by one standard deviation (SD).

This chapter reports analyses involving a nationally-representative sample of students in a way that graphically emphasizes the significant disadvantages that arise when students are not well-prepared. The results reported here underline

S.S. Yeh, *Solving the Achievement Gap*,
DOI 10.1057/978-1-137-58767-1_7

the importance of swiftly adopting the most efficient approaches for raising student achievement. To the extent that existing approaches for raising student achievement are unproductive, inefficient, and disproportionately affect minority students, current policies may serve to depress baccalaureate attainment rates and to perpetuate the disadvantaged status of minorities. The lack of corrective action may be partly attributable to an incomplete understanding of the significance of strong educational preparation as measured by test scores.

Measuring Preparation

It is well-accepted that student achievement, measured by test scores, influences educational outcomes.[3] Test scores are important standardized indicators of the educational preparation that the primary and secondary education system has imparted. A standardized measure is important because it can be used to measure the size of the gap that needs to be filled and to measure the extent to which interventions to raise student achievement will fill that gap.

One issue is the appropriate criterion measure. Studies that use undergraduate freshman GPA (FGPA) as the criterion measure suffer from two problems. Restriction of range occurs when students with high test scores and grades gravitate toward one set of institutions and students with low test scores and grades gravitate toward a second set of institutions. Criterion reliability is affected when FGPA does not reflect the difficulty of each course. Some students elect challenging courses, while others elect easier courses. The grades received do not reflect the difference in level of difficulty. The problem is exacerbated if students with greater aptitude elect challenging courses while students with lesser aptitude elect easier courses. After correcting for predictor restriction of range and criterion unreliability, the validity coefficient for SAT scores is similar to the validity coefficient for high school GPA, indicating that SAT scores are approximately as valid as high school GPA for the purpose of predicting FGPA.[4] However, the problems with FGPA as a criterion outcome suggest that college graduation is a better criterion.

In perhaps the most influential study, Bowen and Bok investigated the relationship between combined SAT scores and baccalaureate graduation rates for students who matriculated at 28 academically-selective colleges and universities and found that both black and white students graduated at higher rates at more selective institutions.[5] This result suggested that it may be desirable to remove barriers that impede minorities from enrolling at selective institutions. However, Bowen and Bok's data indicate that there was a positive relationship between SAT scores and graduation rates.

A subsequent study by Bowen and two co-authors found that SAT scores explained much of the variation in student performance at 19 academically-selective postsecondary institutions.[6] After controlling for SAT scores, students from disadvantaged socioeconomic backgrounds did not underperform their more advantaged counterparts, graduated at comparable rates, and were equally successful in attaining lucrative law and business degrees.[7] Bowen and his colleagues concluded that strong preparation, captured by math and reading test scores, can overcome socioeconomic disadvantages and "is the major determinant of differences in educational attainment" between advantaged and disadvantaged young people.[8]

While a third study by Bowen and two co-authors[9] concluded that the influence of test scores is relatively small and high school GPA is a better predictor of baccalaureate attainment rates, other studies suggested that academic preparation as measured by test scores has an important influence on postsecondary outcomes. A synthesis of research studies involving 400 institutions and 82,000 students found that SAT scores are a better predictor of baccalaureate attainment rates than high school GPA.[10] A national study of 1,429 baccalaureate-granting institutions found correlations ranging from .62 to .73 between graduation rates and SAT or ACT scores.[11] A national study of 262 baccalaureate-granting institutions found that students with combined SAT scores of 1300 or above were about three times more likely to attain a baccalaureate degree within four years compared to students scoring below 800.[12] An analysis of data involving 12,144 students from the nationally-representative National Education Longitudinal Study found that only 16.1 percent of students scoring in the bottom quartile on the eighth-grade mathematics test subsequently completed a baccalaureate degree, compared to 67.0 percent of those scoring in the top quartile.[13] An analysis of longitudinal data from the nationally-representative High School and Beyond Study of 10,470 high school students found that only three percent of those scoring in the bottom quintile of a short version of the SAT administered during their senior year subsequently completed a baccalaureate degree, compared to 64 percent of those scoring in the top quintile.[14] An analysis of data from 22,652 high school seniors who participated in the National Longitudinal Study of the High School Class of 1972 found that college students ranking at the 25th percentile of their high school class were twice as likely to drop out if their combined SAT score was 700, compared to 1300.[15] College students ranking at the 100th percentile of their high school class were over three times as likely to drop out if their SAT score was 700, compared to 1300.[16]

A study of academically-competitive colleges found that fewer than 14 percent of students with combined SAT scores of 1000 or lower compiled a college freshman grade-point average (FGPA) of 3.5 or higher while over half of the students with SAT scores over 1200, and 77 percent of the students with SAT scores over 1400, reached this standard.[17] A study of academically-selective colleges found a similar pattern for four-year college GPAs: none of the students scoring below 800 had four-year GPAs of 3.5 or higher while over 50 percent of students with SAT scores equal to or exceeding 1410 met this standard.[18]

These results suggest that test scores do indeed measure significant differences in academic preparation that are reflected in important outcomes. If the purpose is to predict the impact on these outcomes of raising student achievement by one standard deviation (SD), it is reasonable to use test scores as a standardized measure of achievement.

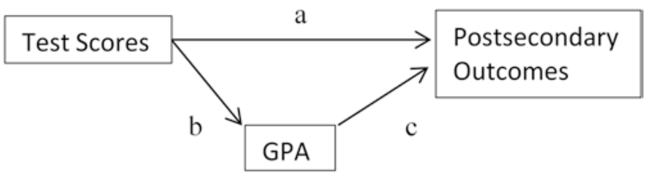

Figure 7.1 Path model of relationships among test scores, GPA, and postsecondary outcomes

Even the best available studies, however, are inadequate for this purpose. These studies typically regress baccalaureate attainment on SAT score and high school GPA and are intended to predict the college performance of a particular individual. However, these studies are not adequate for understanding the potential impact of an intervention that raises student achievement in the K-12 years because these studies hold GPA constant, whereas an intervention that raises student achievement may be expected to raise both test scores and GPA in tandem. What is needed is a study that aggregates the direct effect of an intervention that operates through test scores directly on postsecondary outcomes, plus the indirect effect of the intervention operating through GPA to raise postsecondary outcomes, to arrive at the total effect (Figure 7.1). This type of path analysis was not conducted in any of the previous studies reviewed above. Thus, the purpose of the analysis reported in this chapter is to estimate the total effect of a hypothetical intervention that raises student achievement by one SD.

Key outcomes include the probability of dropping out of high school, compiling a rigorous high school record, completing algebra 2 in high school, completing calculus in high school, enrolling at a four-year institution, compiling an undergraduate GPA of 3.75 or above, attaining a baccalaureate degree, and aspiring to a doctorate or professional degree. Students who drop out of high school are economically disadvantaged. They are not able to earn as much as individuals who complete high school. Similarly, students who do not attain a baccalaureate degree are economically disadvantaged. On average, they do not earn as much as individuals who complete a baccalaureate degree. Students who do not complete a rigorous high school curriculum or algebra 2 are disadvantaged when applying for college. The completion of calculus is an indicator of a student's level of preparation in mathematics and preparation for careers in science, technology, engineering, and mathematics. Enrollment at a four-year institution and compilation of an undergraduate GPA above 3.75 are important indicators of progress toward successful completion of a baccalaureate degree. Aspiring to a doctorate or professional degree is an important indicator of a student's educational aspirations. If a student does not aspire to a doctorate or professional degree it is unlikely that the student would apply to a doctoral or professional degree program and unlikely that the student would complete a doctoral or professional degree program.

NELS

The National Education Longitudinal Study (NELS), sponsored by the National Center for Education Statistics, followed a nationally-representative cohort of 27,394 individuals who were surveyed as eighth-grade students in 1988, tenth-grade students in 1990, twelfth-grade students in 1992, and again in 1994 and 2000, eight years after their expected high school graduation date. In addition to the student survey data, information was collected from parents, teachers and school administrators, high school transcripts, postsecondary institutions and transcripts, and standardized reading and math tests developed by the Educational Testing Service specifically for NELS. The analysis employed the results of tests administered during the sophomore year (1990). To guard against ceiling and floor effects and improve accuracy, each sample member was administered a test form that aligned the difficulty level of the mathematics and reading questions to the examinee based on his or her scores on the eighth-grade base year mathematics and reading tests. Unlike the SAT or ACT, the tests were administered to all students in the nationally-representative sample and therefore do not suffer from restriction of range. The measure of socioeconomic status was constructed from parent and student surveys using parental education levels and occupations and family income. The dropout indicator was constructed from transcript and survey information. Postsecondary transcripts and institutional data were used to determine baccalaureate attainment and to construct the indicator of attendance at a four-year institution. High school transcripts were used to determine whether a student received an AP exam score in any subject and to determine course enrollment in specific courses and patterns of courses. Undergraduate GPA and educational aspirations by age 30 were each constructed from student survey responses at the last follow-up in 2000.

Analysis

High school GPA was calculated from transcript data. GPA and test scores were standardized. Subgroup weighted means for test scores, GPA, and the socioeconomic status index, plus test scores one SD above the mean, were calculated by race. In stage one, high school GPA was regressed on test scores, socioeconomic status, sex, and categorical variables for black, Hispanic, Asian, and American Indian/Alaska Native students (the excluded category was white students). In stage two, logistic regression was employed to regress each of eight outcomes on test scores, GPA, socioeconomic status, sex, and categorical variables for black, Hispanic, Asian, and American Indian/Alaska Native students and the northcentral, south, and west regions of the United States (the excluded categories were white students and northeast region). Insignificant predictors were dropped to arrive at final models for the eight outcomes. The eight categorical outcomes were: a.) ever dropped out of high school, b.) ever attended a four-year postsecondary institution, c.) attainment of baccalaureate degree, d.) completed a highly-rigorous high school curriculum, e.) took calculus in high school, f.) took algebra 2 in high school, g.) attained an undergraduate GPA above 3.75, and h.) aspired to a PhD or professional degree. Predicted percentages for each of the eight outcomes were generated by race and sex, using corresponding values of mean test scores, predicted GPA from the stage one regression, socioeconomic status, and the race, sex, and region categorical variables. Predictions were repeated using values of test scores at one SD above the corresponding mean values, including the increase in predicted GPA from the stage one regression that would be expected if test scores were increased by one SD. Alternative estimates employing hierarchical linear modeling (HLM) are reported in Appendix D. Partial correlations (a measure of effect size) were calculated for the outcomes and factors included in the stage one and stage two regressions.

Results

Results of the stage one ordinary least squares (OLS) regression are reported in Table 7.1. As expected, GPA is strongly predicted by test scores. The magnitude of the effect exceeds the magnitude of the effect of socioeconomic status by 74 percent. For white students, a one SD increase in test scores is predicted to increase GPA by 0.45 SD. At the mean test score for black males, individuals in this group are predicted to have GPAs that are 0.63 SD below the predicted GPA for white males with mean test scores for white males. At one SD above the mean test score for black males, individuals in this group are predicted to have GPAs that are 0.28 SD below the predicted GPA for white males with mean test scores for white males.

Results of the stage two logistic regressions are reported in Tables 7.2a and 7.2b and partial correlations are reported in Appendix E. The effect size for test scores exceeds the effect size for socioeconomic status in every comparison, and exceeds the effect size for GPA in all but one comparison. For each outcome, results for the full model as well as the final reduced model are reported. Table 7.3 reports the predicted percentages of black and Hispanic students attaining each outcome when test scores are one SD above the mean score for the corresponding racial group, in comparison with the predicted percentages of black and Hispanic students meeting each outcome when test scores are at the mean of the corresponding group (if test scores are normally distributed, a one SD increase above the mean score corresponds to an increase from the 50th to the 84th percentile of test scores for the relevant group). By every measure, high-scoring students are significantly better prepared and perform significantly better than students who score at the mean. In most cases, the difference in educational preparation and performance is large; in some cases, it is extremely large.

Table 7.1 OLS Estimates and Predicted Values of GPA at Mean Scores and 1 SD Above Mean Scores, by Race and Sex

	Full Model	*Reduced Model*
test scores	0.45***	0.45***
	(0.02)	(0.02)
SES	0.10***	0.09***
	(0.02)	(0.02)
sex	0.18***	0.18***
	(0.03)	(0.03)
Asian	0.18***	0.18***
	(0.04)	(0.04)
Hispanic	0.07	
	(0.05)	
black	−0.25***	−0.26***
	(0.06)	(0.06)
AmIndian/AK	−0.11	
	(0.25)	
constant	−0.02	−0.01
n	9652	9652
R^2	0.29	0.29
Predicted GPA at mean score for males (females)		
black		−0.55 (−0.37)
Hispanic		−0.31 (−0.13)
Asian		0.25 (0.43)
white		0.08 (0.26)
Predicted GPA at 1 SD above mean for males (females)		
black		−0.20 (−0.02)
Hispanic		0.14 (0.32)
Asian		0.89 (1.07)
white		0.50 (0.69)

Notes. **p*<.05, ***p*<.01, ****p*<.001 (two-tailed tests)

Top panel: Linearized standard errors in parentheses

Predicted GPA is standardized

Ever Dropped Out

Table 7.2a reports logit estimates and Table 7.3 reports predicted probabilities that a student ever dropped out of high school, by race and sex. For black males, the predicted probability of ever dropping out of high school decreased from 21.5 percent for students scoring at the mean test score to 10.4 percent for students scoring one SD above the mean. For Hispanic males, the predicted probability of ever dropping out of high school decreased from 20.1 percent for students scoring at the mean test score to 7.7 percent for students scoring one SD above the mean. The interpretation is that raising black student test scores from the mean to one SD above the mean is predicted to reduce the number of black male students who ever drop out of high school by 51.6 percent; the corresponding reduction for Hispanic male students is 61.7 percent.

Table 7.2a Logit Estimates for Selected Outcomes

	(1a)	*(1b)*	*(2a)*	*(2b)*	*(3a)*	*(3b)*	*(4a)*	*(4b)*
	Dropped Out	*Dropped Out*	*Attended 4-Year*	*Attended 4-Year*	*Attained BA*	*Attained BA*	*Rigorous Curriculum*	*Rigorous Curriculum*
SES	−0.65***	−0.58***	0.90***	0.91***	0.93***	0.93***	0.51	0.59*
	(0.11)	(0.11)	(0.08)	(0.08)	(0.07)	(0.07)	(0.28)	(0.27)
test scores	−0.70***	−0.69***	0.86***	0.86***	0.72***	0.71***	1.54***	1.98***
	(0.09)	(0.10)	(0.06)	(0.06)	(0.08)	(0.08)	(0.35)	(0.22)
GPA	−0.98***	−0.93***	0.50***	0.50***	0.59***	0.58***	0.82	
	(0.09)	(0.09)	(0.07)	(0.07)	(0.11)	(0.11)	(0.48)	
Asian	−1.39***	−1.10***	0.70*	0.71*	0.37		0.89*	0.92*
	(0.32)	(0.31)	(0.33)	(0.33)	(0.21)		(0.42)	(0.40)
Hispanic	−0.48*		0.64***	0.63***	0.10		−0.70	
	(0.23)		(0.18)	(0.18)	(0.16)		(0.65)	
black	−0.49		0.65**	0.63**	0.08		1.76*	1.81*
	(0.32)		(0.21)	(0.20)	(0.19)		(0.88)	(0.86)
AmIndian/Alaskan	0.15		−0.35		−0.80		[a]	
	(0.81)		(0.40)		(0.44)			
sex	−0.00		−0.08		0.35***	0.35***	−.63	
	(0.14)		(0.10)		(0.10)	(0.10)	(0.44)	
northcentral	−0.16		−0.39*	−0.26*	−0.31*		−1.33*	−1.32*
	(0.24)		(0.17)	(0.12)	(0.12)		(0.59)	(0.53)
south	−0.19		−0.21		−0.21		0.41	
	(0.26)		(0.17)		(0.13)		(0.40)	
west	0.33		−0.97***	−0.84***	−0.68***	−0.45**	−0.38	
	(0.27)		(0.18)	(0.14)	(0.16)	(0.14)	(0.50)	
constant	−2.29***	−2.39***	0.93***	0.75***	−0.84***	−1.01***	−6.82***	−6.88***
n	9636	9652	7988	7988	7959	7960	9535	9648

Notes. *p<.05, **p<.01, ***p<.001 (two-tailed tests)

Linearized standard errors in parentheses

Neither R-squared nor pseudo-r-squared is computed with logistic regression using survey data

[a] No AmIndian/Alaskan student experienced a highly-rigorous curriculum; these cases were dropped

Table 7.2b Logit Estimates for Selected Outcomes

	(5a)	*(5b)*	*(6a)*	*(6b)*	*(7a)*	*(7b)*	*(8a)*	*(8b)*
	Took Calculus	*Took Calculus*	*Took Algebra 2*	*Took Algebra 2*	*UGPA > 3.75*	*UGPA > 3.75*	*PhD or Prof'l*	*PhD or Prof'l*
SES	0.37***	0.38***	0.31***	0.30***	–0.17*	–0.16*	0.70***	0.72***
	(0.08)	(0.08)	(0.07)	(0.07)	(0.07)	(0.07)	(0.12)	(0.12)
test scores	2.09***	2.20***	0.85***	0.85***	0.36***	0.44***	0.80***	0.90***
	(0.18)	(0.15)	(0.07)	(0.06)	(0.09)		(0.15)	(0.09)
GPA	0.26		0.53***	0.53***	0.20		0.20	
	(0.18)		(0.07)	(0.07)	(0.11)		(0.20)	
Asian	0.77***	0.70***	0.37		0.15		0.60**	0.63**
	(0.19)	(0.18)	(0.28)		(0.31)		(0.21)	(0.21)
Hispanic	0.38		0.44*	0.43*	–0.54**	–0.54***	0.88**	0.88**
	(0.43)		(0.18)	(0.18)	(0.16)	(0.15)	(0.32)	(0.31)
black	0.82**	0.85**	0.11		–0.69*	–0.75**	0.79**	0.78**
	(0.32)	(0.32)	(0.20)		(0.28)	(0.28)	(0.26)	(0.26)
AmIndian/Alaskan	–0.48		–0.55		0.69		–0.23	
	(0.67)		(0.38)		(0.85)		(0.74)	
sex	–0.14		–0.02		0.29*	0.34**	0.01	
	(0.14)		(0.09)		(0.12)	(0.12)	(0.13)	
northcentral	–0.60**	–0.51**	0.26		0.06		–0.00	
	(0.20)	(0.17)	(0.18)		(0.17)		(0.21)	
south	0.12		0.70***	0.54***	0.09		0.44*	0.41*
	(0.17)		(0.18)	(0.11)	(0.19)		(0.20)	(0.16)
west	–0.40		0.19		0.10		–0.02	
	(0.24)		(0.19)		(0.20)		(0.23)	
constant	–3.72***	–3.79***	–0.49**	–0.31***	–2.10***	–1.99***	–4.30***	–4.26***
n	9636	9648	9636	9636	7315	7329	9636	9648

Notes. *$p<.05$, **$p<.01$, ***$p<.001$ (two-tailed tests)

Linearized standard errors in parentheses

Neither R-squared nor pseudo-r-squared is computed with logistic regression using survey data

Table 7.3 Predicted Probabilities at Mean Test Scores and 1 SD Above Mean Test Scores, by Race and Sex

	(1b)	*(2b)*	*(3b)*	*(4b)*	*(5b)*	*(6b)*	*(7b)*	*(8b)*
	Dropped Out	*Attended 4-Year*	*Attained BA*	*Rigorous Curriculum*	*Took Calculus*	*Took Algebra2*	*UGPA > 3.75*	*PhD or Prof'l*
Predicted percent at mean scores for males (females)								
black	21.5 (18.8)	57.6 (59.8)	11.4 (16.8)	0.17 (0.17)	1.37 (1.37)	23.6 (25.4)	5.1 (7.0)	1.4 (1.4)
Hispanic	20.1 (17.5)	55.1 (57.4)	10.5 (15.6)	0.02 (0.02)	0.55 (0.55)	33.4 (35.6)	6.5 (8.8)	1.3 (1.3)
Asian	1.9 (1.6)	86.9 (87.9)	36.0 (47.0)	0.38 (0.38)	6.33 (6.33)	50.0 (52.4)	12.3 (16.4)	3.4 (3.4)
white	6.7 (5.7)	73.6 (75.3)	32.0 (42.5)	0.15 (0.15)	3.29 (3.29)	47.5 (50.0)	12.6 (16.8)	1.7 (1.7)
Predicted percent at 1 SD above mean for males (females)								
black	10.4 (8.9)	76.1 (77.7)	21.5 (30.1)	0.80 (0.80)	7.1 (7.1)	41.8 (44.2)	7.0 (9.5)	2.8 (2.8)
Hispanic	7.7 (6.6)	78.6 (80.1)	23.7 (32.8)	0.18 (0.18)	4.8 (4.8)	59.9 (62.2)	9.7 (13.1)	3.2 (3.2)
Asian	0.4 (0.3)	96.9 (97.2)	69.0 (77.8)	5.93 (5.93)	60.3 (60.3)	82.3 (83.7)	20.7 (26.8)	11.1 (11.1)
white	2.5 (2.1)	88.7 (89.6)	54.1 (65.0)	0.98 (0.98)	21.5 (21.5)	71.7 (73.6)	18.0 (23.5)	4.0 (4.0)

Ever Attended a Four-Year Institution

Table 7.2a reports logit estimates and Table 7.3 reports predicted probabilities that a student ever attended a four-year postsecondary institution, by race and sex. For black males, the predicted probability of ever attending a four-year institution increased from 57.6 percent for students scoring at the mean test score to 76.1 percent for students scoring one SD above the mean. For Hispanic males, the predicted probability of ever attending a four-year institution increased from 55.1 percent for students scoring at the mean test score to 78.6 percent for students scoring one SD above the mean. The interpretation is that raising black student test scores from the mean to one SD above the mean is predicted to increase the number of black male students who ever attend a four-year institution by 32.1 percent; the corresponding increase for Hispanic male students is 42.6 percent.

Attained BA Degree

Table 7.2a reports logit estimates and Table 7.3 reports predicted probabilities that a student attained a baccalaureate degree, by race and sex. For black males, the predicted probability of attaining a baccalaureate degree increased from 11.4 percent for students scoring at the mean test score to 21.5 percent for students scoring one SD above the mean. For Hispanic males, the predicted probability of attaining a baccalaureate degree increased from 10.5 percent for students scoring at the mean test score to 23.7 percent for students scoring one SD above the mean. The interpretation is that raising black student test scores from the mean to one SD above the mean is predicted to increase the number of black male students who attain a baccalaureate degree by 88.6 percent; the corresponding increase for Hispanic male students is 125.7 percent.

Highly-Rigorous High School Curriculum

Table 7.2a reports logit estimates and Table 7.3 reports predicted probabilities that a student completed a highly-rigorous high school curriculum, by race and sex. "Highly rigorous" was defined as a minimum curriculum including no less than four years of English, three years of math, three years of science, three years of social science, two years of a foreign language, one or more Advanced Placement (AP) test scores in any subject, and all of the following courses: precalculus, biology, chemistry, and physics. For black males, the predicted probability increased from 0.17 percent for students scoring at the mean test score to 0.80 percent for students scoring one standard deviation above the mean. For Hispanic males, the predicted probability increased from 0.02 percent for students scoring at the mean to 0.18 percent for students scoring one SD above the mean. The interpretation is that raising black student test scores from the mean to one SD above the mean is predicted to increase the number of black male students who complete a highly-rigorous high school curriculum by 370.6 percent; the corresponding increase for Hispanic male students is 800 percent.

Completed Calculus

Table 7.2b reports logit estimates and Table 7.3 reports predicted probabilities that students had completed calculus in high school, by race and sex. For black males, the predicted probability increased from 1.37 percent for students scoring at the mean test score to 7.1 percent for students scoring one SD above the mean. For Hispanic males, the predicted probability increased from 0.55 percent for students scoring at the mean to 4.8 percent for students scoring one SD above the mean. The interpretation is that raising black male student test scores from the mean to one SD above the mean is predicted to increase the number of black male students who complete calculus in high school by 418.2 percent; the corresponding increase for Hispanic male students is 772.7 percent.

Completed Algebra 2

Table 7.2b reports logit estimates and Table 7.3 reports predicted probabilities that students had completed algebra 2 in high school, by race and sex. For black males, the predicted probability increased from 23.6 percent for students scoring at the mean test score to 41.8 percent for students scoring one SD above the mean. For Hispanic males, the predicted probability increased from 33.4 percent for students scoring at the mean to 59.9 percent for students scoring one SD above the mean. The interpretation is that raising black male student test scores from the mean to one SD above the mean is predicted to increase the number of black male students who completed algebra 2 in high school by 77.1 percent; the corresponding increase for Hispanic male students is 79.3 percent.

Attained 3.75 GPA

Table 7.2b reports logit estimates and Table 7.3 reports predicted probabilities that a student had compiled a cumulative undergraduate GPA above 3.75, by race and sex. For black males, the predicted probability increased from 5.1 percent for students scoring at the mean test score to 7.0 percent for students scoring one SD above the mean. For Hispanic males, the predicted probability increased from 6.5 percent for students scoring at the mean to 9.7 percent for students scoring one SD above the mean. The interpretation is that raising black male student test scores from the mean to one SD above the mean is predicted to increase the number of black males who compile a cumulative undergraduate GPA above 3.75 by 37.3 percent; the corresponding increase for Hispanic male students is 49.2 percent.

Aspired to PhD or Professional Degree

Table 7.2b reports logit estimates and Table 7.3 reports predicted probabilities that a student aspired to a doctorate or a professional degree, by race and sex. For black males, the predicted probability increased from 1.4 percent for students scoring at the mean test score to 2.8 percent for students scoring one SD above the mean. For Hispanic males, the predicted probability increased from 1.3 percent for students scoring at the mean to 3.2 percent for students scoring one SD above the mean. The interpretation is that raising black male student test scores from the mean to one SD above the mean is predicted to increase the number of black males who aspire to a doctorate or a professional degree by 100.0 percent; the corresponding increase for Hispanic males is 146.2 percent.

A Large Impact

The results of the study predict that raising student test scores from the mean score in each racial group to scores that are one SD above the mean in each group would increase the number of black male students who attain baccalaureate degrees by 88.6 percent and the number of Hispanic male students who attain baccalaureate degrees by 125.7 percent. The number of black female students who attain baccalaureate degrees would increase by 79.2 percent and the number of Hispanic female students who attain baccalaureate degrees would increase by 110.3 percent. The results predict that black and Hispanic students would be less likely to drop out of high school, much more likely to complete a highly-rigorous high school curriculum and to complete algebra 2 and calculus in high school, more likely to attend a four-year institution, more likely to compile an undergraduate GPA above 3.75, and much more likely to aspire to a doctorate or professional degree. Thus, an intervention that raises student achievement by one SD is predicted to have a large impact on black and Hispanic students.

A question that arises is whether the predictions would change if the ability of disadvantaged students to afford the cost of college is incorporated. However, while net tuition and fees at four-year public institutions increased significantly for full-time dependent students in the top half of the income distribution, these costs barely increased for students in the bottom half of the income distribution, from $893 in 1999-2000 to $1,163 in 2011-2012.[19] For students in the bottom quartile of the income distribution, net tuition and fees remained at zero from 1999-2000 through 2011-2012.[20] For students in the bottom half of the income distribution, the net cost of attendance, including living costs, increased modestly from $11,059 in 1999-2000 to $13,843 in 2011-2012: an increase of $2,784 over the 12-year period.[21] For students in the bottom quartile of the income distribution, the net cost of attendance increased only $2,234 over the same period.[22]

Thus, while list prices have increased, those prices are not what well-prepared black and Hispanic students pay. These students are highly sought by selective colleges that offer scholarships and grants to attract qualified minorities. As a consequence, the average net price of tuition for full-time, dependent, in-state freshmen in the bottom half of the income distribution at 20 prestigious state flagship universities, including the University of California-Los Angeles, the University of Wisconsin-Madison, the University of Illinois at

Urbana-Champaign, and the University of Minnesota-Twin Cities was negative $1,570; in other words, these students received more grant money than they paid in tuition, offsetting their living costs.[23] The average net price of tuition for full-time, dependent, in-state freshmen in the bottom half of the income distribution at 15 less-selective state universities, including the University of North Carolina-Charlotte, Appalachian State University, George Mason University, and Virginia Commonwealth University was negative $1,280.[24]

While listed tuition and fees at private colleges seemingly place them out of reach of low- and middle-income students, net prices are much more affordable. The average net cost of attendance in 2012-2013, including tuition, fees, room and board, after taking into account federal, state, and institutional financial aid for students who come from households earning between $30,000 and $48,000 a year and qualifying for federal aid, was $3,000 at Harvard, $3,500 at Columbia, and $4,300 at Stanford.[25] Students raised in households earning less would pay even less than those amounts.

The actual tuition paid by well-prepared minorities is typically a fraction of the list price. As a result, it appears that students are not being forced to enroll in inexpensive colleges that are inappropriate for their level of preparedness.[26] Instead, it appears that students from high- and medium-high-income families who have low SAT scores and high school grades are being replaced by highly-prepared students from low-income families.[27] Less than eight percent of all students are prevented from enrolling by their inability to pay[28] and federal Pell grants have no significant impact on enrollment,[29] leading Bowen and his colleagues to conclude that "family finances have a fairly minor direct impact on a student's ability to attend a college."[30] Math and verbal SAT scores "are much more important factors in the college [application] process than financial variables such as family income."[31] Therefore, it appears that high tuition is not what prevents minorities from attaining baccalaureate degrees. Instead, the results reported here suggest that low average test scores significantly depress rates of attainment. If the test scores of minorities are raised by one SD, it appears that a much larger number of minorities would qualify for selective colleges, accelerating the substitution of highly-qualified minorities for white students.

The significance of the results reported here is that they underline the importance of swiftly adopting the most efficient approaches for raising student achievement in the K-12 grades—well before students enroll in college. A one SD increase in test scores would boost the average performance of every school, creating more high-performing schools and permitting more students—including more minority students—to experience the benefits of attending high-performing schools. In addition, students would also benefit from the improvement in educational outcomes that would occur whether students attend a high- or low-performing school.

The results suggest that when underprepared students enroll in college, their prospects for success are greatly diminished. However, the results should not be interpreted to imply that college admissions officers should emphasize test scores when they select students, nor should they be interpreted to imply that there is no room for colleges to improve baccalaureate attainment rates. Instead, the policy implication is that efforts to identify and adopt efficient approaches for raising student achievement should be redoubled because the payoff appears to be large.

This conclusion may seem unsurprising. The call to raise student achievement has been sounded for at least 30 years, since the publication of *A Nation At Risk: The Imperative of Educational Reform*, by the National Commission on Excellence in Education.[32] The report issued a call to improve what the Commission viewed as the mediocre level of educational performance. In response, enormous effort has been invested over the past 30 years toward the goal of identifying effective approaches for raising student achievement. The Institute of Education Sciences alone budgets over $671 million annually for education research.[33]

What remains unclear is whether this effort has been productive. While there has been some improvement with regard to trends in achievement by 9-year-old and 13-year-old students, the achievement of 17-year-old students has remained flat over the past 40 years.[34] If existing educational strategies have been unsuccessful, perhaps a new strategy is required.

Rapid performance feedback is a strategy that has largely been ignored, yet potentially offers large gains. A review of research regarding feedback found an average effect size of 0.79 SD.[35] The results suggest that feedback is most effective when it is nonjudgmental, involving frequent testing (two-to-five times per week), and presented immediately after a test. Under these conditions, the meta-analyses and reviews of feedback interventions suggest that the effect size for testing feedback is no lower than 0.7 SD,[36] equivalent to raising the achievement of an average nation such as the United States to the level of the top five nations.[37] When teachers were required to follow rules about using the assessment information to change instruction for students, the average effect size exceeded 0.9 SD, and when students were reinforced with material tokens, in addition to the frequent testing, the average effect size increased even further, exceeding 1.1 SD.[38] Emotionally neutral (i.e., testing) feedback that is void of praise or criticism "is likely to yield impressive gains in performance, possibly exceeding 1 SD."[39] Lysakowski and Walberg[40] reported an effect size of 1.13, Walberg[41] reported an effect size of 0.82, and Tenenbaum and Goldring[42] reported an effect size of 0.74, all of which are substantial effects.

These effect sizes were typically obtained over periods of one year or less. Presumably, the implementation of rapid performance feedback throughout the entire academic careers of students, from kindergarten through twelfth grade, would result in even larger effect sizes. The studies suggest that more efficient approaches for raising student achievement are available. To the extent that existing approaches for raising student achievement are relatively unproductive, inefficient, and disproportionately affect minority students; current policies may serve to perpetuate the disadvantaged status of minorities. The results reported in this chapter suggest that attention to these issues should receive high priority.

CHAPTER 8

No-Excuses Charter Schools

The No Child Left Behind Act created enormous pressure to find effective ways to raise student achievement and close the achievement gap between economically-disadvantaged students and their more privileged peers. Many districts have struggled to raise achievement, and national trends are discouraging. However, dramatic results from evaluations regarding the Knowledge is Power Program (KIPP) and the Harlem Children's Zone (HCZ) charter school programs have raised hope that these types of reform-oriented charter schools may be especially effective in raising student achievement. A rigorous evaluation concluded that "the effects of attending an HCZ middle school are enough to close the black-white achievement gap in mathematics. The effects in elementary school are large enough to close the racial achievement gap in both mathematics and English language arts."[1] Similarly, the best available evaluation of KIPP concluded that "estimated impacts are frequently large enough to substantially reduce race- and income-based achievement gaps within three years of entering KIPP."[2] Significantly, the federal government's Race-to-the-Top initiative emphasizes expanded implementation of these types of charter schools to address the achievement gap.

The Knowledge is Power Program (KIPP), begun in 1994, comprises a national network of charter schools that aim to equip students, who are drawn primarily from low-income and minority families, with the knowledge, skills, and character traits needed to succeed in top-quality high schools and colleges. KIPP emphasizes high expectations and standards for students, and seeks to achieve its goals by recruiting highly-dedicated teachers who are willing to work long hours.

The Harlem Children's Zone (HCZ) is a nonprofit organization that funds and operates a neighborhood-based system of education and social services in a 97-block area in central Harlem, New York. HCZ combines reform-minded charter schools with a web of community services created to support children from birth through college. HCZ charter schools are similar to KIPP schools in emphasizing high expectations and standards for students and reliance on the recruitment of highly-dedicated teachers to improve student achievement.

S.S. Yeh, *Solving the Achievement Gap*,
DOI 10.1057/978-1-137-58767-1_8

Thus, a core assumption of both charter school approaches is that raising student achievement requires highly-dedicated teachers who are willing to work long hours. However, both approaches exhibit high teacher attrition. To scale up these approaches, a secondary assumption is required: there must be a sufficient supply of highly-dedicated teachers to fill the large number of vacant teaching positions that arise as a consequence of these models. This assumption appears to be incorrect since principals of "no-excuses" charter schools such as KIPP and HCZ report that they have to "scour the country" for suitable teachers.[3]

The analysis presented in this chapter indicates that the core assumption, in combination with empirical evidence regarding high teacher attrition and evidence that principals have to scour the country for suitable teachers, implies an internal contradiction: if this assumption is correct, it is *logically impossible* to replicate the impact results on a national level, given the high level of teacher attrition. While other researchers have suggested that the HCZ and KIPP models may be *difficult* to scale up,[4] the contribution of the analysis presented here is that it demonstrates the existence of an internal contradiction that would effectively *prevent* the models from being scaled up.

The distinction between a *difficult* and a *logically impossible* task is important. If it is logically impossible to scale up KIPP and HCZ, the task should not be attempted, because it would divert scarce resources from improvement strategies that are feasible and worthy. If it is logically feasible to scale up KIPP and HCZ, then the effort may be worthwhile. However, the analysis reported here suggests reasons why scaling appears to be impractical. Unless and until KIPP/HCZ supporters can demonstrate that the analysis is incorrect, the appropriate policy conclusion is that scale-up should be delayed until research has been performed that would substantiate the secondary assumption that there is a sufficient supply of highly-dedicated teachers to fill the large number of vacant teaching positions that arise as a consequence of these models.

While the validity of the current analysis depends on the models' core assumption, the assumption is derived from statements by the founders of the HCZ and KIPP models, as described below in the first section of this chapter. If it can be shown that any intervention based on this assumption, in combination with empirical evidence of high teacher attrition, cannot be scaled up in a way that maintains the putative impacts of the HCZ/KIPP models, then serious questions arise about the external validity of the models. Either the assumption is correct, and the models cannot be scaled up nationally, or the assumption is incorrect and the models are based on an incorrect assumption.

The analysis proceeds in two parts. The first part of this chapter reviews the results of the key impact studies and empirical evidence of high teacher attrition, draws upon statements by the founders of the HCZ and KIPP models articulating the core assumption, and presents a narrative explanation of the internal contradiction in these models. This analysis suggests why the impact results are likely due to artifacts stemming from high teacher attrition and hoarding of a disproportionate share of the nation's limited pool of highly-dedicated teachers, rather than gains that could be sustained when the programs are scaled up and implemented nationwide. The second part of this chapter formalizes the analysis. Under the assumption that any gains depend on the proportion of highly-dedicated teachers, the gains would fall to zero once these programs are implemented in every school across the nation. The contribution of the formal analysis presented in part two is that it demonstrates that neither HCZ nor KIPP can be scaled up nationally while maintaining their putative impacts on student achievement, given the founders' core assumption, empirical evidence regarding high teacher attrition, and evidence that the characteristics of HCZ and KIPP are insufficient to generate the waiting list of highly-dedicated teachers that would be necessary for scale-up—the analysis suggests that no other assumption or evidence is required to reach this conclusion. To refute the analysis, it would be necessary to show that additional assumptions or empirical evidence are required. The final section of this chapter discusses this result.

The purpose of the analysis presented in this chapter is to investigate the possibility of an internal contradiction in the assumptions underlying the KIPP and HCZ models, given the available evidence of high teacher attrition. This possibility suggests the need for extreme caution in extrapolating the results of the impact studies of KIPP and HCZ. The purpose of the analysis is not to resolve empirical questions about the level of teacher attrition, or whether teachers who stay are truly more dedicated or more effective than teachers who leave, or the reasons why teachers leave the KIPP and HCZ schools.

A Threat to Validity

Researchers who evaluate the effectiveness of various interventions are well aware that sample attrition can lead to false conclusions that an intervention is effective. Attrition occurs when, for example, a greater number of less-motivated students drop out of a treatment group, compared to a control group. What is not widely-recognized is that a similar threat to the validity of a research study may occur when an intervention causes less-motivated teachers to drop out, thereby raising the measured performance of the remaining group of teachers and the putative effectiveness of any intervention based on that performance.[5] If student achievement depends on having motivated teachers, and if an intervention causes less-motivated teachers to drop out, then it necessarily follows that the measured performance of the remaining group of teachers will be higher, resulting in the false conclusion that the intervention caused student achievement to improve. The conclusion is false because the intervention did not cause student achievement to improve; instead, the improvement is an artifact of teacher attrition.

At first glance, it may seem desirable if a treatment causes less-motivated teachers to drop out. If it is always possible to substitute highly-motivated teachers for less-motivated counterparts, then this phenomenon could be considered a desirable feature of the intervention, and the measured impact of the intervention is a valid measure of program impact. However, a problem arises if the intervention relies entirely on teacher attrition to improve student performance and there are insufficient teacher candidates who meet the selection criterion and are available to fill the teaching slots that become available as a consequence of the treatment's high attrition rate. This may not be a problem when the intervention has only been implemented in a few scattered schools across the nation because these schools are free to pull highly-motivated teachers from neighboring schools and districts, as well as distant schools from across the nation. These schools may also pull a disproportionate number of the limited supply of highly-dedicated college graduates seeking to become teachers. Significantly, principals of "no-excuses" charter schools such as KIPP and HCZ report that they have to "scour the country" for suitable teachers, which suggests that the model of teacher recruiting employed by these charter schools may not be scalable.[6]

This siphoning effect may be hidden in the large flows of teachers from school to school and district to district across the nation every year. Typically, researchers do not pay attention to this effect and do not observe or measure it. They focus on controlling bias in the selection of students, rather than bias in the retention of highly-dedicated teachers. When researchers focus on attrition, they focus on students, not the attrition of less-dedicated teachers.[7]

Recently, several research studies have evaluated the impact on student achievement of the Knowledge is Power Program (KIPP) and the Harlem Children's Zone (HCZ). These studies sought to control for sample bias, found positive impacts on student achievement, and concluded that the interventions are promising.[8] However, these studies did not examine the possibility that the results may be artifacts of high teacher attrition.

Research on HCZ

Dobbie and Fryer provided the first and perhaps the most highly regarded empirical test of the causal impact of HCZ on student achievement. Using lottery-based randomization and two-stage least squares (2SLS), they found that Promise Academy middle-school lottery winners gained 0.229 SD per year in math and 0.047 SD per year in English language arts, compared to students who did not win admission.[9] Promise Academy elementary school lottery winners gained 0.191 SD per year in math and 0.095 SD per year in English language arts.[10] Significantly, while HCZ includes a variety of social services in addition to the charter school component, a Brookings Institution analysis found no evidence that HCZ influences student achievement through those social services, suggesting that the impact of HCZ on student achievement results primarily from the influence of the HCZ charter schools.[11]

The Promise Academy schools emphasize the recruitment and retention of high-quality teachers in order to raise student achievement, and use a test-score value-added measure to incentivize and evaluate current teachers.[12] The view that teacher quality and teacher motivation are paramount is clear from an interview with Geoffrey Canada, the founder of the HCZ Promise Academies, on the CBS news program *60 Minutes*. Canada stated that he would “fire the teachers” if they do not raise student achievement to a level where students are college-ready.[13] In the documentary, “Waiting for Superman,” he stated, “I want to be able to get rid of teachers that we know aren’t able to teach kids.”[14] Elsewhere, he has stated that “Finding great teachers is the ‘secret sauce’ of great schools and, in particular, great charter schools.”[15] This indicates that the theory of action underlying the Academies is that teachers are the main factor influencing student achievement and low student achievement is due to poor teachers. If, instead, Canada believed that the most important factor was the length of the school day and the school year, or the level of resources, or any factor other than the effort and quality of the teachers, it would make more sense to simply address those other factors. It would not make sense to threaten to fire teachers if the most important factors are beyond their control.

At present, it is not possible to determine whether the impact of the Academies on student achievement is due primarily to the recruitment of highly-dedicated teachers. This type of analysis would require a large sample of HCZ schools and a statistical analysis of all of the factors influencing student achievement. The limited number of HCZ schools makes this analysis impossible. However, it is possible to test the *implications* of Canada’s assumption that finding great teachers is the “secret sauce” of great charter schools. The formal analysis, below, suggests that this assumption contains an internal contradiction—if Canada is correct, then the impacts identified by Dobbie and Fryer cannot be replicated when the intervention is scaled up. While Canada’s analysis may be faulty, and it is possible that the impacts are due to factors other than highly-dedicated teachers, serious questions arise about the viability of the intervention if it can be demonstrated that the founder’s most important assumption implies that impacts fall toward zero when the intervention is scaled up nationally. Either the assumption is correct, and the HCZ model cannot be scaled up nationally, or the assumption is incorrect and the model is based on an incorrect assumption.

Canada’s assumption, however, is fully consistent with the available evidence. Highly-dedicated teachers are needed because the HCZ Promise Academy approach requires that teachers work long hours. HCZ Promise Academy students who are performing below grade level attend school for approximately twice as many hours as a traditional public school student in New York City.[16] As a consequence, HCZ teachers work several hundred more hours than regular New York City teachers. These extra hours are accumulated through a longer work day, a lengthened school year and, for many teachers, a summer school session that is added to the school year and runs through the first week of August.[17] Perhaps as a consequence, both schools have had high teacher turnover as they search for highly-dedicated teachers willing to work long hours: 48 percent of Promise Academy teachers did not return for the 2005 – 2006 school year, 32 percent left before 2006 – 2007, and 14 percent left before 2007 – 2008, suggesting very high three-year attrition rates that almost certainly exceed 48 percent.[18] The decline in the rate of teacher turnover in each subsequent year is consistent with the hypothesis that turnover primarily serves to rid the school of less-dedicated teachers. As the proportion of highly-dedicated teachers increases each year, the proportion of less-dedicated teachers declines, resulting in a decrease in turnover during each subsequent year. Thus, the cumulative three-year teacher attrition rate is a more valid indicator of the degree of teacher replacement that is occurring than any single-year attrition figure.

Qualitative evidence suggests that highly-dedicated teachers are needed for a second reason, in addition to long hours. The Promise Academies require HCZ teachers to undertake strenuous efforts to raise student achievement. According to Terri Grey, the former principal at the Promise Academy middle school, the HCZ approach requires not only long hours but an emphasis on test preparation, leading to teacher burnout and attrition.[19] Opposed to the heavy emphasis on test preparation, she was fired.[20] As recounted by Paul Tough in his book about HCZ:

> Test prep was under way by the third week in September. There were morning test-prep sessions, a test-prep block during the school day, test prep in the afterschool program, and test prep on Saturdays…As the year went on, the time dedicated to test prep only grew, and the time dedicated to everything else was forced to shrink further.[21]

A month later superintendent Doreen Land quit and more than a dozen teachers followed Grey and Land out the door.[22] However, despite all of the test-prep, the HCZ board of directors feared that the eighth-grade cohort of students would perform poorly on standardized tests. In order to protect the reputation of the Harlem Children's Zone, the cohort was disbanded and all of the students in the cohort were reassigned to other schools.[23] This depressed teacher morale. Chris Finn, the dean of students, reported that the experience was exhausting and engendered a feeling of failure.[24] In sum, the evidence suggests that HCZ relies heavily on highly-dedicated teachers who are willing to work long hours, endure conditions that cause less-dedicated teachers to quit, and undertake strenuous efforts to raise student achievement. Teachers who are unable to maintain the exhausting pace either quit or are fired.

Research on KIPP

Studies of KIPP suggest a parallel narrative: promising impact results as a consequence of highly-dedicated teachers working long hours under exhausting conditions, resulting in high teacher attrition. While early evaluations of KIPP were limited to a small number of schools or included only weak controls for selection bias, Tuttle, Teh, Nichols-Barrer, and their colleagues used a matched comparison group design with achievement data for a nationwide sample of 22 KIPP schools and found that, after three years, students in half of the KIPP schools gained at least 0.16 SD per year in math.[25] Similarly, students in half of the KIPP schools gained at least 0.09 SD per year in reading.[26]

The researchers did not examine teacher attrition. However, an SRI study investigated teacher attrition at five San Francisco Bay Area KIPP schools and found that among the 84 teachers who taught in the five schools in 2006-07, nearly half (49 percent) left the classroom before the start of the 2007-08 school year.[27] In the spring of 2007, the median tenure of teachers at the Bay Area KIPP schools was only two years.[28]

Thus, evaluations of both HCZ and KIPP suggest positive impacts on student achievement, but at the cost of teacher attrition that ranges up to 49 percent each year as a consequence of the "heroic" efforts that are required of teachers.[29] While the SRI report points out that high teacher turnover is not uncommon in urban schools serving poor and minority students, studies suggest that the annual turnover rate in these schools is substantially lower than attrition at the HCZ and KIPP schools—about 20 percent.[30] The SRI study is useful because it explores the reasons for KIPP teacher attrition in detail. While SRI studied the KIPP schools, the explanation may also apply to the HCZ schools as well because both approaches rely on highly-dedicated teachers working long hours. The SRI study suggests that the long work hours and personal sacrifices demanded by this type of approach cause high teacher attrition. Excerpts from the interviews conducted by SRI with KIPP teachers and school leaders included the following:

> "I can't do this job very much longer. It is too much. I don't see any solution… No one has really presented any way to solve that problem."[31]
>
> "The big question is the sustainability question…We are really tired."[32]
>
> "You're taking on the place of the family, giving up your own family. I respect and admire that in others, and I don't know that I can do that again."[33]
>
> "Turnover is so high that teachers are constantly coming in and reinventing the wheel."[34]
>
> "That's the biggest KIPP challenge: how do you keep teachers coming back here year after year? A lot of the workload I have I put on myself…When do I stop worrying about them and take care of me? It's hard to find that balance. That's going to be the most challenging thing, retaining teachers and keeping them rested and healthy."[35]

According to the SRI report, veteran KIPP teachers in every school, including the founders, expressed similar sentiments, and regret that they need to choose between teaching at KIPP and finding balance in their lives.[36] Bay Area KIPP teachers spent a median 65 hours per week on all school-related activities (a range of 60 to 67 hours), whereas urban middle-school teachers worked an average of 52 hours per week nationally.[37]

At KIPP, teachers carry cell phones and are expected to be available 24 hours a day to respond to any concerns that students may have.[38] Nine-and-a-half-hour days, class on Saturday, and school during the summer "are all non-negotiable."[39]According to one of the co-founders, the whole KIPP framework is built around maximizing teaching time and teacher accountability.[40] KIPP teachers are personally held accountable for student progress and are "contractually obligated to see that their students succeed. They know they *have* to teach until the kids get it."[41] KIPP teachers sign commitments to do "whatever it takes" to get students to learn.[42] KIPP teachers regularly visit students in their homes to teach parents the importance of checking homework, reading with their children, and fostering aspirations to attend college.[43] Both co-founders believe that it is impossible to scale up and replicate the KIPP model on a national scale, given the current pool of teachers.[44] "What we do isn't easy. First, we need to find a way to make this level of commitment the standard. Then we need to make it attractive, livable, and affordable for teachers."[45]

A Formal Analysis

A formal analysis may be conducted regarding the implications of Canada's assumption that the impact of charter schools such as HCZ and KIPP depends on the recruitment of highly-dedicated teachers. This analysis indicates that the gains of HCZ and KIPP would fall to zero once these programs are implemented in every school across the nation.

Underlying both KIPP and HCZ is the core assumption that the recruitment and retention of a highly-dedicated teaching staff is central to student achievement. Let f[x] describe this theory of student achievement, where f[x] is a linear, monotonically increasing function that depends on the proportion (a) of the teaching staff (S) that is composed of highly-dedicated teachers:

$$f\left[x\right]=f\left[aS\right], f'\left[x\right]>0$$

If f[x] = f[aS] is assumed to be linear, then f[aS] = kaS, for some k > 0. If the units of achievement are chosen such that k = 1, then student achievement = f[aS] = aS.

Let b = the proportion of highly-dedicated teachers before implementation of KIPP or HCZ

Let c = the increase in the proportion of highly-dedicated teachers after implementation of KIPP or HCZ (pulled from non-KIPP/HCZ schools or college graduates who would otherwise be hired by non-KIPP/HCZ schools), where:

$$a=b+c$$

and

$$f\left[(b+c)S\right]>f\left[bS\right]$$

Then the initial level of achievement is

$$bS$$

before implementation of KIPP or HCZ. Achievement increases to

$$(b+c)S$$

after implementation of KIPP or HCZ.

Thus, student learning increases by an amount equal to

$$(b+c)S-bS=cS$$

as a consequence of implementing KIPP or HCZ.

If, however, every school implements KIPP or HCZ, the value of (c) must (by definition) equal zero ($c = 0$). The proportion of the teaching staff that is highly-dedicated must equal (b), in other words, $a = b$, and the increase in student achievement as a consequence of implementing KIPP or HCZ must fall to zero:

$$(b+c)S-bS=bS-bS=0$$

The key to understanding this result is that once every school is a KIPP or HCZ school, it is no longer possible for any school to pull teachers from non-KIPP/HCZ schools: recruitment necessarily pulls from the supply of highly-dedicated teachers to other KIPP/HCZ schools, reducing the performance of the other schools. In essence, recruitment becomes a zero-sum game in which any single KIPP/HCZ school can only recruit additional highly-dedicated teachers if some other KIPP/HCZ school loses highly-dedicated teachers.[46] Take, for example, three schools, each with a number of highly-dedicated teachers equal to (bS). School 1 introduces KIPP or HCZ and attracts cS highly-dedicated teachers, pulled equally from schools 2 and 3. The boost to school 1 is equal to cS, consistent with the published evaluation results. School 2 loses cS/2, as does school 3, resulting in a total loss of achievement at schools 2 and 3 equal to $cS/2 + cS/2 = cS$. The aggregate gain in achievement across the three schools is therefore zero.

Implications

The results of the formal analysis suggest that it is worth considering the full implications of high teacher attrition. While the KIPP co-founders concede that it may be difficult to scale up and replicate the KIPP model on a national scale, and other researchers have suggested the same conclusion, no previous analysis has suggested that it would be impossible to scale up the KIPP/HCZ models, given their core assumption and evidence of high teacher attrition.

To explain why the putative impacts of KIPP and HCZ necessarily fall to zero when the programs are implemented nationally, under Canada's assumption that finding great teachers is the "secret sauce" of great charter schools, it is useful to consider an analogy. If KIPP and HCZ rejected 49 percent of the least-motivated students every year and only accepted the most-motivated students, the measured impact of the KIPP and HCZ approaches would be artificially boosted, simply because of this creaming effect. A similar effect would occur if KIPP and HCZ screen out the least-motivated 49 percent of teacher recruits by setting grueling work hours and conditions. This creaming effect would leave only the most-motivated, dedicated teachers, artificially raising the measured performance of KIPP and HCZ in any impact evaluation. Unfortunately, it is impossible to detect this bias in any published evaluation of KIPP or HCZ because those evaluations simply state the gains achieved by the corps of teachers who remain with KIPP or HCZ after substantial creaming has already occurred; the results do not tell us what would happen if the KIPP/HCZ approaches were to be used with an average group of teachers. However, it is clear that the impact results would be biased in any case where the least-motivated teachers were regularly screened out.

The resulting bias in existing KIPP/HCZ impact evaluation results has been hidden because KIPP and HCZ schools can pull highly-motivated teachers from non-KIPP/HCZ schools (or from the supply of those teachers to non-KIPP/HCZ schools) in order to fill slots that open up due to attrition. However, the bias would become clear if every school across the nation adopted KIPP or HCZ. Under that condition, KIPP/HCZ schools could only pull teachers from other KIPP/HCZ schools, or from the supply of new college graduates who would ordinarily go to those other schools. The effectiveness of every KIPP/HCZ school would be reduced as every school lost dedicated teachers recruited by other KIPP/HCZ schools. Without the ability to stock their teaching staffs with highly-dedicated teachers, each KIPP/HCZ school would be forced to recruit from the less-dedicated corps of teachers that was rejected when KIPP and HCZ were only implemented in a few schools across the nation. The performance of each KIPP and HCZ school would necessarily decline to a normal level—the level that prevails when highly-dedicated teachers are evenly distributed across all schools instead of being concentrated in a few KIPP/HCZ schools. Evaluations of the KIPP and HCZ schools would reflect this lower level of performance. The magnitude of the reduction in KIPP and HCZ school performance would provide information about the magnitude of bias in current estimates of KIPP/HCZ impact. The results of the formal analysis imply that the reduction would equal the difference in performance between current KIPP/HCZ schools and non-KIPP/HCZ schools, under Canada's assumption that the main reason for the outstanding performance of current KIPP/HCZ schools is their recruitment of a staff of unusually dedicated teachers. In sum, the formal analysis implies that if the unusually dedicated staff goes away, the unusually high performance goes away.[47]

A question that arises is whether the national implementation of the HCZ and KIPP charter schools might attract a much larger pool of highly-dedicated individuals to the teaching profession, thereby filling the empty teaching slots that are created by high teacher attrition. In fact, a central claim of market-oriented reformers is that the use of merit pay and the practice of retaining teachers on the basis of student learning gains might attract a more talented pool of individuals to the teaching profession. Hanushek, for example, makes this argument.[48] However, he concedes that "we do not know how teacher quality responds to different levels of salaries."[49] In other words, his claim that merit pay might elicit a stronger pool of teachers is not based on empirical data. Similarly, it is possible that KIPP and HCZ may, in the long run, induce a larger flow of highly-dedicated individuals into the teaching profession—individuals who are attracted by the KIPP/HCZ emphasis on good teaching. In addition, KIPP and HCZ might inspire and transform teachers who are currently not highly dedicated, such that they become teachers who are highly dedicated. Either effect could serve to address the problem of high teacher attrition. At present, however, there is no evidence that either of these effects is significant. In the absence of evidence, it would not be appropriate to assume that KIPP or HCZ would elicit a substantially greater flow of talented individuals into the teaching profession. On the contrary, reports from principals of "no-excuses" charter schools such as KIPP and HCZ that they have to "scour the country" for suitable teachers suggests that, to date, neither KIPP nor HCZ has inspired an adequate flow of such individuals to take up the profession of teaching.[50] This is consistent with the view of the KIPP co-founders that it is impossible to scale up and replicate the KIPP model on a national scale, given the current pool of teachers.[51] Whatever KIPP and HCZ have accomplished, they have not inspired a vast pool of talented individuals to switch to the teaching profession. Given that the principals of the KIPP and HCZ schools have to scour the country for suitable teachers, the only way that the schools can acquire the necessary teachers is by hoarding.

The validity of the current analysis is independent of explanations about why some teachers fail to thrive in KIPP/HCZ schools. It does not matter whether failure is a consequence of their lack of dedication or commitment or their ineffectiveness relative to their peers who remain in the KIPP/HCZ schools. For the purpose of the current analysis, it is sufficient to demonstrate three conditions: 1.) the KIPP/HCZ models are based on the assumption that high performance is a consequence of highly-dedicated teachers, 2.) there is high attrition among KIPP/HCZ teachers, and 3.) there is a shortage of highly-dedicated teachers who can replace teachers who leave. All of the conclusions of the analysis follow from these three conditions. No other assumption is necessary. It is conceivable, for example, that teachers who leave are just as talented as teachers who stay, but simply lack the necessary endurance and fortitude. The end result is that there is a shortage of the type of teacher that is required by the KIPP/HCZ models. If there is a shortage, then—by definition—the models cannot be scaled up nationally. They can only succeed in scattered examples here and there because they can pull a disproportionate share of dedicated teachers. If an attempt is made to scale up the models nationally and the option of pulling teachers from other schools is eliminated, it becomes logically impossible to maintain the gains that were achieved in the small-scale research studies.

It is not helpful to compare the current gain scores of the students taught by teachers who stay, versus teachers who leave. Lower gain scores for teachers who leave would support the interpretation presented throughout this chapter—that creaming of highly-effective teachers occurs because less-dedicated, lower-performing teachers quit. However, equal gain scores would not be inconsistent with the thesis that creaming occurs. Highly-effective teachers may be attracted to the KIPP and HCZ schools in disproportionate numbers (rather than being culled from a larger set of teachers, some of whom are less-dedicated and less-effective). When the highly-effective teachers quit, they scatter across the rest of the nation's schools. Researchers who compare the performance of KIPP/HCZ schools to non-KIPP/HCZ schools would continue to find that the KIPP/HCZ schools outperform the non-KIPP/HCZ schools—until these models are scaled up nationwide.

Finally, the validity of the analysis presented here is independent of evidence that there are plenty of entrants to the teaching profession, but poor organizational conditions, including high pressure from accountability systems, cause many teachers to leave the profession.[52] Whatever the reasons that teachers leave the profession, the existence of KIPP and HCZ schools has clearly not reversed this trend. If the organizational changes implemented by KIPP and HCZ were sufficient to reverse this trend, then we should observe long waiting lists of teachers seeking to transfer to KIPP and HCZ schools. This is clearly not the situation if charter schools such as KIPP and HCZ report that they have to "scour the country" for suitable teachers.[53] Thus, there is no reason to think that scaling up the KIPP and HCZ models would reverse the enormous teacher attrition that occurs across the nation.

The basic problem with the KIPP and HCZ models is that if they are implemented in every school, there are no longer any non-KIPP/HCZ schools where teachers may be pulled. Recruiting schools can only pull teachers from the limited national supply of highly-dedicated teachers that is available to all KIPP/HCZ schools. If one school garners a disproportionate number of those teachers, other KIPP/HCZ schools must lose those teachers. Recruitment becomes a zero-sum game. Since it would no longer be possible for all KIPP/HCZ schools to maintain a teaching force of the most highly-dedicated individuals from across the nation, performance would inevitably decline. KIPP/HCZ schools would lose the essential character that made them successful when implemented in a few scattered schools across the nation, simply because there is no waiting, unemployed army of individuals seeking to become teachers whose level of performance exceeds the 49th percentile. This is true whether teachers are pulled directly from other KIPP/HCZ schools, or indirectly, out of the supply of brand new teacher candidates.

In essence, existing evaluations of the KIPP/HCZ approach lack external validity. To address the issue of external validity, it would be necessary to implement KIPP/HCZ in every school within a defined geographical area and prohibit the schools from recruiting teachers from outside of that area. To the extent that the schools recruit nationally, they would be drawing down the small corps of highly-dedicated teachers that would otherwise be available to schools outside of the area.

The KIPP/HCZ approach can only be scaled up if working conditions are changed so the life of a KIPP/HCZ teacher is less exhausting and more attractive. However, relaxing the bar would permit less-dedicated teachers to remain. The bar would have to be relaxed substantially if the KIPP/HCZ approach were to be extended to all schools. It is likely that the bar would have to be lowered to the current level for non-KIPP/HCZ schools. If every school is a KIPP or HCZ school and teachers have no option but to remain at a KIPP or HCZ school or leave the profession, a bar that is any higher would drive teachers out of the profession and exacerbate the current shortage of teachers, increasing class sizes or leaving large numbers of students without teachers.

The policy implication that has been drawn from published impact evaluations of KIPP and HCZ is that they represent promising approaches for raising student achievement, if only the issue of teacher burnout can be addressed.[54] However, this issue cannot be treated as a minor issue. The high burnout rate is not only a fundamental flaw in the approach, but suggests that positive interpretations of KIPP's and HCZ's effects on student achievement are based on flawed analyses. Reported gains in student achievement are most likely artifacts due to the attrition of up to 49 percent of the teaching force every year, leaving the most-dedicated teachers and reflecting brief spurts of 65-hour work weeks that cannot be sustained over time.

The KIPP and HCZ approaches work by recruiting highly-dedicated teachers, creating a work environment that only the top half of all teachers can survive, and constantly recruiting additional teachers across the nation in a wide-ranging attempt to fill the empty teaching slots. The consequence, however, is that this process inevitably pulls the best teachers from across the country, leaving an insufficient number of those teachers to implement the same approach in every school nationwide.

The implication is that the KIPP/HCZ model can only be scaled up if the number of teacher applicants is twice the size of the current teaching force, permitting KIPP and HCZ schools to reject the bottom-half of those applicants. However, there is currently a teacher shortage, and the shortage is projected to become increasingly severe over the foreseeable future.[55] The vast army of unemployed, highly-qualified, and highly-dedicated teachers that is required to implement KIPP and HCZ on a nationwide basis simply does not exist.

This issue has largely been overlooked by previous researchers. At first glance, the promising impact findings held out hope that the KIPP/HCZ recipe had solved the problem of low student achievement. All that was required was to scale up and implement this model nationwide. Researchers claimed that this approach would close the achievement gap between poor minority students and their more advantaged peers.[56] However, the evaluations of KIPP and HCZ overlooked the need to pull highly-dedicated teachers from non-KIPP and non-HCZ schools in order to fill the large number of empty teaching slots created each year when teachers are unable to endure the exhausting KIPP and HCZ workdays. Thus, the results of the research studies lacked external validity—the type of validity that is required if KIPP and HCZ are to be successfully scaled up and implemented nationwide. The problem is not a research issue that can be addressed in future studies—it is inherent in the KIPP/HCZ approach. The only way the issue can be addressed is by making KIPP/HCZ less grueling for teachers and relaxing the bar for hiring and retaining teachers. That bar must be relaxed to the point where teachers do not leave the KIPP/HCZ schools in disproportionate numbers. However, if the bar is relaxed, those schools cease to maintain the essential character of KIPP and HCZ—the character established by recruiting the most-dedicated teachers in the nation.

Policymakers may wonder how it is possible that KIPP and HCZ can draw the most-dedicated teachers, produce impressive results as measured by carefully controlled research studies, but not be scalable. To draw an analogy, an automobile that is not working can be pushed by a team of very strong, highly-dedicated athletes. In a few cases, these athletes may even sustain an impressive speed for a short period of time. However, this type of athlete is rare, and it is impractical to recruit a sufficient number of these athletes to push all of the stalled cars nationwide.

The task of raising student achievement is not simple. However, it may be useful to diagnose the reasons for low student achievement, much as it is useful to diagnose the reasons when a car is stalled, before attempting a solution. A proper diagnosis might suggest that attention should be focused in a different direction, and might be more efficient, just as it would be efficient to fix a faulty electrical system in order to start a stalled car, instead of recruiting a team of super athletes to push the car. The evidence offered in this book suggests the nature of the direction that may be fruitful.

CHAPTER 9

Better Teachers

Value-added modeling (VAM) may be defined as the use of statistical methods for the purpose of isolating the "value-added" contribution of individual teachers to student achievement.[1] Numerous researchers advocate the use of value-added performance information to make decisions about hiring, firing, rewarding, or promoting teachers.[2] School districts across the nation are adopting VAM and many districts are using VAM for high-stakes decisions.[3] Tennessee uses VAM in teacher evaluations, including decisions regarding teacher tenure and dismissal.[4] The District of Columbia public schools use VAM in teacher evaluations, including decisions regarding teacher dismissal,[5] while 29 District of Columbia public charter schools incorporate VAM into teacher evaluations.[6] Florida and Ohio public schools use VAM in teacher evaluations, including decisions regarding dismissal.[7] Colorado public schools use VAM in teacher evaluations, including decisions regarding tenure.[8] The Pittsburgh public schools use VAM in teacher evaluations.[9] The states of New York and Louisiana are incorporating VAM into teacher evaluations.[10] North Carolina and the Los Angeles school district are incorporating VAM into teacher evaluations.[11] Federal policy endorses this approach by directing federal funds to states that adopt the approach.[12]

Given the rapid adoption of VAM in districts across the nation and its apparent endorsement by researchers and policymakers, there is a need to evaluate the effectiveness of policies based on VAM. One study in particular has received a great amount of attention.[13] Raj Chetty, John Friedman, and Jonah Rockoff suggest that the use of VAM to identify the lowest-performing five percent of all teachers and replace them with average teachers would increase student achievement and would translate into sizable gains in the lifetime earnings of their students: "The estimated total undiscounted earnings gains from this policy are approximately $50,000 per child and $1.4 million for the average classroom."[14] These startling figures have been cited to justify the use of VAM and appear likely to accelerate the adoption of VAM by school districts nationwide.[15]

S.S. Yeh, *Solving the Achievement Gap*,
DOI 10.1057/978-1-137-58767-1_9

If policies based on VAM are indeed as effective as indicated by the Chetty et al. study, then national implementation would appear to be warranted. However, Jesse Rothstein examined the study's assumptions and concluded that VAM-based estimates of teacher quality are biased, unreliable, and invalid:

> My results are sufficient to re-open the question of whether high-value added elementary teachers have substantial causal effects on their students' long-run outcomes…[there is] no strong basis for conclusions about the long-run effects of high- vs. low-value added teachers, which in the most credible estimates are not distinguishable from zero.[16]

Rothstein found that teachers with high value-added scores may have been systematically assigned students whose increased earnings are attributable to factors other than differences in teacher value-added.[17] Rothstein concluded that Chetty et al.'s tests to detect bias in their estimates were inadequate. It is not appropriate to attribute increased earnings to measured differences in teacher value-added. This chapter explores the issue of bias as well as numerous problematic assumptions that are implicit when VAM is used to identify and replace low-performing teachers.

An important distinction is whether VAM is only used to replace existing indicators of teacher quality such as principal judgments, for existing decisions (regarding merit pay, for example) that are already being made, or whether it is used to justify a large expansion of teacher termination and replacement, as in the case of the Chetty et al. proposal. It may be the case that VAM is a better predictor than other indicators that are currently used to make decisions about pay, promotion, or hiring and, therefore, can be justified as a better substitute for those other indicators. However, the controversial aspect of VAM is its expanded use as an explicit strategy for terminating teachers who would not otherwise be terminated, in an effort to improve student achievement. What is missing from this discussion are analyses to determine whether this strategy is a cost-effective use of society's scarce resources, compared to alternative strategies for raising student achievement. VAM might be justified as the best predictor of teacher quality for decisions that are already being made using less reliable indicators, but may not be justifiable for policies that involve vast expansion of teacher termination and replacement, unless this strategy has been shown to be the most cost-effective approach for raising student achievement.

The first section of this chapter reviews literature regarding the reliability and validity of VAM. The second section of this chapter analyzes several key assumptions underlying the Chetty et al. study and suggests that these assumptions are common to studies that evaluate the effectiveness of policies based on VAM. The second section includes cost-effectiveness and benefit-cost analyses of the Chetty et al. proposal. The third section of this chapter concludes that VAM is neither reliable nor valid for the purpose of high-stakes decisions regarding teacher hiring and firing, and VAM-based policies are not cost-effective strategies for raising student achievement. In view of the need to consider alternatives, the third section compares VAM-based policies to rapid performance feedback (RPF), which appears to be far more cost-effective and suggests an alternative way of thinking about strategies for improving student achievement.

Reliability and Validity Issues

Interest in VAM was stimulated by Sanders and Rivers, who used statistical methods to isolate the contribution of individual teachers to student achievement two years into the future.[18] The results suggested that teachers have persistent effects on their students' achievement and the accumulation of these effects can be substantial. The following year, Sanders and his colleagues published an article asserting that teachers are the most important factor influencing student achievement.[19] Interest in VAM grew as subsequent studies indicated that the contribution of teachers to student achievement is large and value-added estimates of teacher contributions predict their students' measured achievement.[20] Teacher ratings based on VAM are moderately correlated with ratings obtained from portfolio evidence and classroom observations conducted by trained evaluators.[21] The evidence that teachers have significant effects on student achievement led many researchers to advocate the use of VAM to identify and replace low-performing teachers.[22] Using Monte Carlo simulations, Staiger and Rockoff asserted that "80 percent of teachers should be dismissed after their first year" based on VAM estimates of their effectiveness.[23] Advocates of using VAM for high-stakes decisions regarding teacher hiring and firing argue that concerns about falsely identifying low-performing teachers can be addressed by using multiple years of data to estimate each teacher's ranking, and an excessive concern with false identifications serves the interests of teachers, rather than their students.[24]

Any VAM-based policy to identify and replace low-performing teachers, however, requires the strong assumption that specific teachers *cause* the observed gains or losses in their students' achievement.[25] The critical assumption is that any differences among classes, schools, or programs that are not captured by the predictor variables used in the VAM model are captured by the student-fixed-effect components.[26] However, using data from North Carolina, Rothstein found that the value-added scores for fifth-grade teachers predicted their students' prior performances.[27] For the purpose of predicting their students' fourth-grade scores, the value-added scores for the fifth-grade teachers were nearly as strong a predictor as the value-added scores for their students' fourth-grade teachers.[28] Since it is impossible for fifth-grade teachers to cause performance that occurred prior to the fifth grade, this result implies there is nonrandom selection of students into teacher classrooms that is not controlled through the inclusion of time-invariant student characteristics. Therefore, the central assumption underlying VAM appears to be invalid.[29] Rothstein concluded: "Results indicate that even the best feasible value-added models may be substantially biased."[30] This surprising result suggests that the use of VAM to identify and terminate low-performing teachers is not warranted. When teachers are assigned students who achieved high gains in performance the previous year, existing VAM models erroneously subtract a portion of the gain that is properly attributed to these teachers, making them look like bad teachers.[31] This problem may be exacerbated if VAM is used to identify and terminate teachers because the high stakes may cause teachers to lobby principals for students who are predicted to post large gains in the coming year, and principals may be tempted to use their control over classroom assignments to reward favored teachers.[32]

Using data from San Diego, Koedel and Betts corroborated Rothstein's primary finding, demonstrating that the effect is not unique to North Carolina.[33] However, Koedel and Betts also found that sorting bias can be almost completely mitigated when a complex value-added model is used that restricts the analysis to teachers for whom at least three contiguous cohorts of student test scores are available.[34] A major difficulty, however, is that it would not be uncommon for data to be missing in a way that would prevent the use of this technique with large numbers of teachers.[35] Not only would it be necessary for teachers to have three contiguous cohorts of student test scores, but most VAM models are restricted to complete cases of data, which is only appropriate if the missing data are missing completely at random.[36] This assumption is inappropriate because systematic factors influence missing school data. For example, students who move may be more likely to be students who perform at lower levels.

Ishii and Rivkin identified specific parent and school influences on student assignment to classrooms that may systematically bias VAM estimates even when the models incorporate student-fixed effects.[37] Highly-educated parents are more likely to request that their children be assigned to particular teachers. Highly-educated parents may also hire tutors during time periods when they perceive that their children's teachers are inadequate. Also, teachers tend to prefer classrooms with higher-achieving students, and principals might assign high-performing teachers to classrooms with high-achieving students as an incentive for the high-performing teachers to remain at a school. Not all of these influences could be controlled using student-fixed effects because the purposeful nature of these choices almost certainly introduces correlations among teacher quality and family/student characteristics.[38]

Employing the same tests used by Rothstein, Briggs and Domingue analyzed the VAM model developed by the RAND Corporation and used by the Los Angeles Unified School District to rank teachers.[39] Briggs and Domingue found that estimates produced by the model are significantly biased and teacher rankings are highly dependent on the specification of the model.[40] An alternative specification controlling for a longer history of each student's performance, peer influences, and school-level factors produced different teacher ratings: in reading, 53.6 percent of teachers did not retain the same effectiveness rating under both specifications; in math, 39.2 percent of teachers did not retain the same effectiveness rating.[41] This suggests that teacher ratings using VAM are highly sensitive to details regarding the model's implementation.

Ballou, Sanders, and Wright point out that the inclusion of socioeconomic status (SES) in an effort to control for differences in family/student characteristics would bias any estimate of teacher effectiveness toward zero because of the likely correlation between SES and teacher quality.[42] For this reason, the Education Value-Added Assessment System (EVAAS), a popular variant of VAM, omits student covariates including SES. However, McCaffrey et al. found that this would likely confound estimated teacher effects, and teacher rankings based on these effects, when different schools serve distinctly different student populations.[43] Ballou et al. point out that EVAAS, which uses each student's test score history to substitute for SES and demographic variables, is not vulnerable to missing SES and demographic data and, in Tennessee, produced teacher rankings that were comparable to rankings when SES and demographic variables were included. However, no systematic study has examined EVAAS rates of false positive and false negative teacher classifications.[44]

Another problem with VAM is that it does not appear possible to separate teacher and school effects using currently available accountability data.[45] Separating these effects would only be possible if each teacher regularly taught at multiple schools where the accountability systems were consistent and the data were available across schools. Currently, however, when VAM is used to estimate individual teacher effects and to rank teachers, these estimates are contaminated by effects that are properly attributed to schools, not teachers. Furthermore, there is no obvious solution to this problem.

A largely-ignored problem is that *true* teacher performance, contrary to the main assumption underlying current VAM models, varies over time.[46] These models assume that each teacher exhibits an underlying trend in performance that can be detected given a sufficient amount of data. The question of stability is not a question about whether *average* teacher performance rises, declines, or remains flat over time. The issue that concerns critics of VAM is whether *individual* teacher performance fluctuates over time in a way that invalidates inferences that an individual teacher is "low" or "high" performing. This distinction is crucial because VAM is increasingly being applied such that individual teachers who are identified as low performing are to be terminated. From the perspective of individual teachers, it is inappropriate and invalid to fire a teacher whose performance is low this year, but high the next year, and it is inappropriate to retain a teacher whose performance is high this year but low next year. Even if average teacher performance remains stable over time, individual teacher performance may fluctuate wildly from year to year.

Goldhaber and Hansen investigated the stability of teacher performance in North Carolina using data spanning ten years and found that 70 percent of a teacher's estimated performance varies over time:

> We estimate that much of a teacher's estimated performance in a single year (classroom) is variable over time—21 percent due to measurement error, 15 percent due to random non-persistent performance fluctuations over time, and 34 percent due to dynamic changes in performance that evolve slowly over time.[47]

Goldhaber and Hansen found that 95 percent of the variation in a teacher's performance is due to unobservable factors such as effort, motivation, and class chemistry that are not easily captured through VAM.[48]

Goldhaber and Hansen's results invalidate the assumption of stable teacher performance that is embedded in Hanushek's[49] and Gordon, Kane, and Staiger's[50] VAM-based policy proposals, as well as VAM models specified by McCaffrey, Sass, Lockwood, and Mihaly[51] and Staiger and Rockoff.[52] [53] The implication is that standard estimates of impact when using VAM to identify and replace low-performing teachers are significantly inflated.[54] Furthermore, Goldhaber and Hansen's results indicate that the variation in performance attributable to unobservable factors such as effort, motivation, and class chemistry is large. The variation due to these factors is not measured and controlled in existing value-added models. This suggests that the models are incorrectly specified and specification bias is a problem. Since there is no simple way to measure effort, motivation, and class chemistry, there is no easy solution to the problem.

Another problem arises when, for example, a pretest score measures pre-algebra but the posttest score measures geometry skills, or when a teacher emphasizes pre-algebra but not geometry. Improvements in learning may not be captured by the assessment. A mismatch between instruction and assessment would tend to invalidate VAM-based teacher rankings.[55] VAM requires the use of vertically-scaled achievement data that span wide grade, developmental, and content ranges; however, the shift in constructs that are measured from grade to grade introduces remarkable distortions: effective teachers may be identified as ineffective and vice-versa, and effects contributed by prior teachers may be erroneously attributed to later teachers.[56] Martineau concluded: "With current technology, there are no vertical score scales that can be validly used in high-stakes analyses for estimating value added to student growth in either grade-specific or student-tailored construct mixes…A serious (but reasonable) implication of this study is to all but eliminate the high-stakes use of value-added accountability systems based on vertically scaled student achievement data."[57] Even when instruction and assessment are matched, differences in the particular achievement tests that are used produce substantially different answers about individual teacher performance and do not rank teachers consistently.[58]

Assumptions

The preceding review of literature suggests numerous reasons for caution in using the results of any VAM model to identify and replace low-performing teachers. These concerns are magnified when VAM is used, as it is used in the Chetty et al. study, to make assertions about the long-term economic benefits to students who are taught by teachers identified as "high-performing" teachers according to the VAM analysis. The analysis presented in this chapter suggests that the findings of the Chetty et al. study depend on numerous assumptions that may be questioned. Significantly, these assumptions are common to studies that predict positive benefits of policies based on VAM. Therefore, the analysis presented here has implications for VAM-based policies in general, whenever they are used to make predictions about the long-term benefits of identifying and replacing low-performing teachers.

Fixed Teacher Quality?

A key assumption of the Chetty et al. analysis is that true teacher quality is fixed over time. Consider their statistical model,[59] where Y_{it} denotes the earnings of student "i" at time "t" (during adulthood), and m_{jt} denotes the value-added score of the student's teacher (during the period when student "i" is a student):

$$Y_{it} = a + k_g m_{jt} + \eta_{it}$$

Chetty et al. use this model to estimate the impact of having a high value-added teacher in grade "g" while "holding fixed future teacher quality."[60] The coefficient of teacher quality is denoted by the constant k_g, and Chetty et al. state that "the ultimate earnings impact of retaining teachers on the basis of their value-added depends on k_g."[61]

The fact that the coefficient k_g is fixed over time indicates that Chetty et al. are assuming that true teacher quality remains fixed over time—in other words, Chetty et al. are assuming that a teacher who appears to boost student achievement during the one-, two-, or three-year base period when data are collected about the performance of the teacher will also provide the same boost to student achievement in the future, after a decision is made by the teacher's school principal to retain and continue to employ the teacher.

The Chetty et al. analysis assumes that a high-quality teacher this year will remain a high-quality teacher next year, and a low-quality teacher this year will remain a low-quality teacher next year. In a companion article, however, Chetty et al. conclude that, in fact, "teacher quality fluctuates over time."[62] Estimates of teacher quality obtained during a base period are not reliable predictors of teacher quality in future years: "Because teacher quality drifts over time, the predicted effect differs from past performance."[63] As a consequence, the authors conclude that the actual gains that might be expected if a low-quality teacher is replaced with a high-quality teacher are significantly smaller after accounting for "drift in teacher quality": "The gains from deselecting teachers based on estimated value-added are significantly smaller because of noise in value-added estimates and drift in teacher quality."[64] When using one year of performance data, the gains are nearly halved.[65]

The authors' primary analysis assumed that teacher quality is fixed over time but their own data suggest that teacher quality, as measured by teacher value-added, is not in fact time-invariant, consistent with the results reported by Goldhaber and Hansen.[66] While Chetty et al. minimize the significance of "drift" in teacher quality, the available evidence indicates that teacher quality varies significantly over time, invalidating the assumption that a high-quality teacher this year will remain a high-quality teacher next year, and a low-quality teacher this year will remain a low-quality teacher next year.

The intertemporal reliability of value-added teacher rankings was investigated by Aaronson et al.,[67] Ballou,[68] Koedel and Betts,[69] and McCaffrey et al.[70] In each study, VAM was used to rank teacher performance from high to low. In each study, a majority of teachers who ranked in the lowest quartile or lowest quintile shifted out of that quartile (or quintile) the following year (see Tables 9.1 and 9.2). Furthermore, a majority of teachers who ranked in the highest quartile or quintile shifted out of that quartile (or quintile) the following year (see Tables 9.1 and 9.2).

What this means is that value-added teacher rankings are insufficiently reliable for the purpose of high-stakes decisions regarding hiring and firing. High-stakes decisions are clearly unwarranted if this volatility in the rankings is due to unmeasured variables or random measurement error. However, even in the unlikely event that there are no unmeasured variables and measurement error is zero, implying that all volatility is due to true variation in teacher

Table 9.1 Instability of Value-Added Teacher Rankings in Chicago and Tennessee

	Teacher Rankings	
Locale	*Bottom 25% in Year t;* *Top 75% in Year t+1*	*Top 25% in Year t;* *Bottom 75% in Year t+1*
Chicago, IL	67%	59%
Tennessee	60%	52%

Note: Chicago data are from Aaronson, et al. (2007), Table 7, for high school math teachers, with controls for student, peer, and neighborhood covariates. Tennessee data are from Ballou (2005), Figure 5b, for math teachers in grades 3 through 8 in a single large district.

Table 9.2 Instability of Value-Added Teacher Rankings in San Diego and 5 Florida Counties

	Teacher Rankings	
Locale	*Bottom 20% in Year t;* *Top 80% in Year t+1*	*Top 20% in Year t;* *Bottom 80% in Year t+1*
San Diego, CA	65%	71%
Dade County, FL	70%	67%
Duval County, FL	67%	61%
Hillsborough County, FL	67%	67%
Orange County, FL	59%	65%
Palm Beach County, FL	69%	68%

Note: San Diego data are from Koedel and Betts (2007), Table 9, based on elementary school math teachers, with controls for student- and school-fixed effects. Data for Florida counties are from McCaffrey et al. (2009), Table 4, based on elementary school math teachers with 15 or more students per year, with controls for student-fixed effects.

performance, it would not be appropriate to hire or fire based on the ranking in a given year (designated "year t"). In over half of all instances, performance would have either improved or declined the following year (designated "year t+1") by such an extent as to invalidate the year t ranking. If VAM is used to identify and fire the bottom quartile (or quintile) of teachers, the results in Tables 9.1 and 9.2 indicate that this decision is incorrect, according to the year t + 1 teacher rankings, between 59 and 70 percent of the time. These results suggest that productive teachers would be culled more frequently than unproductive bottom-quartile (or bottom-quintile) teachers.[71]

In the case of value-added rankings, *it is inappropriate* to infer that a teacher should be hired or fired based on the rankings from any given year. Since this inference would be inappropriate, the results of value-added teacher rankings are not valid for the purpose of high-stakes decisions regarding hiring and firing.[72] In short, VAM lacks validity for the purpose of high-stakes decisions regarding individual teachers.

While some researchers suggest averaging two or more years of rankings to improve reliability, averaging may introduce significant bias—raising the issue of validity once again.[73] Furthermore, it would not be uncommon for data to be missing in a way that would prevent averaging. For large numbers of teachers, it would be impractical to average their rankings across two or more years.[74] Regardless, when two years of rankings are used for tenure decisions, intertemporal reliability remains low: in reading, data from North Carolina indicate that 68 percent of teachers ranked in the bottom quintile shift out of that quintile after tenure (indicated by a weighted average of all post-tenure observations), and 54 percent of teachers ranked in the top quintile shift out of that quintile post-tenure.[75] When three years of rankings are used, reliability is even worse: 74 percent of teachers ranked in the bottom quintile shift out of that quintile post-tenure, and 56 percent of teachers ranked in the top quintile shift out of that quintile post-tenure.[76] In math, reliability is somewhat better, but over half of all teachers in the bottom and top quintiles shift out of those quintiles post-tenure.[77]

These results were confirmed and extended by an analysis, also using data from North Carolina, which found that more than half of all teachers who ranked in the bottom quintile shifted out of that quintile the following year, regardless of whether one, two, three, four, or five years of data were used to predict future performance, regardless of the subject area (math or reading), and regardless of whether a simple or complex Bayes estimator was used to improve predictive accuracy.[78]

Unbiased?

Chetty et al. interpret their results as if their coefficients are unbiased and the impact of an individual teacher can be isolated: "k represents the mean impact of having a higher value-added teacher for a single grade between grades 4–8."[79] Chetty et al. interpret the coefficient as follows: "A one standard deviation increase in teacher value-added in a single grade increases earnings at age 28 by $350, 1.65 percent of mean earnings in the regression sample."[80] After incorporating additional variables into their model, Chetty et al. adjusted this figure downward and concluded that "a one standard deviation increase in teacher value-added raises earnings by 1.34 percent."[81] To calculate the increase in the earnings of students taught by the teacher, Chetty et al. "assume that the percentage impact of a one standard deviation improvement in teacher value-added on earnings observed at age 28 is constant at b = 1.34 percent (Table 3, column 2) over the life cycle" of each student taught by the teacher.[82] This assumption is the foundation for their statement that replacing a low-quality teacher with a high-quality teacher would result in a large lifetime gain in income for each student taught by this teacher: "Under these assumptions [and assuming average present value lifetime earnings equal to $522,000 for each student taught by this teacher], the financial value of having a one standard deviation higher value-added teacher (i.e., a teacher at the eighty-fourth percentile instead of the median) is 1.34 percent × $522,000 ≃ $7,000 per grade. The undiscounted lifetime earnings gain (assuming a 2 percent growth rate but 0 percent discount rate) is approximately $39,000 per student."[83] This equals $1,099,800 for a classroom of 28.2 students.[84]

However, Chetty et al. acknowledge, due to limitations in their analytical method, that it is not valid to interpret the coefficient for teacher quality as if the impact of teacher quality has been isolated from the influence of all other inputs: "k_g cannot be interpreted as the *structural* impact of teacher quality holding fixed all other inputs."[85] As a consequence, factors other than teacher quality may explain the 1.34 percent gain in earnings observed at age 28. Some of the impact may be due, for example, to the influence of parental social connections that permit children from wealthier families to obtain higher-paying jobs. This influence was not controlled in the Chetty et al. analysis, nor is there an obvious methodological remedy that could be applied by other researchers—suggesting that the problem is not easily corrected. The need to control for social connections is especially important because even a weak influence from connections might explain a small 1.34 percent difference in annual earnings.

Chetty et al. devoted a companion article to the topic of bias but concluded, after various tests, that "our quasi-experimental estimates show that the degree of bias due to selection on unobservables turns out to be negligible."[86] In October 2015, however, Jesse Rothstein posted results from analyses indicating that Chetty et al.'s tests for bias are inadequate:

> I find that teacher switching does not create a valid quasi-experiment. The treatment—the change in the average value-added of the teaching staff in a school-grade cell from one year to the next—is not as good as randomly assigned but rather is correlated with pre-determined student characteristics that are predictive of outcomes.[87]

As a consequence, a teacher's value-added score may be overestimated for teachers whose students are systematically stronger than expected, given their observed characteristics, and a teacher's value-added score may be underestimated for teachers whose students are systematically weaker than expected, given their observed characteristics.[88] Rothstein concluded that none of the tests for bias conducted by Chetty et al. alter the conclusion that their VAM model is a biased, unreliable predictor of long-run teacher effects:

> Using Chetty et al.'s methods and drawing on their [statistical] programs (CFR 2014f), I successfully reproduce all of the key results of each paper. Further investigation, however, indicates that neither North Carolina nor New York data support Chetty et al.'s substantive conclusions regarding value-added bias or teachers' long-run effects.[89]

Rothstein's results suggest that bias may be inherent in every VAM model, suggesting the possibility that previously-published VAM-based estimates of the contribution of individual teachers to student achievement may be biased, over-optimistic, and invalid. In Rothstein's view, the most credible estimates of the true magnitude of teacher contributions "are not distinguishable from zero":

> My results are sufficient to re-open the question of whether high-value added elementary teachers have substantial causal effects on their students' long-run outcomes…[there is] no strong basis for conclusions about the long-run effects of high- vs. low-value added teachers, which in the most credible estimates are not distinguishable from zero.[90]

If no existing VAM-based estimate of teacher quality is demonstrably free from the type of bias that Rothstein has discovered, the basis for conclusions about the effects of high- versus low-value-added teachers is weak. Any statement, based on VAM, that teacher contributions to student achievement are significant is open to question. Rothstein's analysis challenges not only the Chetty et al. analysis, but also the previous body of studies, based on VAM, suggesting that teachers make significant contributions to student achievement.[91] The conclusion that VAM is flawed raises serious questions about the broad assertion that teachers make significant contributions to student achievement. The main foundation for this assertion rests on studies, based on VAM, suggesting that the contribution of teachers to student achievement is large. The identification of a serious flaw in VAM suggests a flaw in the foundation of a core belief that is driving national as well as state education policy.

Persistent Effects?

Chetty et al. assumed, arbitrarily, that the 1.34 percent increase in earnings observed at age 28 would persist at every age throughout an adult's life, resulting in large cumulative lifetime gains in earnings: "All values in these figures are based on our estimate that a one standard deviation increase in true teacher value-added increases earnings by 1.34 percent."[92] This assumption is not consistent with evidence that a teacher's impact quickly fades. Other researchers employed stronger analytical methods and found that the fade-out is large, quick, and any persistent effect is small. For example, Kane and Staiger[93] employed random-assignment of teachers to students and found that half of a teacher's impact fades after one year and an additional 50 percent fades after the second year, implying that no more than 25 percent of a teacher's impact persists after two years.[94] Therefore, Chetty et al.'s ad hoc assumption may be questioned. Perhaps a more reasonable assumption, one that is more consistent with the evidence regarding fade-out, is that the 1.34 percent increase in income observed at age 28 fades by 50 percent in each subsequent year.[95]

Consequential?

Given Rothstein's results and the absence of adequate controls for factors that contribute to bias in Chetty et al.'s estimates, plus the evidence that VAM rankings lack adequate stability for operational decisions, Chetty et al.'s estimate of the impact of raising teacher quality by one standard deviation (SD) may be questioned. In any case, the estimated impact is quite small. With regard to student achievement, a one unit increase in teacher quality is associated with a 0.993 SD increase in student test scores that declines to 0.221 SD four years after students receive instruction from their teachers.[96] Chetty et al. reported that the standard deviation of teacher quality ranges from 0.042 SD (for teachers of middle school English) to 0.116 SD (for teachers of elementary school math).[97] This implies that a one SD increase in teacher quality is associated with an increase in student test scores that declines from 0.115 SD to 0.026 SD for elementary-level students in math and from 0.042 SD to 0.009 SD for middle school students in English.[98]

If VAM is used to replace the lowest ten percent of all teachers, any gains in student performance would be limited to ten percent of all students. A hypothetical 0.115 SD gain in performance for ten percent of all students would translate, in the aggregate, to an average 0.0115 SD gain for all students, or approximately six days of learning over one academic year.[99]

With regard to earnings, Chetty et al. estimated that a one SD increase in teacher quality is associated with a 1.34 percent increase in income at age 28, equal to $284 for a single person. Assuming that this differential persists at every age throughout a person's life, Chetty et al. estimated that the cumulative lifetime gain for a single person would equal $6,995 after discounting the gains at an annual rate of three percent.[100] Once again, if VAM is used to replace the lowest ten percent of all teachers, any gains would be limited to ten percent of all students. The policy would translate, in the aggregate, to an average gain in lifetime earnings of $700 per person, averaged across all students.

Chetty et al. estimated that a larger 2.063 SD increase in teacher quality is associated with a $14,500 cumulative lifetime gain for a single person after discounting the gains at an annual rate of three percent, equal to $408,900 for an entire class of 28.2 students.[101] However, as indicated by Chetty et al., only five percent of all teachers may be expected to fall 2.063 SD below the mean level of teacher quality.[102] If VAM is used to identify and replace these teachers, any gains would be limited to five percent of all students. The policy would translate, in the aggregate, to an average gain in lifetime earnings of $725 per person, averaged across all students.

Chetty et al. acknowledge that this figure is overstated because estimation error and drift in teacher quality over time would reduce expected gains.[103] Under more realistic assumptions, a 2.063 SD increase in teacher quality is associated with an expected gain of $226,000 per class of 28.2 students, equal to a $8,014 cumulative lifetime gain for a single person after discounting the gains at an annual rate of three percent.[104] Once again, however, only five percent of all teachers may be expected to fall 2.063 SD below the mean level of teacher quality.[105] If VAM is used to identify and replace these teachers, any gains would be limited to five percent of all students. The policy would translate, in the aggregate, to an average gain in lifetime earnings of $401 per person, averaged across all students.

Chetty et al. found that the lifetime gain for an entire classroom of students equals $266,000 if three years of data are available but, as noted above in Section 3.1, it would not be uncommon for data to be missing in a way that would prevent averaging. For large numbers of teachers, it would be impractical to average their rankings across two or more years.[106]

While the preceding analysis suggests that the impact on lifetime earnings averaged over all students would be small, newspaper accounts focused on the claim that the use of VAM to identify and replace the lowest-performing five percent of teachers with average teachers would translate into much larger gains in the lifetime earnings of their students.[107] Chetty et al. contributed to this confusion by stating:

> The undiscounted cumulative lifetime earnings gains from deselection are 5.5 times larger than these present value gains ($80,000 per student and $2.25 million per classroom)…These simple calculations show that the potential gains from improving the quality of teaching…are quite large.[108]

How can this be reconciled with the view that gains are small? The explanation is that the $80,000 and $2.25 million figures were not discounted to reflect the time value of money. Income received many years in the future is not as valuable as income that is received today. For this reason, economists discount future income streams, effectively reducing the amounts to account for the time value of money. Chetty et al. reported that after discounting at a three percent annual rate, and after accounting for estimation error and drift in teacher quality over time, the lifetime gain of $80,000 per child shrinks to

$8,014; the lifetime gain of $2.25 million for an entire classroom of 28.2 students shrinks to $226,000. The smaller amounts are the appropriate amounts to use in any economic analysis of the benefits and costs of VAM-based policies. Once the $8,014 figure is averaged over all students, it shrinks further, to $401 per person.

Stable Quality Differentials?

As noted, Chetty et al. estimated that substituting an average teacher for a teacher in the bottom five percent of all teachers would result in a lifetime gain, after discounting, equal to $226,000 for a class of 28.2 students taught by that teacher. It may be argued, regardless of the above analysis, that a gain of $226,000 remains significant. However, the working assumption is that a teacher in the bottom five percent consistently performs at a level that is 2.063 standard deviations below an average teacher.[109] This assumption may be questioned.

A 2.063 standard deviation increase in performance might be possible if rankings were stable and rankings in the current year predicted performance in the following year. As Tables 9.1 and 9.2 indicate, however, teacher rankings bounce up and down from year to year. A teacher who ranks in the lowest quartile this year is more likely to rank in the upper three quartiles the next year than to remain in the bottom quartile. Conversely, a teacher who ranks in the highest quartile this year is more likely to drop into the bottom three quartiles the next year than to remain in the top quartile. In a prepublication version of their article, Chetty et al. stated that "one-quarter of the variance in the mean test score residual for a single classroom is driven by teacher quality, with the remaining variance due to classroom and student level noise," implying that 75 percent of the variance in teacher rankings is attributable to random measurement error or performance fluctuations over time, rather than persistent differences in teacher performance.[110] Other researchers found that one-third to one-half of the differentials in teacher performance are driven by random measurement error, rather than true differences in teacher performance.[111] Thus, a teacher who appears to rank 2.063 standard deviations above another teacher is not likely to maintain that differential the following year, and it would not be appropriate to assume that substituting a high-performing teacher for a low-performing teacher would result in the same differential in performance next year. The view that teacher rankings are stable over time and actual gains in student achievement next year would equal the measured differential in performance this year is not supported by the evidence in Tables 9.1 and 9.2. For this reason, it is unlikely that substituting a teacher who performs highly this year would translate into the expected 2.063 standard deviation gain in performance next year. If that gain is not achieved, then the estimated $226,000 gain in lifetime earnings would not be achieved.

Adequate Teacher Supply?

Chetty et al. assume that there is an adequate supply of unemployed teachers who are ready and willing to be hired and would perform at a level that is 2.063 standard deviations above the performance of teachers who are fired based on value-added rankings. Chetty et al. do not justify this assumption with empirical data. The assumption may be questioned. A simple example illustrates that the vacant teaching positions created when low-performing teachers are fired must ultimately be filled with novice teachers whose performance is significantly worse than the performance of experienced teachers.[112] The reason that novice teachers must be hired is because there is a teacher shortage.[113] In the aggregate, there are more positions than qualified teachers and overall teacher demand is projected to exceed supply by 35 percent over the next two decades.[114]

To simplify, suppose that there are ten teaching positions in the entire nation. Suppose that nine of the positions are currently filled with teachers (that is, there is one vacancy). Suppose, further, that value-added methods could be used to reliably identify the lowest-performing teacher ("Teacher Number Nine"), who performs at a level that happens to be 2.063 standard deviations below the performance of Teacher Number One. If Teacher Nine is fired (and not re-hired by any other school), a second vacancy is created. Teacher Nine potentially could be replaced with Teacher Number One, but this action simply shifts the second vacancy to Teacher Number One's school. The process of teacher substitution may continue but, at the end of the process, Teachers One through Eight remain employed. There are now two vacant teaching positions that can only be filled with novice teachers. This is true whether the novice teachers arrive as fresh graduates from teaching colleges or as individuals previously employed in nonteaching occupations who choose to switch into the teaching profession through the alternative certification path. If one of those vacant positions is filled with a novice teacher, then any gain in the average level of student achievement (across Teachers One through Nine) depends entirely on the difference in performance between (the fired) Teacher Nine and the newly-hired novice teacher. If the newly-hired novice teacher outperforms Teacher Nine, then there is a gain in performance; if the novice performs worse than Teacher Nine, there is a loss.

One might ask why low-performing teachers cannot be replaced with experienced teachers who leave the teaching force temporarily, then rejoin at a later date. Suppose, for example, that two experienced teachers rejoin the teaching force every year. Why is it not possible for those teachers to fill the two vacant teaching slots?

This is only possible if there is no teacher shortage. If the supply of teachers to the profession equals the number of vacancies, then no teacher shortage exists. A shortage can only exist if the supply of teachers is less than the number of vacancies. In the example given above, if it is the case that two experienced teachers rejoin the teaching force every year, then it must be the case that two experienced teachers leave the force every year, leaving a single vacancy in the absence of Chetty et al.'s proposal. If two experienced teachers rejoin the teaching force but only one teacher leaves, then the single remaining vacancy would be filled by Teacher Number Ten: all ten teaching positions would now be filled and, therefore, there would be no teacher shortage. Recall, however, that there is currently a teacher shortage, which means that it cannot be true that the net inflow of experienced teachers rejoining the teaching force equals or exceeds the number of vacancies.[115] Furthermore, if there is an inflow of novice teachers, the inflow of novice plus experienced teachers must be less than the number of vacancies, if indeed there is a teacher shortage.

Currently, in the absence of Chetty et al.'s proposal, some of the existing vacancies across the nation are being filled with novice teachers, some vacancies are being filled by experienced teachers who rejoin the teaching force, and at least one vacancy remains (because there is a teacher shortage)—implying that any *extra* vacancies created by Chetty et al.'s proposal *must* be filled with novices. There is no other possible source. *In the presence of a teacher shortage, it cannot be the case that any of the extra vacancies created by Chetty et al.'s proposal will be filled by experienced teachers. Ultimately, after the type of shuffling described above, all of the extra vacancies must necessarily be filled with novices. Therefore, any policy that involves firing low-performing teachers must acknowledge that the vacant positions will ultimately be filled with novices, not experienced teachers.*

Significantly, when value-added methods are used to identify low-performing teachers, replacing these teachers with novice teachers can have unexpectedly negative effects. For example, McCaffrey et al. controlled for student-fixed effects and found that a policy of replacing the bottom 40 percent of all teachers would raise student achievement by 0.04 SD if fired teachers are replaced with teachers performing in the top 60 percent.[116] However, a more realistic assumption is that replacements are novices whose performance is lower than the performance of experienced teachers.[117] Under the assumption that fired teachers are replaced with novice teachers, the overall impact on student achievement across all students would be negative 0.055 SD.[118] The poor result is a direct consequence of the lack of stability in teacher rankings. The use of value-added methods is unreliable in identifying the bottom 40 percent of all teachers; when those methods are employed, many teachers who do not "belong" in the low-performing category are fired, while many teachers who do not "belong" in the high-performing category are retained. The result is a very small gain in aggregate performance that is completely offset by the well-established decrease in performance when large numbers of novice teachers are hired to replace experienced teachers.[119]

Chetty et al.'s main analysis excluded the impact of replacing fired teachers with novices. However, in a footnote, they estimated that the students of novice teachers score 0.05 SD below the students of experienced teachers.[120] This would reduce the previously estimated 0.115 SD increase in math test scores for elementary-level students to 0.065 SD, for every one SD increase in teacher quality. The reduction in impact is significant, but less than alternative estimates. For example, Gordon, Kane, and Staiger estimated that the average "value-added" of novices is about four percentile points lower than teachers with two years of experience, equal to a negative effect size of 0.171 SD.[121] This would reduce Chetty et al.'s estimated 0.115 SD increase in student test scores to negative 0.056 SD. Thus, average student achievement would decrease by 0.056 SD as a consequence of replacing low-performing teachers with novice teachers.

The negative effect of replacing low-performing teachers with novice teachers would decrease as novice teachers gain experience, but any argument that long-term gains would be positive is contingent on the reliability and stability of teacher rankings. There is no empirical evidence that long-term gains are positive, and there is no evidence that long-term gains would outweigh the immediate losses that are incurred when novice teachers replace experienced teachers.

Is VAM Cost-Effective?

Chetty et al. implicitly assume that the use of VAM to identify and replace low-performing teachers is a cost-effective approach for improving student outcomes, where cost-effectiveness is defined by the resulting gain in student achievement for each dollar of resources invested by society. However, two cost-effectiveness studies indicate that VAM is not cost-effective relative to alternative approaches for raising student achievement.[122] Both studies suggest that there are large costs to society of implementing any scheme to replace low-performing teachers: the costs to society of educating new teachers (including their foregone wages), costs incurred by hiring school districts and schools, costs incurred by new teachers, costs incurred by terminated teachers, the reduced output of terminated teachers while learning a new occupation, the opportunity cost of the labor of newly-hired teachers, the costs of adjudicating terminations based on VAM, the cost to raise salaries for all teachers by an amount that would be necessary to attract more individuals to the teaching profession, and the additional cost to implement VAM assessments.[123] These costs would be offset by the output of terminated teachers in new occupations after a period of retraining and job search, but would be substantial.

The termination of a single teacher would create net social costs equal to $314,825.57 (Table 9.3). The annual cost per student equals $15,741, assuming 20 students per teacher.

The largest cost to society is the opportunity cost of replacing terminated teachers with newly-minted college graduates who obtain teaching certification after one additional year of college coursework. The cost to society includes the value of their foregone output in the next best use of their labor. This may be imputed based on the average beginning teacher salary of $40,049.[124] The present value of this stream over the expected career duration of a new teacher (9.11 years), adjusted for a total compensation-to-salary ratio of 1.43, and

Table 9.3 Cost to Society of Replacing One Teacher

Cost to Society of Replacing One Teacher	*2006 Dollars*
1. Cost to Educate the New Teacher	78,952.94
2. Costs to Hiring District and School	
A. Recruiting, administrative, hiring, training	10,625.64
B. Reduced output of new teacher	(unknown)
3. Costs to New Teacher	
A. Job search (cash costs)	705.70
B. Time devoted to job search	5,189.24
C. Relocation	2,684.13
4. Costs to Terminated Teacher	
A. Job search (cash costs)	705.70
B. Time devoted to job search	5,189.24
C. Retraining	2,890.23
D. Relocation to new job	2,684.13
E. Lost output (income)	41,586.89
5. Additional Costs to Society	
A. Reduced output of terminated teacher	5,506.74
B. Foregone output of replacement teacher	456,082.06
C. Social costs of adjudication	116,957.52
6. Gain of Terminated Teacher's New Output	(414,934.59)
TOTAL	$314,825.57

Note: All figures adjusted for inflation to August, 2006, dollars using the consumer price index. Except where indicated in the text, all figures are from Yeh, S. S. (2012), The reliability, impact, and cost-effectiveness of value-added teacher assessment methods, Journal of Education Finance, 37(4), 374-399, Table 4.

assumed to grow at two percent per year (including increases in real income as living standards rise over time as well as seniority-related increases in compensation) but discounted at five percent per year for the present value calculation, is $456,082.06.[125]

This cost to society is offset by the gain in the output of the terminated teachers once they have been retrained and have transitioned into new occupations. While it is not possible to know exactly what occupations the former teachers will transition into, it is reasonable to assume that they will be occupations that require the same level of education (a college degree) and provide roughly the same value of output as teaching. Assuming that retrained workers start in a new occupation at a salary equivalent to a new teacher's salary of $40,049, assuming a compensation-to-salary ratio of 1.43, assuming that wages grow at two percent per year (including increases in real income as living standards rise over time as well as seniority-related increases in compensation) but discounted at five percent per year for the present value calculation, the gain in output to society equals $414,934.59. The income stream begins after an average of 27.36 weeks of retraining[126] and an average of 10.4 weeks to find a new position,[127] lasts a period of 8.38 years, and ends 9.11 years after the date of termination. Thus, the income stream is calculated over the same overall time period as the average duration of the new teacher's expected teaching career.

Society would also incur the costs of adjudicating any disputed terminations. Unlike the proposal by Gordon et al.[128] to use VAM to identify and fire the bottom quartile of novice, untenured teachers (approximately two percent of all teachers), the proposal that is the focus of the current analysis would involve firing a larger percentage of all teachers, a majority of whom would necessarily be tenured teachers who could not be fired without adequate cause. As previously noted, VAM is not reliable for the purpose of categorizing high- and low-performing teachers.[129] Thus, the use of VAM to terminate teachers is likely to result in an avalanche of lawsuits by terminated teachers. The evidence overwhelmingly favors litigants who assert that results based on VAM do not meet the legal standard of adequate cause for termination, suggesting that terminated teachers would be likely to win almost every case, since it would be nearly impossible for school districts to show "adequate cause" for termination based on VAM. Districts would have to fall back on existing methods for identifying poor teachers, which currently result in the involuntary termination of a very small percentage of all teachers. In New York, for example, only 88 out of approximately 80,000 city schoolteachers lost their jobs for poor performance over a three-year period—a rate of 0.037 percent per year.[130] In Los Angeles, only 112 of 43,000 tenured teachers faced termination between 1995 and 2005, a rate of 0.026 percent per year.[131] In New Jersey, 47 of 100,000 teachers were fired over a 10-year period, a rate of 0.005 percent per year.[132] The annual termination rate is 0.01 percent in Chicago, 0.04 percent in Cincinnati, and 0.01 percent in Toledo.[133] In Akron, OH; Denver, CO; Elgin, IL; Jonesboro, AR; and Pueblo, CO; no teachers were formally dismissed over periods that ranged from two to four years.[134] Even if all of these terminated teachers are drawn from the bottom five percent of all teachers subject to termination based on VAM, only small percentages of the VAM-based terminations could be justified based on methods that are independent of value-added rankings: 2.2 percent in New York, 5.2 percent in Los Angeles, 0.94 percent in New Jersey, 0.2 percent in Chicago, 0.8 percent in Cincinnati, and 0.2 percent in Toledo. This implies that litigants who were terminated on the basis of VAM might be expected to prevail in over 94.8 percent of all cases, assuming that courts agree with the National Academies' Board on Testing and Assessment, which concluded that it is not appropriate to use VAM to make operational decisions regarding teacher hiring and firing.[135] As a consequence, and as a consequence of provisions that require teachers to receive their normal pay during the termination process, school districts could expect nearly every termination to be challenged, resulting in enormous costs.[136]

The cost of litigation is high, regardless of the outcome. Tenured teachers often must be provided with names of witnesses, the power of subpoena to compel production of documents and testimony of witnesses, the right to counsel at all stages of the process, and the right to appeal.[137] It is estimated that the average cost of terminating a teacher in California is approximately $200,000.[138] In San Diego, a single termination proceeding took more than four years and cost more than $300,000 in legal fees.[139] In New York, section 3020-a of the state education law allows a tenured school district employee who has been charged with incompetence or misconduct to request that a hearing officer review the district's charges, and make findings of fact and recommendations as to penalty or punishment, if warranted. In general, "the cost and time required to terminate a permanent teacher are extreme."[140]

On average, a full 3020-a hearing costs New York districts $216,588 and takes 502 days, according to a New York State School Boards Association survey of 400 districts from 2004 to 2008.[141] This survey provided a breakdown of costs that permits adjustments to reflect the true social costs. The largest expense was the salary and fringe benefits paid to the suspended employees, accounting for 52 percent of costs. Salaries and benefits for substitute teachers represented 30 percent of the costs, while legal expenses represented 12 percent of the costs. Other expenses included other staff costs (five percent) and miscellaneous costs such as the cost of outside investigators, expert witnesses, transcription, photocopying, and travel (one percent). However, since the salary and benefits of the suspended employees would have been paid in the absence of the disciplinary hearings, I reduced the total cost figure by 52 percent to reflect the real social cost incurred by each district, equal to $103,962.24.

In addition to the costs incurred by each district, the suspended employees (or their unions) incurred legal expenses that may be expected to average approximately half the legal expenses incurred by their school districts, equal to $12,995.28 per case. The total social cost of each hearing equals $103,962.24 plus $12,995.28, or a total of $116,957.52 per terminated teacher. This excludes psychic costs incurred by terminated teachers, as well as the cost of any appeals, which could double the cost.

The annual cost of implementing a value-added assessment system may be estimated from the costs of administering and scoring the assessments for Tennessee's Value-Added Assessment System (TVAAS): $5.60 per student, adjusted for inflation and including the cost of the TVAAS reports.[142] The total cost of assessing all students nationwide equals the national student population multiplied by $5.60. After spreading the total cost over the five percent of all students that would benefit from Chetty et al.'s proposal and calculating the cost per student in this five percent subset of all students, the cost increases to $112 per student in the five percent subset (20 × $5.60).

In addition to the cost of the assessments, salaries must be raised for all teachers in order to attract more individuals to the profession of teaching. This cost is above and beyond the cost to educate and train the new teachers, since there is no army of unemployed teachers waiting to fill the empty teaching slots. On the contrary, there are shortages in many subject specialties and overall teacher demand is projected to exceed supply by 35 percent over the next two decades.[143]

The increase in teacher salaries required to attract a sufficient number of new individuals to the teaching profession may be estimated using conservative assumptions. Suppose, for example, that VAM is used to identify and replace the bottom five percent of all teachers. If the stock of teachers is denoted by Q, the value-added proposal implies that Q must be increased by 5.26 percent to an amount equal to 1.0526Q (elimination of the bottom five percent of teachers reduces 1.0526Q to Q). The cost is determined by the elasticity of teacher supply, defined as the percentage change in the quantity of teachers that is elicited by a one percent change in annual teacher salary: $\%\Delta Q/\%\Delta S$. I assumed a supply elasticity of three, which is near the top of the range of ordinary supply elasticities estimated by Manski.[144] However, the correct elasticity is likely to be lower because the use of value-added methods to fire the bottom five percent of all teachers increases the risk of being fired, making teaching a less desirable career choice. Thus, a one percent salary increase is likely to be insufficient to induce a three percent increase in the supply of new teachers, implying that the estimate of the required increase in teacher salaries is likely to be a lower-bound estimate of the true cost.

A supply elasticity of three implies that teacher salaries must increase by 1.753 percent (5.26/3) to elicit the number of new teachers required to replace the bottom five percent of all teachers. Assuming an average teacher salary of $51,055.19 per year after adjusting for inflation,[145] a compensation-to-salary ratio of 1.43[146] and assuming 20 students per teacher, it would cost an extra $64 per student per year to raise salaries sufficiently to attract the teachers necessary to replace the bottom five percent of all teachers.

The total cost to raise salaries for all teachers nationwide equals the national student population multiplied by $64. After spreading the total cost over the five percent of all students that would benefit from Chetty et al.'s proposal and calculating the cost per student in this five percent subset, the cost increases to $1,280 per student in the five percent subset (20 × $64).

The total annual cost of implementing this proposal is the cost to society of replacing a terminated teacher through a fifth-year teacher education program ($15,741), plus the cost of the assessments ($112), plus the cost to raise salaries sufficiently to replace the bottom five percent of all teachers ($1,280), or a total of $17,133 per student. This figure is underestimated to the extent that fired teachers incur psychic losses, and to the extent that the increased occupational risk of entering the teaching profession that is implied by firing five percent of all teachers each year would drive teacher salaries upward, raising the cost of hiring new teachers as well as the cost of employing existing teachers.

Cost-Effectiveness Results

If terminated elementary school math teachers who are currently performing 2.063 SD below the level of an average teacher are replaced with novice teachers whose students perform 0.05 SD below the students of an average teacher, the gain in achievement equals $[(0.993 \times 0.116 \times 2.063) - 0.05] = 0.188$ SD, under the assumption that VAM reliably identifies low-performing teachers in the tail of the teacher performance distribution. The effectiveness-cost ratio, 0.188 SD divided by $17,133, equals 0.00001. If terminated teachers are replaced with novice teachers whose students perform 0.171 SD below the students of experienced teachers, the average gain is 0.067 SD per year and the effectiveness-cost ratio equals 0.000004.

To determine whether teacher replacement is a cost-effective strategy, it is necessary to compare the approach with other strategies. With regard to the field of education, a cost-effective intervention may be defined as the approach that offers the largest impact with regard to student achievement in math and reading for each dollar invested by society in that intervention.[147] Using this definition, teacher replacement is not cost-effective. The effectiveness-cost ratio for rapid performance assessment, an alternative strategy for improving student outcomes, ranges from 0.017152 to 0.028571.[148] The smallest ratio for rapid performance assessment is 1,700 times larger than the largest ratio for Chetty et al.'s VAM-based teacher replacement strategy (0.00001), implying that Chetty et al.'s strategy is not a cost-effective approach for raising student achievement.

Benefit-Cost Results

With regard to earnings, Chetty et al.'s proposed intervention does not meet the test of a benefit-cost analysis. As indicated in Section 3.4, if VAM is used to identify and replace the lowest five percent of all teachers, a 2.063 SD increase in teacher quality in a single grade is associated with an average gain in lifetime earnings of $8,014 per student, in 2010 dollars.[149] To maintain consistency with cost calculations which assumed 2006 as the base year, the $8,014 figure was adjusted using the Consumer Price Index to 2006 dollars.[150] The resulting figure ($7,485), divided by the cost per student in 2006 dollars to implement this intervention ($17,133), produces a benefit-cost ratio equal to 0.44. Society would gain $0.44 for every dollar invested in the intervention, implying that the costs of the intervention exceed the benefits by a ratio of 2.3 to one.[151] If tenure is eliminated and all teachers are employed at-will, litigation costs may be excluded. The cost per student falls to $11,285, but the benefits of the intervention ($7,485) remain less than the costs. This analysis also applies to the case where the proposed policy is only applied to novice teachers who are not tenured. If the policy is applied to a mixture of experienced and novice teachers, the costs of the policy would range between $11,285 and $17,133 per student. However, the policy does not meet a benefit-cost test at any point in this range. This verdict holds whether the proposed policy is implemented once, or on an ongoing basis, because both the costs and the presumed benefits are incurred every time the policy is implemented, so the ratio of benefits to costs (or the ratio of effect size to costs) would remain unchanged.

Implications

The studies reviewed in the first section of this chapter, together with the analysis in the second section, suggest that the use of value-added statistical methods to identify and replace low-performing teachers is not warranted.

VAM lacks sufficient reliability and validity for the purpose of hiring and firing teachers. Once gains are averaged over all students, they would be very small. Furthermore, it appears that any gains would fade away very quickly. Significantly, the approach is neither cost-effective nor does it meet the test of a benefit-cost analysis.

While the preceding analysis is based on Chetty et al.'s study, much of the analysis applies to any proposal to use value-added methods to replace low-performing teachers. Studies of the stability of VAM-based teacher rankings have found inadequate reliability for operational decisions regarding the hiring and firing of teachers.[152] Even when studies of VAM are taken at face value, the results suggest small impacts on student achievement of policies that would employ VAM to identify and fire low-performing teachers.[153] When these results are integrated with analyses of the full social costs of implementing VAM in order to replace low-performing teachers, it becomes clear that VAM is not cost-effective relative to the most promising strategies for raising student achievement.[154]

These results suggest a need to revisit the assumption that large improvements in student outcomes may be achieved by identifying and replacing low-performing teachers. This assumption suggests that high- and low-performing teachers are analogous to "good apples" and "bad apples," and implies that average teacher quality would improve if we get rid of bad apples. The assumption is that teacher quality is a fixed characteristic—that a high-performing teacher this year will be a high-performing teacher next year, and a low-performing teacher this year will be a low-performing teacher next year. As indicated above, this assumption is not supported by the available data. Teacher quality is not a fixed, inherent characteristic but instead fluctuates over time and is variable in a way that is not captured by a model that categorizes workers as "good apples" and "bad apples."[155] Much of a teacher's performance varies over time due to unobservable factors such as effort, motivation, and class chemistry that are not easily captured through VAM.[156]

Advocates of using VAM for high-stakes decisions regarding teacher hiring and firing argue that an excessive concern with false identifications of low-performing teachers serves the interests of teachers, rather than their students.[157] Framing the issue in this way, however, sets up a false dichotomy. The question is not whether society should serve the interests of teachers rather than their students, or the proper balance between false positive and false negative identifications, but *what is the most efficient approach for raising student achievement*? A number of cost-effectiveness analyses have now been performed that permit comparison of 22 of the leading approaches for raising student achievement.[158] The results from the cost-effectiveness analysis suggest that the most efficient approach—rapid performance feedback—is approximately 1,700 times as efficient as the use of VAM to identify and replace low-performing teachers.

This result may appear to be improbable. There are two reasons for the tremendous disparity in efficiency. First, the particular variant of rapid performance feedback that is the focus of the comparison (labeled "rapid assessment" programs) involves changes in the way learning material is individualized and presented to students, in combination with performance feedback in the form of individualized daily assessments. These changes apparently alter students' perceptions of their abilities to improve their performances, so that low-performing students begin to believe that they can achieve academic success through their own efforts.[159] Students appear to acquire an internal locus of control, exerting more effort than students who do not receive rapid performance feedback.[160] This approach offers a different way of thinking about how student performance may be improved. In contrast, VAM-based teacher replacement policies attempt to improve student achievement without addressing the psychology of student learning.

A second reason for the disparity in efficiency between VAM and rapid assessment is that rapid assessment is primarily implemented with the aid of computer software, the cost of which can be amortized over multiple years and spread over hundreds of students in each school building. The annual cost per student is very low. In contrast, as indicated by the cost-effectiveness analysis, the use of VAM to identify and replace low-performing teachers is tremendously costly.

In contrast to the rapid performance feedback model, the use of VAM to identify and replace low-performing teachers relies on the conventional model of instruction, which fails to individualize task difficulty and therefore fails to change the tedious experience of schooling for students who are above-average and the discouraging experience of schooling for students who are below-average. Failing to address these dynamics, VAM-based policies place the entire burden of raising student achievement on teachers who are locked into systems that appear to inadvertently undermine student engagement and achievement. The available evidence suggests that this approach is neither effective nor efficient.

CHAPTER 10

22 Strategies

This chapter draws upon the author's previously-published evaluation studies and book-length evaluation of 22 interventions for raising student achievement.[1] The focus in this chapter is on the major results of the book-length evaluation. Details regarding the evaluation methods and the interventions that were evaluated are described in the book and associated journal articles and are not repeated here.

The purpose of this chapter is to give the reader a basic understanding of what is currently known about the relative cost-effectiveness of the major interventions. The key finding is that a technology-based intervention that individualizes task difficulty, provides rapid performance feedback, and is based on the theory of the achievement gap described in this book is dramatically more cost-effective than any of the other 22 interventions for which data are available. This result offers key evidence supporting the proposed theory of the achievement gap. It implies that a given amount of social resources would provide maximum benefits in closing the gap if it is directed toward the implementation of a technology-based intervention that individualizes task difficulty and provides rapid performance feedback, rather than the implementation of any of the other 22 interventions. The efficiency of this approach implies that the underlying theory is correct. It is extremely unlikely that the underlying theory could be incorrect, yet yield the most efficient approach for raising student achievement. In general, only a correct theory can produce an efficient solution.

The results that are reported in this chapter are significant because the acid test of any theory is whether an effective intervention or useful prediction can be developed based upon the theory. The ability to develop a highly-efficient intervention largely precludes the possibility that the theory omits a key variable, incorrectly specifies the model of achievement, or mistakes correlation for causation. The ability to manipulate key variables through an efficient intervention and obtain the predicted result is prima facie evidence that the key variables are operating as predicted to cause the predicted result. This can only occur if the underlying theory is correct.

S.S. Yeh, *Solving the Achievement Gap*,
DOI 10.1057/978-1-137-58767-1_10

The results reported in this chapter will undoubtedly raise questions about the dramatic differences in reported efficiency. In some cases, the individualized difficulty/rapid assessment intervention appears to be thousands of times more efficient than the alternatives. How is this possible? I previously addressed this question:

> In essence, I argue that researchers have overlooked a pattern of research findings that point away from current beliefs regarding "what works" and points toward a very specific problem underlying poor student engagement and achievement. The current failure to identify this problem is analogous to a hypothetical failure of medical professionals to recognize an epidemic disease and, instead, to attribute high rates of mortality to inadequate resources, accountability, and competition. However, if the cause of illness is a specific disease, it makes little sense to increase funding indiscriminately, increase accountability for doctors and patients, increase competition, reform hospitals, stiffen credentialing of doctors, replace the bottom quartile of doctors, and reduce overcrowding in hospitals. While an argument can be made for each of these proposals, they clearly would be less effective than an intervention based on a precise identification of the specific causative disease.[2]

In the given example, one would expect that indiscriminate increases in medical funding, increased accountability for doctors and patients, increased competition among hospitals, hospital reform, stiffened credentialing of doctors, replacement of the bottom quartile of doctors, and reduced overcrowding in hospitals would be extraordinarily inefficient strategies in combating an epidemic disease such as Ebola. One would expect that a vaccine that targets the Ebola virus would be thousands of times more efficient in reducing mortality. While the cost of developing a vaccine may be substantial, once a vaccine is developed, the marginal cost of producing an additional dose is small. In this regard, it is significant that the previously-described technology-based intervention to individualize task difficulty and provide rapid performance feedback has already been developed. The costs of development are sunk and have already been incorporated into the unit price of the product and the cost calculations utilized in the cost-effectiveness comparisons. However, the main explanation for the vast difference in reported efficiency is that individualization of task difficulty and rapid performance feedback appear to be based upon an accurate theory of the cause of the achievement gap.

The analogy with epidemic disease is apt because the 22 alternative strategies for addressing the achievement gap are analogous to indiscriminate increases in medical funding, increased accountability for doctors and patients, increased competition among hospitals, hospital reform, stiffened credentialing of doctors, replacement of the bottom quartile of doctors, and reduced overcrowding in hospitals. The notion that minority children attend schools that are underfunded has a long history, has received enormous attention, and has been reinforced through landmark court decisions that found that the states of California, Kentucky, New Jersey, Ohio, North Carolina, Maryland,

and New York violated constitutional provisions guaranteeing adequate educational opportunities to children in those states.[3] The results reported in this chapter indicate, however, that the strategy of raising educational expenditures is an extremely inefficient approach for raising student achievement.

The notion that it is necessary to enforce accountability for teachers has gained currency to the point where it is perhaps the most popular and widely-accepted theory about the cause of low achievement. Accountability might be enforced through various strategies, including the use of value-added statistical modeling to identify and replace the bottom quartile of teachers, the use of NBPTS teacher certification to identify and retain high-quality teachers, or the use of a test score requirement to screen teachers at the point when they are hired. The results reported in this chapter indicate, however, that all three of these strategies would be extremely inefficient approaches for raising student achievement.

The notion that it would be helpful to enforce accountability for students has a long history and has been implemented in the form of exit examinations that students must pass in order to graduate from high school. The results reported in this chapter indicate, however, that enforcing accountability for students is an extremely inefficient strategy for raising student achievement.

The notion that it would be helpful to raise the level of competition experienced by schools has been implemented in the form of voucher programs and charter schools. Voucher programs link government funding to students who are free to enroll at any school of their choice within a designated catchment area. The receiving school receives the funding. Therefore, schools must compete with each other in order to attract students and the funding that is linked to each student. Charter schools are open to any student within their home school districts. Therefore, all schools within those districts must compete to attract and retain students and the district funding that accompanies the students. The results reported in this chapter indicate, however, that the strategy of promoting competition among schools is an extremely inefficient approach for raising student achievement.

The notion that it would be helpful to reform schools has been implemented through federal schoolwide reform programs that provide funding to schools that adopt a federally-approved school reform model such as the *Success For All* program.[4] The results reported in this chapter indicate, however, that even the most effective and efficient school reform model is an extremely inefficient approach for raising student achievement.

The notion that it would be helpful to stiffen the credentialing process for teachers receives little support. In particular, the results reported in this chapter indicate that NBPTS certification—perhaps the most rigorous and promising of the credentialing approaches—is an extremely inefficient approach for raising student achievement.

Finally, the notion that it would be helpful to reduce class sizes has a long history and receives support from studies indicating that class size reduction does indeed improve student achievement. However, the results reported in this

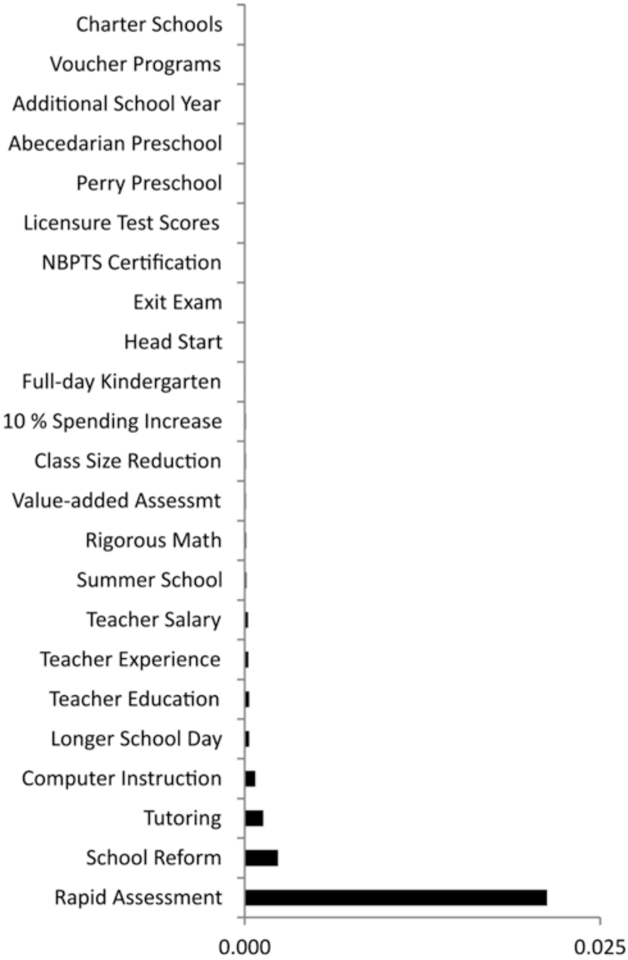

Figure 10.1 Relative magnitude of 23 effectiveness-cost ratios

chapter indicate that class size reduction is an extremely inefficient approach for raising student achievement.

Figure 10.1 compares the cost-effectiveness of 23 approaches for raising student achievement. Each bar represents the magnitude of the corresponding effectiveness-cost ratio, defined as the annualized effect size in standard deviation units of student achievement, divided by the annualized opportunity cost per student of implementing the corresponding approach for raising student achievement.

Figure 10.1 indicates that the technology-based individualized task difficulty/rapid performance feedback intervention (labeled "rapid assessment") is dramatically more efficient than any of the other 22 approaches for raising student achievement. The dramatic difference in efficiency can only be explained if the approach reflects an accurate understanding of the cause of low student achievement and the alternative approaches reflect inaccurate understandings, much as physicians would expect a vaccine against Ebola to be dramatically more efficient in reducing mortality from Ebola, compared to indiscriminate increases in medical funding, increased accountability for doctors and patients, increased competition among hospitals, hospital reform, stiffened credentialing of doctors, replacement of the bottom quartile of doctors, and reduced overcrowding in hospitals.

Of the 23 approaches for raising student achievement, individualized task difficulty/rapid performance feedback is the only approach that conceptualizes the problem of low achievement as a problem of student demoralization caused by existing grading and testing practices. By analogy, a public health strategy that focuses on the development and broad administration of an Ebola vaccine conceptualizes the problem of Ebola as a problem involving a specific virus requiring a specific vaccine, instead of indiscriminate increases in medical funding, increased accountability for doctors and patients, increased competition among hospitals, hospital reform, stiffened credentialing of doctors, replacement of the bottom quartile of doctors, and reduced overcrowding in hospitals. It should not be surprising that unfocused, poorly-targeted strategies based on false understandings of the nature of the problem are extraordinarily inefficient in addressing the problem, whether the problem is Ebola or low student achievement. Policymakers would not expect indiscriminate increases in medical funding, increased accountability for doctors and patients, increased competition among hospitals, hospital reform, stiffened credentialing of doctors, replacement of the bottom quartile of doctors, and reduced overcrowding in hospitals to have any effect in reducing mortality from Ebola. Similarly, policymakers should not expect indiscriminate increases in education funding, increased accountability for teachers and students, increased competition among schools, school reform, stiffened credentialing of teachers, replacement of the bottom quartile of teachers, and reduced overcrowding in classrooms to have any effect in addressing the demoralization, disengagement, and reduction in achievement that stems from existing grading and testing practices and is occurring among almost all American students and particularly among low-achieving black and Hispanic students.

It is difficult to explain the demoralization and disengagement among students that are documented in Chapter Four, except as a consequence of existing grading and testing practices. Any parent or teacher can attest to the psychological impact when a child receives poor grades. There is arguably no other experience that is as powerful and occurs on a regular basis in American classrooms across the nation, across grades K through 12, across race, ethnicity, and gender, in urban, suburban, and rural classrooms, and across every demographic category and geographic locale. Students are regularly subjected to classroom experiences and homework assignments that may be boring and uninspiring, but these experiences do not compare to bad grades. Bullying exists, as does the phenomenon of oppositional peer culture, and can be the cause of demoralization and disengagement among a fraction of students, but cannot explain the persistence of the achievement gap over the K-12 years throughout the entire nation. Perhaps the only other viable explanation is the possibility that disadvantaged minority students experience significantly worse teachers and schools. However, the results of the analyses reported in Chapters Three, Four, Five, and Six and the cost-effectiveness comparison reported in Figure 10.1 do not support the conclusion that teacher and school quality can explain the persistence of the achievement gap. Identifying and replacing low-performing teachers using VAM is not cost-effective. The strategy of identifying and retaining NBPTS certified teachers is not cost-effective. Screening teachers using teacher licensure scores is not cost-effective. The analysis in Chapter Six suggests that the most likely explanation for the existence of teachers who exhibit large estimated contributions to student achievement is that the statistical models used to calculate these contributions omit one or more important variables. The most likely explanation is that the models mistakenly lump together the effect that occurs when a portion of teachers receive, by chance, entire classrooms filled with high-self-efficacy students, together with the effect of teacher contributions to student achievement, in a way that conflates the two effects and mistakenly attributes the entire effect to teacher skill and effort.

If one accepts that the teacher rankings produced by VAM are not reliable for the purpose of predicting future teacher performance and the cause of the lack of reliability is that VAM omits a measure of self-efficacy, the results displayed in Figure 10.1 begin to make more sense. The failure to identify a strong predictor of teacher performance, and the failure to identify an efficient intervention to improve teacher quality, is not simply due to a deficiency in existing technology that can be corrected. Instead, the focus on teacher quality reflects a basic misunderstanding of the primary factors influencing student achievement. The primary factor appears to be the degree to which students are immersed in an environment that is individualized and designed to present challenging but not overwhelmingly difficult tasks to each student each day, in combination with rapid performance feedback that signals when progress has been achieved, so that all students can achieve high math accuracy scores and high reading comprehension scores on a daily basis, feel a strong sense of accomplishment, and feel a sense of satisfaction that they are competent, capable individuals who are learning new material and rapidly advancing every day. The conventional model of schooling and all strategies for raising student achievement other than the rapid assessment intervention are characterized by a failure to create this task environment. The rapid assessment intervention provides technology that helps teachers to create this task environment. This may explain why the rapid assessment intervention is dramatically more efficient than any other strategy for the purpose of raising student achievement.

CHAPTER 11

Solving the Achievement Gap

Chapter One reviewed studies by Nobel laureate James Heckman and his co-authors. Those studies focused on the hypothesis that personality traits, presumably shaped at an early age by sociocultural influences and parenting style, were responsible for differences in achievement that exist at entry into kindergarten—differences that persist throughout the K-12 years and are assumed by Heckman to be caused by parenting style rather than school influences. How can Heckman's arguments be reconciled with the evidence presented in this book that the persistence of the gap is due to school-related factors—i.e., the harmful psychological effects that occur when students are graded, tested, and compared to their same-age classmates?

The answer lies in Heckman's published studies. The key study upon which Heckman relies for his conclusions about the influence of personality was published in 2011 as a 182-page report in the *Handbook of the Economics of Education*, co-authored by Heckman with Mathilde Almlund, Angela Duckworth, and Tim Kautz, titled "Personality Psychology and Economics."[1] The relevant portion is section seven, titled "The Predictive Power of Personality Traits," which discusses the empirical evidence regarding the power of personality in predicting life outcomes. Heckman and his co-authors acknowledge that "many studies" covered in their review "do not address the question of causality, that is, does the measured trait cause (rather than just predict) the outcome?" and acknowledge that simple "empirical associations are not a reliable basis for policy analysis."[2]

The subsection titled "Educational Attainment" focuses on three studies involving nationally-representative samples of individuals. Heckman and his co-authors rely on these three studies for the assertion that personality traits predict level of schooling. In these studies, however, personality was not assessed during early childhood. Instead, personality was assessed as an adult, after all K-12 schooling had been completed and, thus, after any influences from K-12 schooling had shaped personality.[3]

S.S. Yeh, *Solving the Achievement Gap*,
DOI 10.1057/978-1-137-58767-1_11

Suppose that grading practices during the K-12 years erode self-efficacy, demoralize low-performing students, and convert many individuals who enter kindergarten as happy, enthusiastic children into apathetic, depressed, withdrawn 18-year-olds. This hypothesis is consistent with Heckman's data that demonstrate an association between personality traits measured in adulthood and level of schooling. Apathetic, depressed, withdrawn adults are unlikely to pursue advanced levels of schooling. In this example, however, it is the harmful effects of K-12 schooling that cause reductions in educational attainment, rather than personality traits acquired at a young age.

Heckman and his co-authors review few studies involving preschool children, relying heavily on studies where measures of personality were obtained during or after the K-12 years. In any case, none of the studies cited by the authors establish a link between personality traits established prior to the harmful effects of K-12 grading practices, and levels of education attained as adults. Instead, the studies cited focus on academic achievement measured by test scores obtained during the K-12 years. The authors cite one study involving preschool children where teacher ratings of effortful control predicted standardized achievement test scores in kindergarten.[4] Another study found an association between preschool performance on a task that required children to touch their head, toes, knees, and shoulders, and later performance on standardized achievement tests.[5] Perhaps the most famous study of effortful control demonstrated that the number of seconds a child waits for a preferred treat in a preschool test of delay of gratification predicts the child's SAT college admission test score more than a decade later.[6]

Heckman's strongest evidence is a study by Greg Duncan and his colleagues, who analyzed six large longitudinal datasets and found that attention skills, measured at school entry, predict subsequent achievement test scores, measured in various grades after school entry.[7] However, it is worth quoting directly from the report of this study:

> Given that teachers emphasize the importance of attention skills and socioemotional behavior for school readiness and the possibility that these skills shape classroom learning processes, it might be expected that these early skills would have crossover effects on subsequent reading and math achievement. With the important exception of attention skills, we did not find evidence that changes in these skills during the preschool years predict later achievement . . . The average effect sizes of externalizing and internalizing problem behaviors and social skills were close to zero . . . Our results suggest that attention skills, but not problem behavior or social skills, predict achievement outcomes.[8]

Duncan's report indicates that attention skills, but not other personality traits, predict achievement test scores.

To summarize, three studies demonstrate a link between effortful control, measured prior to school entry, and later academic achievement, and one large study demonstrates a link between attention skills, measured prior to school entry, and later academic achievement. However, these studies are consistent with the thesis that early differences in personality influence achievement at entry into kindergarten but this initial influence is maintained, perpetuated, and magnified after school entry because grading and testing practices systematically demoralize and depress low-performing children over the entire 13-year K-12 period. The studies that Heckman relies upon are consistent with the thesis that in the absence of those grading and testing practices, much if not all of the gap in achievement that exists upon entry into kindergarten would disappear. The possibility remains that the primary reason that researchers observe correlations between personality measured prior to school entry and academic achievement measured after school entry is that grading and testing practices demoralize low-achieving students and drive a wedge between high- and low-achieving students starting at entry into kindergarten and continuing throughout the entire 13-year K-12 period.

If attention is confined solely to correlational studies, it is not possible to determine whether Heckman's thesis is correct or whether the thesis presented in this book is correct. To determine which theory is correct, it is necessary to perform rigorous evaluation studies of interventions that are based on each theory and then compare not merely the relative effectiveness of each intervention but the relative cost-efficiency, defined as the annualized effect size of student achievement gains divided by the annualized opportunity cost per student, in other words, the gain in student achievement per dollar of social resources that are required to implement the corresponding intervention. That information was provided in Chapter Nine. The results indicate that a technology-based intervention that individualizes task difficulty and provides rapid performance feedback (labeled "rapid assessment") is dramatically more efficient than either the Perry preschool program or Abecedarian preschool program. These preschool programs are typically classified as high-quality preschool programs. The results indicate that an intervention based on the thesis presented in this book is dramatically more efficient than interventions based on Heckman's theory. The efficiency of the rapid performance feedback approach implies that the underlying theory is correct. It is extremely unlikely that the underlying theory could be incorrect, yet yield the most efficient approach for raising student achievement. In general, only a correct theory can produce an efficient solution. The existence of a highly-efficient intervention largely eliminates the possibility that the theory omits a key variable, incorrectly specifies the model of achievement, or mistakes correlation for causation. The ability to manipulate key variables through an effective intervention and obtain the predicted result is prima facie evidence that the key variables are operating as predicted to cause the predicted result. In general, this can only occur if the underlying theory is correct.[9]

Why does this matter? It matters because policymakers need information regarding the most efficient strategy for addressing the achievement gap. If early differences in effortful control and attention create initial differences in achievement that exist upon entry into school, but the primary factor causing those differences to be magnified and to persist through the end of high school is the demoralizing effect of existing grading and testing practices, then the most efficient strategy for addressing the achievement gap is likely to be the implementation of technology that helps teachers to individualize task difficulty, permits all students to receive high math accuracy scores and high reading comprehension scores on a daily basis, permits all students to receive high grades and permits all students to be successful and feel successful on a daily basis, instead of funding extraordinarily expensive high-quality preschool programs whose positive effects would essentially be nullified by demoralizing grading and testing practices inflicted on students over the 13-year period from kindergarten through the end of high school. The latter approach is analogous to an attempt to fix poor performance by purchasing and installing an extremely expensive engine in a car where the brakes are always engaged. It would be more efficient to simply release the brakes. And once the brakes are released, the driver may find that there is no need to buy a new engine.

Conclusion

Why has the perspective offered in this book not received more attention? The differences that are observed when children enter school in kindergarten and the apparent differences in the quality of schools experienced by black and white children make it easy to conclude that culture, family advantages, and school quality are sufficient to explain the achievement gap. The impression is that there is no need to reexamine the evidence.

My hope is that the evidence presented in this book has persuaded the reader that there are numerous discrepancies indicating that sociocultural and socioeconomic explanations and explanations focusing on school and teacher quality are inadequate for the purpose of explaining the persistence of the achievement gap. While culture and family advantages may explain why certain children enter kindergarten with an academic head start, they do not explain why that academic advantage persists and grows throughout the K-12 years. If white children enter kindergarten with a head start, why does their self-efficacy and engagement decline, why does the decline accelerate, and how can this be reconciled with the persistence of the achievement gap? If Asian children have parents who emphasize the value of education, why does their self-efficacy and engagement decline, why does the decline accelerate, and how can this be reconciled with the persistence of the achievement gap? If, on average, high-performing children experience higher-quality schools and teachers than low-performing children, why does the self-efficacy and engagement of high-performing children decline, why does the decline accelerate, and how can this be reconciled with the persistence of the achievement gap?

One might be tempted to cobble together an explanation that avoids the loss of control hypothesis. One could argue, for example, that children naturally enter kindergarten full of optimism and an inflated sense of their capabilities. A decline in self-efficacy over the K-12 years might be expected and might accelerate as children encounter increasingly difficult academic tasks. Engagement might be expected to decline for the same reasons. It might be expected that the decline would be more severe for children who enter

S.S. Yeh, *Solving the Achievement Gap*,
DOI 10.1057/978-1-137-58767-1

kindergarten performing below their same-age peers, resulting in a gap in achievement that persists over the K-12 years. In this view, a decline in self-efficacy and engagement is inevitable and unavoidable if children are exposed to increasingly difficult academic tasks.

However, it is clear from the research regarding the rapid assessment programs that a decline in self-efficacy and engagement is not inevitable, and the performance of low-achieving students can be remedied, simply by using technology that adjusts the level of difficulty of the books and math problems that are presented to each student and ensures that each student receives rapid performance feedback regarding his or her reading comprehension and math accuracy. The view that a decline in self-efficacy and engagement is inevitable and unavoidable if children are exposed to increasingly difficult academic tasks is incorrect. What was previously assumed to be fixed has been shown to be malleable.

Research demonstrates that it is possible to increase engagement, effort, and performance, even when the tasks presented to each student become progressively more difficult, if technology is used to individualize task difficulty and supply performance feedback on a daily basis. Conversely, it is possible to take the same students and reduce their engagement, effort, and performance by supplying tasks that are too difficult or too easy, and by eliminating performance feedback. Essentially, it is possible to turn engaged, motivated students into disengaged, unmotivated students by supplying conditions that mimic conditions that characterize the conventional model of schooling: uniform task difficulty and lack of performance feedback.

The hypothesized mechanism explains the broad decline in self-efficacy and engagement that occurs across race, ethnicity, and levels of achievement during the K-12 years. The mechanism explains the puzzling decline in self-efficacy and engagement for white students, Asian students, and high-performing students. The same mechanism explains the achievement gap. The mechanism explains why a disproportionate fraction of minority students become disengaged and drop out. The mechanism explains why black males drop out at high rates and why their dropout rates fall below the rates for white and Asian females after controlling for grades. The mechanism is consistent with the array of empirical results described in Chapters Three, Four, Five, and Six. This array of results forms a fine-grained pattern that is difficult to explain in any other way, except in terms of the hypothesized mechanism underlying the achievement gap.

The same phenomena are difficult to explain if race, ethnicity, sociocultural influences, or socioeconomic factors are assumed to be the primary causes of the student achievement gap. Explanations based on these assumptions fit some but not all of the facts. If sociocultural or socioeconomic factors are the cause, it is difficult to understand why black children enter the school system with a level of self-efficacy that exceeds the level of white children but become progressively demoralized and disengaged and drop out at high rates, yet drop out at rates lower than white or Asian children at every level of GPA.

Instead of a convoluted explanation involving culture and socioeconomic influences, a much simpler explanation is that there is something about the conventional structure of schooling that causes self-efficacy and engagement to decline broadly across race, ethnicity, and every level of student achievement. Furthermore, it cannot be the case that low-quality schools are the culprit. If low-quality schools were the cause, then the pattern would not be exhibited in high-quality schools. This is not the case.

The evidence points to a structural factor that is embedded in the conventional model of schooling, rather than parenting style or the quality of individual schools. Only a structural factor could exert a systematic influence across race, ethnicity, and level of achievement. This factor must be something that is hidden in plain view because it would not be possible to hide anything this powerful. It is hidden only in the sense that we do not recognize it.

Fixed task difficulty and the use of grades to rank, compare, and categorize students as either above- or below-average characterize the conventional model of schooling and are known to depress engagement and achievement. These features are invisible to the casual observer because they are structural features of every school. It is difficult for most observers to imagine schools without these characteristics.

This book has sought to demonstrate that these characteristics are not benign. Instead, they have extremely powerful unintended negative effects on children. Recognition of this central fact is the key to unraveling the mystery of the achievement gap, understanding the source of the problem, and formulating effective public policies to address it.

Appendices

Appendix A

S.S. Yeh, *Solving the Achievement Gap*,
DOI 10.1057/978-1-137-58767-1

Table A.1 Partial Correlations Between Math Achievement and Math Self-Efficacy, Controlling for Race and Sex

	Grade 3 Math Achievement	*Grade 5 Math Achievement*	*Grade 5 Math Achievement*	*Grade 8 Math Achievement*	*Grade 8 Math Achievement*	*Grade 5 Math Self-Efficacy*	*Grade 8 Math Self-Efficacy*
American Indian/ Alaska Native	–0.084	–0.064	–0.063	–0.072	–0.073	–0.005	–0.022
Pacific Islander	–0.027	–0.007	–0.006	–0.008	–0.003	–0.041	–0.007
black	–0.253	–0.282	–0.283	–0.282	–0.280	–0.017	–0.020
Hispanic	–0.136	–0.106	–0.104	–0.094	–0.094	–0.017	–0.048
Asian	0.069	0.085	0.084	0.082	0.082	0.012	0.047
sex	–0.078	–0.079	–0.079	–0.021	–0.019	–0.121	–0.085
Grade 3 self-efficacy	0.181	0.150					
Grade 5 self-efficacy			0.292	0.261			
Grade 8 self-efficacy					0.311		
Grade 3 math achievement						0.283	
Grade 5 math achievement							0.303

Notes: Each column reports partial correlations from a regression of the variable named in the header on the row variables in the first column. The partial correlations are measures of effect sizes. As predicted, they indicate sizable reciprocal effects of self-efficacy on math achievement, as well as effects of math achievement on self-efficacy, relative to the effects of race.

Table A.2 Partial Correlations Between Reading Achievement and Reading Self-Efficacy, Controlling for Race and Sex

	Grade 3 Reading Achievement	*Grade 5 Reading Achievement*	*Grade 5 Reading Achievement*	*Grade 8 Reading Achievement*	*Grade 8 Reading Achievement*	*Grade 5 Reading Self-Efficacy*	*Grade 8 Reading Self-Efficacy*
American Indian/Alaska Native	–0.100	–0.084	–0.082	–0.068	–0.070	–0.005	–0.051
Pacific Islander	–0.008	–0.001	–0.001	–0.120	–0.009	–0.026	–0.044
black	–0.194	–0.235	–0.233	–0.254	–0.249	0.026	–0.011
Hispanic	–0.185	–0.158	–0.157	–0.155	–0.157	–0.088	–0.085
Asian	0.051	0.057	0.057	0.071	0.071	0.006	0.021
sex	0.098	0.091	0.093	0.120	0.119	0.089	0.143
Grade 3 self-efficacy	0.235	0.209					
Grade 5 self-efficacy			0.316	0.279			
Grade 8 self-efficacy					0.279		
Grade 3 reading achievement						0.314	
Grade 5 reading achievement							0.280

Notes: Each column reports partial correlations from a regression of the variable named in the header on the row variables in the first column. The partial correlations are measures of effect sizes. As predicted, they indicate sizable reciprocal effects of self-efficacy on reading achievement, as well as effects of reading achievement on self-efficacy, relative to the effects of race.

Appendix B

Table B.1 Partial Correlations Between GPA and Engagement, and Between GPA and Self-Efficacy, Controlling for Race and Sex

	Grade 9 Math GPA	*Grade 9 English GPA*	*Grade 11 Math GPA*	*Grade 11 English GPA*	*Grade 12 Math GPA*	*Grade 12 English GPA*	*Grade 10 Math Self-Efficacy*	*Grade 10 English Self-Efficacy*
sex	0.066	0.159	0.073	0.196	0.079	0.217	–0.086	0.122
Asian	0.062	0.064	0.030	0.073	0.021	0.060	0.011	0.022
Hispanic	–0.082	–0.071	–0.080	–0.098	–0.080	–0.109	–0.041	–0.039
black	–0.113	–0.153	–0.106	–0.132	–0.118	–0.120	0.081	0.030
American Indian/Alaska	–0.067	–0.059	–0.028	–0.059	–0.028	–0.032	0.015	–0.010
Grade 8 engagement	0.062	0.060						
Grade 10 self-efficacy			0.302	0.234	0.245	0.217		
Grade 9 math GPA							0.305	
Grade 9 English GPA								0.291

Notes: Each column reports partial correlations from a regression of the variable named in the header on the row variables in the first column. The partial correlations are measures of effect sizes. As predicted, they indicate sizable reciprocal effects of engagement and self-efficacy on math and English achievement, as well as effects of math and English achievement on self-efficacy, relative to the effects of race.

Appendix C

Models of achievement typically include controls for race. This controls for the possibility that the effects of other covariates depend on race. For example, if black and Hispanic children are born with levels of self-efficacy that are depressed relative to the self-efficacy of white children, this might explain why black and Hispanic children exhibit depressed levels of self-efficacy and depressed levels of achievement. Alternatively, if black and Hispanic families raise their children in ways that cause self-efficacy to be depressed, this might explain why black and Hispanic children exhibit depressed levels of self-efficacy and depressed levels of achievement.

However, it appears that black and Hispanic children enter the school system with higher—not lower—levels of self-efficacy, compared to white children, but their levels of self-efficacy are depressed and fall below the level of white children after entry into the school system (see Figures 4.5 and 4.6). This contradicts the hypothesis that black and Hispanic children are born with levels of self-efficacy that are depressed relative to the self-efficacy of white children. In addition, this contradicts the hypothesis that black and Hispanic families raise their children in ways that cause self-efficacy to be depressed.

Similarly, models of achievement typically include controls for level of income or socioeconomic status. This controls for the possibility that the effects of other covariates depend on the level of income or socioeconomic status. For example, if low-income children (family income below the poverty level) are born with levels of self-efficacy that are depressed relative to the self-efficacy of middle-income children (family income above the poverty level), this might explain why low-income children exhibit depressed levels of self-efficacy and depressed levels of achievement. Similarly, if low-SES children (who fall in the bottom quintile of the SES distribution) are born with levels of self-efficacy that are depressed relative to the self-efficacy of high-SES children (who fall in the top four quintiles of the SES distribution), this might explain why low-SES children exhibit depressed levels of self-efficacy and depressed levels of achievement. Alternatively, if low-income families raise their children in ways that cause self-efficacy to be depressed, this might explain why low-income children exhibit depressed levels of self-efficacy and depressed levels of achievement. Similarly, if low-SES families raise their children in ways that cause self-efficacy to be depressed, this might explain why low-SES children exhibit depressed levels of self-efficacy and depressed levels of achievement.

However, it appears that low-income children enter the school system with higher—not lower—levels of self-efficacy, compared to middle-income children, but their levels of self-efficacy are depressed and fall below the level of middle-income children after entry into the school system (see Figure 4.10). Similarly, low-SES children enter the school system with higher—not lower—levels of self-efficacy, compared to high-SES children, but their levels of self-efficacy are depressed and fall below the level of high-SES children after entry into the school system (see Figure 4.11).

This contradicts the hypothesis that low-income (or low-SES) children are born with levels of self-efficacy that are depressed relative to the self-efficacy of middle-income (or high-SES) children. In addition, this contradicts the hypothesis that low-income (or low-SES) families raise their children in ways that cause self-efficacy to be depressed.

Appendix D

When students are nested within schools, linear mixed modeling (including hierarchical linear modeling or HLM) may be employed to analyze the effects of school-level factors on the coefficients of student-level relationships. When outcomes are significantly correlated within level-two units, linear mixed modeling (LMM) is generally preferred. LMM corrects the standard errors of the prediction parameters and partitions the variance in outcomes into within- and between-school components.

However, the software used to implement LMM and HLM typically employs an empirical Bayes estimation strategy because it results in a smaller mean square error.[1] Parameter estimates are "shrunk" toward their estimated conditional mean vectors.[2] Estimates with the least precision experience the most shrinkage.[3] Low precision and high shrinkage would occur whenever analyses are conducted with a truncated sample, for example, minority subgroups experiencing low-frequency outcomes such as electing a rigorous curriculum or electing calculus. In practice, the shrinkage estimator is often biased, especially in small sample situations involving a logit link and Bernoulli sampling model.[4]

With this in mind, regression equations for eight educational outcomes were estimated using HLM. The level-two group variable was each student's high school. Parameter estimates are reported in Table D.1.

Predicted probabilities for each outcome at mean student-level test scores, and one SD above mean student-level test scores, are reported in Table D.2. School test score means when student-level test scores are increased by one SD were predicted from regression of school scores on student-level test scores, socioeconomic status, student grade-point average, sex, race, and geographic region (Table D.3) and were combined with predictions of grades when student-level test scores are increased by one SD (Table 7.1) to estimate the total impact on eight educational outcomes when student-level test scores are raised by one SD.

The overall pattern of predicted probabilities reported in the top panel of Table D.2 is similar to the pattern reported in the top panel of Table 7.3: Asian students outperformed white students, who outperformed black and Hispanic students. Relative magnitudes of performance across race were roughly comparable. However, with regard to minority subgroups experiencing low-frequency outcomes such as electing a rigorous curriculum or electing calculus, the HLM estimates produced predicted frequencies that departed more sharply from the actual frequencies exhibited in the raw data, compared to the predicted frequencies produced by the generalized linear model (GLM) estimates. This difference is attributable to the application of empirical Bayes estimation, which shrinks estimates for each level-two group intercept toward the grand mean (which is zero when values are standardized). Shrinkage is expected to be large when analyzing outcomes that are infrequent (because the

Table D.1 Logit Estimates for Selected Outcomes

	Dropped Out	*Attended 4-Year*	*Attained BA*	*Rigorous Curriculum*	*Took Calculus*	*Took Algebra2*	*UGPA > 3.75*	*PhD or Prof'l*
SES		0.64***	0.67***		0.07**	0.21***	–0.16***	0.30***
		(0.04)	(0.03)		(0.02)	(0.03)	(0.04)	(0.02)
school score mean	–0.57***	0.76***		0.14***	0.59***	0.73***	0.18***	0.18***
	(0.06)	(0.06)		(0.04)	(0.05)	(0.06)	(0.05)	(0.03)
centered test		0.59***	0.37***	0.03***	0.28***	0.58***	0.24***	
		(0.03)	(0.03)	(0.01)	(0.02)	(0.03)	(0.03)	
GPA	–0.87***	0.76***	0.92***	0.02*	0.44***	0.56***	0.21***	0.21***
	(0.03)	(0.04)	(0.04)	(0.01)	(0.03)	(0.03)	(0.03)	(0.02)
Asian		0.34***	0.25**	0.09**	0.61***	0.18*	–0.17*	0.33***
		(0.08)	(0.08)	(0.03)	(0.07)	(0.08)	(0.08)	(0.07)
Hispanic		0.30***				0.19**	–0.19*	0.57***
		(0.08)				(0.07)	(0.08)	(0.04)
black	–0.25**	0.75***	0.47***			0.35***	–0.19**	0.57***
	(0.09)	(0.09)	(0.08)			(0.07)	(0.07)	(0.05)
AmIndian/Alaskan				[a]		0.34*		0.74***
						(0.17)		(0.06)
sex		–0.15**	0.10*		–0.20***		0.24***	
		(0.05)	(0.04)		(0.03)		(0.05)	
constant	–1.97***	0.57***	–0.97***	–1.97***	–1.77***	–0.20***	–1.93***	–2.53***
n	9478	9478	9478	9478	9478	9478	9478	9478

Notes: *$p<.05$, **$p<.01$, ***$p<.001$ (two-tailed tests). Robust standard errors in parentheses.

Estimation method: EM Laplace approximation.

Neither R-squared nor pseudo-r-squared is computed with logistic regression using survey data.

School score mean = school-level test score associated with each student-level observation.

Centered test = student-level test score minus school score mean.

[a] No AmIndian/Alaskan student experienced a highly rigorous curriculum; these cases were dropped.

Table D.2 Predicted Probabilities at Mean Test Score and 1 SD Above Mean Test Score, by Race and Sex

	Dropped Out	*Attended 4-Year*	*Attained BA*	*Rigorous Curriculum*	*Took Calculus*	*Took Algebra2*	*UGPA > 3.75*	*PhD or Prof'l*
Predicted percent at mean scores for males (females)								
black	11.4(8.4)	59.6(59.3)	9.3(13.2)	0.00(0.00)	0.1(0.1)	29.3(34.4)	2.3(3.4)	1.3(1.6)
Hispanic	12.8(9.4)	44.0(43.6)	6.0(8.7)	0.00(0.00)	0.1(0.1)	26.5(31.2)	5.8(8.5)	0.6(0.7)
Asian	3.1(2.2)	85.6(85.4)	43.5(53.4)	0.00(0.00)	3.4(2.9)	72.7(77.1)	6.4(9.3)	3.4(3.9)
white	4.3(3.1)	72.2(71.9)	26.8(35.3)	0.00(0.00)	0.8(0.7)	55.6(61.2)	8.5(12.3)	1.4(1.6)
Predicted percent at 1 SD above mean for males (females)								
black	5.3(3.8)	79.1(78.8)	20.3(27.5)	0.00(0.00)	0.6(0.5)	62.7(68.0)	3.5(5.1)	1.9(2.2)
Hispanic	4.7(3.4)	72.5(72.2)	17.3(23.7)	0.00(0.00)	1.4(1.2)	68.6(73.4)	9.5(13.6)	0.9(1.1)
Asian	0.7(0.5)	97.0(97.0)	80.3(85.8)	0.06(0.05)	64.8(60.7)	97.1(97.7)	12.6(17.7)	6.5(7.4)
white	1.6(1.1)	89.1(88.9)	52.8(62.4)	0.01(0.01)	10.6(9.0)	87.3(89.7)	13.3(18.8)	2.1(2.5)

Table D.3 OLS Estimates and Predicted Value of School Score at Mean Student-Level Score and 1 SD Above Mean Student-Level Score, by Race

	Full Model	*Reduced Model*
test score	0.25***	0.25***
	(0.02)	(0.01)
ses	0.13***	0.13***
	(0.02)	(0.02)
GPA	–0.01	
	(0.02)	
sex	–0.01	
	(0.02)	
Asian	0.07	
	(0.05)	
Hispanic	–0.21***	–0.23***
	(0.04)	(0.04)
black	–0.17**	–0.17**
	(0.06)	(0.06)
AmIndian/AK	–0.56**	–0.59**
	(0.20)	(0.21)
northcentral	–0.02	
	(0.04)	
south	–0.19***	–0.16***
	(0.04)	(0.03)
west	–0.08	
	(0.04)	
constant	0.17***	0.14***
n	9418	9428
R^2	0.40	0.40
Predicted school score mean at mean student-level score		
black		–0.21
Hispanic		–0.30
Asian		0.20
white		0.19
Predicted school score mean at 1 SD above mean student-level score		
black		–0.02
Hispanic		–0.06
Asian		0.55
white		0.43

Notes: **p*<.05, ***p*<.01, ****p*<.001 (two-tailed tests).

Linearized standard errors in parentheses.

Predicted school scores are standardized values.

resulting parameter estimates are imprecise). The HLM estimates predicted that no Asian, black, Hispanic, or white student experienced a rigorous curriculum. However, tabulations of frequencies derived from the raw data indicate that small percentages of Asian, black, Hispanic, and white students experienced a rigorous curriculum. The GLM estimates correctly predicted that small percentages of each racial group experienced a rigorous curriculum. In general, across most of the eight outcome measures, frequencies predicted using the GLM estimates were more closely matched to the actual frequencies exhibited in the raw data, compared to frequencies predicted from the HLM estimates.

In addition, it appears that the GLM estimates produced more conservative estimates of the predicted impact of raising test scores by one SD. For almost all outcomes, the predicted impact of raising test scores by one SD was larger when estimated with HLM versus GLM (compare the top and bottom panels of Table D.2 to the top and bottom panels of Table 7.3). The HLM estimates predicted that 9.3 percent of black males attained a BA degree and 20.3 percent of black males who scored one SD above the mean score would attain a BA degree—a 118.3 percent increase in the number of black males who would attain a BA degree. In comparison, the GLM estimates predicted that raising black student test scores from the mean to one SD above the mean would increase the number of black males who attain a BA degree by 88.6 percent. The HLM estimates predicted that 6.0 percent of Hispanic males attained a BA degree and 17.3 percent of Hispanic males who score one SD above the mean score would attain a BA degree—a 188.3 percent increase in the number of Hispanic males who would attain a BA degree. In comparison, the GLM estimates predicted that raising Hispanic student test scores from the mean to one SD above the mean would increase the number of Hispanic males who attain a baccalaureate degree by 125.7 percent.

The HLM estimates imply that a one SD increase in test scores would increase the number of black males who complete calculus in high school by 500 percent. The number of Hispanic males who complete calculus would increase by 1300 percent, and the number of Asian males who complete calculus would increase by 1805.9 percent. In comparison, the GLM estimates predict that raising black student test scores from the mean to one SD above the mean would increase the number of black males who complete calculus in high school by 418.2 percent; the corresponding increase for Hispanic males is 772.7 percent and the increase for Asian males is 852.6 percent.

It might be argued that the HLM estimates should be preferred because they separate effects that are properly attributed to schools, rather than the characteristics of students within schools. However, a one SD increase in test scores that occurs over the K-12 years would have two effects. An increase in test scores would boost the average performance of every school, creating more high-performing schools and permitting more students—including more minority students—to experience the benefits of attending high-performing schools. In addition, students would also benefit from the improvement in educational outcomes that would occur whether students attend a high- or low-performing school. Therefore, it is appropriate to aggregate both effects of test scores on educational outcomes: the effect that occurs when the average performance of every school increases, permitting more students to experience the benefits of high-performing schools, and the effect that occurs regardless of the quality of the high school that a student attended. The aggregate effect is what is reported by the GLM estimates in Table 7.2a and Table 7.2b and the predicted probabilities in Table 7.3.

In principle, if all student-level test scores are increased by one SD, school test score means would simply increase by corresponding amounts. As indicated above, however, school test score means when student-level test scores are increased by one SD were predicted from regression of school scores on student-level test scores, socioeconomic status, student grade-point average, sex, race, and geographic region (Table D.3). The latter approach produces conservative estimates of the impact of student-level test scores on school test score means. The conservative approach reflects the sorting that occurs when high-scoring students gain admission and self-select into high schools with high average test scores.

Sorting occurs in several ways. First, many of the best public high schools only admit students through competitive examinations. This includes Lowell High School in San Francisco; DeBakey High School in Houston; Boston Latin Academy, Boston Latin School, and John D. O'Bryant in Boston; nine selective public high schools in New York including the Bronx High School of Science and the Brooklyn Latin School; and 11 selective-enrollment schools in Chicago. Second, the parents of high-scoring students seek to place their children in the best private and parochial high schools, which only admit students through competitive examinations. These private and parochial schools require applicants to submit scores from the Independent School Entrance Examination, the Secondary School Admission Test, or the High School Placement Test. In New York, students seeking admission to Catholic high schools must submit scores from the Test for Admission into Catholic High Schools. Third, regardless of whether school admission involves a competitive examination, students and their parents are acutely aware of the reputed quality of various schools. Parents of students who achieve at high levels seek to place their children in the best available schools—both public and private. It is not uncommon for children to commute long distances by bus or automobile to attend the best schools. Even after controlling for socioeconomic status and race, a one SD increase in student test score is predicted to increase the quality of the school attended by a student (the regression estimates reported in Table D.3 control for socioeconomic status and race). This occurs because high-scoring students gain admission and self-select into the best high schools. However, it is important to note that the relationship between student-level test scores and school test score means would be accentuated if *all* student-level test scores are raised by one SD because this increase would boost the average performance of *every* high school, creating more high-performing high schools and permitting more students to experience the benefits of attending high-performing high schools *even if no student changed his or her school.*

To summarize, while HLM is often preferred to GLM when students are nested within schools, the purpose of the analysis reported in this chapter is consistent with the use of GLM rather than HLM. The results reported in Tables D.1 and D.2 suggest that raising student test scores by one SD would have a significant positive impact, whether estimated with HLM or GLM. The impact is larger when estimated with HLM (implying that GLM estimates are relatively conservative). Finally, the application of HLM and empirical Bayes estimation when analyzing low-frequency minority group outcomes using a logit link and Bernoulli sampling model appears to introduce a greater degree of bias in the parameter estimates, compared to the application of GLM.

Appendix E

Table E.1 Partial Correlations Between Test Scores and Outcomes, Controlling for High School GPA, SES, Race, Region, and Sex

	High School GPA	*Ever Dropped Out*	*Attended 4-Year College*	*Attained BA Degree*	*Rigorous HS Curriculum*	*Took Algebra 2*	*Took Calculus*	*College GPA > 3.75*	*Aspired to Ph.D. or Prof'l*
test score	.517	–.348	.441	.442	.107	.463	.400	.136	.166
high school GPA		–.404	.325	.366		.376			
socioeconomic status	.297	–.272	.361	.378	.080	.285	.241	.023	.146
sex	.087			.064			–.018	.057	
Asian	.063	–.058	.030		.037		.065		.038
Hispanic			–.102			–.043	–.050	–.048	–.008
Black	–.187		–.055		.014		–.043	–.071	–.005
American Indian/ Alaska Native							–.039		
northcentral			.029	.036	–.034		–.031		
south						.018	.006		.019
west			–.074	–.052			–.012		

Notes: Each column reports a separate set of partial correlations after dropping insignificant covariates. The partial correlations are measures of effect sizes. They indicate sizable effects of test scores on the outcome variables, relative to the effects of socioeconomic status and race

Notes

Chapter 1

[1] John U. Ogbu, "Cultural Problems in Minority Education: Their Interpretations and Consequences--Part One: Theoretical Background," *The Urban Review* 27, no. 3 (1995); John U. Ogbu, "Cultural Problems in Minority Education: Their Interpretations and Consequences--Part One: Case Studies," *The Urban Review* 27, no. 4 (1995); Signithia Fordham and John Ogbu, "Black Students' School Success: Coping with the 'Burden of "Acting White"'," *Urban Review* 18(1986).

[2] James W. Ainsworth-Darnell and Douglas B. Downey, "Assessing the Oppositional Culture Explanation for Racial/Ethnic Differences in School Performance," *American Sociological Review* 63, no. 4 (1998); Douglas B. Downey and James W. Ainsworth-Darnell, "The Search for Oppositional Culture among Black Students," *American Sociological Review* 67, no. 1 (2002); Phillip J. Cook and Jens Ludwig, "The 'Burden of "Acting White"': Do Black Adolescents Disparage Academic Achievement?," in *The Black-White Test Score Gap*, ed. C. Jencks and M. Phillips (Washington, DC: The Brookings Institution, 1998).

[3] George Farkas, Christy Lleras, and Steve Maczuga, "Does Oppositional Culture Exist in Minority and Poverty Peer Groups?," *American Sociological Review* 67, no. 1 (2002).

[4] Roland G. Fryer, Jr. and Steven D. Levitt, "Understanding the Black-White Test Score Gap in the First Two Years of School," *The Review of Economics and Statistics* 86, no. 2 (2004): 457-458.

[5] Ibid., 458.

[6] See Roland G. Fryer, Jr. and Steven D. Levitt, "The Black-White Test Score Gap through Third Grade," *American Law and Economics Review* 8, no. 2 (2006): 250.

[7] Fryer and Levitt, "Understanding the Black-White Test Score Gap in the First Two Years of School", 458.

S.S. Yeh, *Solving the Achievement Gap*,
DOI 10.1057/978-1-137-58767-1

[8] See Fryer and Levitt, "The Black-White Test Score Gap through Third Grade", 271.

[9] See ibid.

[10] See ibid., 251.

[11] See ibid.

[12] Fryer and Levitt, "Understanding the Black-White Test Score Gap in the First Two Years of School", 456.

Chapter 2

[1] See Sean Reardon, "The Widening Academic Achievement Gap between the Rich and the Poor: New Evidence and Possible Explanations," in *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances*, ed. Greg J. Duncan and Richard J. Murnane (New York: Russell Sage Foundation, 2011).

[2] See ibid.

[3] Betty Hart and Todd R. Risley, *Meaningful Differences in the Everyday Experience of Young American Children* (Baltimore, MD: Paul H. Brookes Publishing Co., 1995).

[4] Ibid.

[5] Ibid.

[6] R. Allington et al., "Access to Books: Variations in Schools and Classrooms," *The Language and Literacy Spectrum* 5(1995); S. Neuman and D. Celano, "Access to Print in Low-Income and Middle-Income Communities," *Reading Research Quarterly* 36, no. 1 (2001); C. Di Loreto and L. Tse, "Seeing Is Believing: Disparity in Books in Two Los Angeles Area Public Libraries," *School Library Quarterly* 17, no. 3 (1999); C. Smith, R. Constantino, and S. Krashen, "Differences in Print Environment for Children in Beverly Hills, Compton, and Watts," *Emergency Librarian* 24, no. 2 (1996).

[7] A. B. Anderson and S. J. Stokes, "Social and Institutional Influences on the Development and Practice of Literacy," in *Awakening to Literacy*, ed. H. Goelman, A. Oberg, and F. Smith (Portsmouth, NH: Heinemann, 1984); D. Feitelson and Z. Goldstein, "Patterns of Book Ownership and Reading to Young Children in Israeli School-Oriented and Nonschool-Oriented Families," *The Reading Teacher* 39(1986); M. M. Harris and N. J. Smith, "Literacy Assessment of Chapter 1 and Non-Chapter 1 Homes," *Reading Improvement* 24(1987); S. B. Heath, "What No Bedtime Story Means: Narrative Skills at Home and School," *Language in Society* 11(1982); S. B. Heath, *Ways with Words: Language, Life, and Work in Communities and Classrooms* (Cambridge: Cambridge University Press, 1983); J. Mason and C. McCormick, "An Investigation of Prereading Instruction from a Developmental Perspective: Foundations for Literacy," Technical Report No. 224, (Urbana-Champaign, IL: Center for the Study of Reading, 1981); W. H. Miller, "Home Prereading Experiences and First-Grade Achievement," *The Reading Teacher* 22(1969); A. Ninio, "Picture-Book Reading in Mother-Infant Dyads Belonging to Two Subgroups in Israel," *Child Development* 51(1980); I. S. Raz and P. Bryant,

"Social Background, Phonological Awareness and Children's Reading," *British Journal of Developmental Psychology* 8(1990); J. Swinson, "A Parental Involvement Project in a Nursery School," *Educational Psychology in Practice* 1(1985); G. Wells, "Preschool Literacy-Related Activities and Success in School," in *Literacy, Language, and Learning*, ed. D. R. Olson, N. Torrance, and A. Hildyard (Cambridge: Cambridge University Press, 1985); A. G. Bus, M. H. van IJzendoorn, and A. D. Pellegrini, "Joint Book Reading Makes for Success in Learning to Read: A Meta-Analysis on Intergenerational Transmission of Literacy," *Review of Educational Research* 65, no. 1 (1995); Hollis S. Scarborough and Wanda Dobrich, "On the Efficacy of Reading to Preschoolers," *Developmental Review* 14(1994).

[8] Erika Hoff-Ginsberg, "Mother-Child Conversation in Different Social Classes and Communicative Settings," *Child Development* 62, no. 4 (1991).

[9] Susan Sidney Smith and Rhonda G. Dixon, "Literacy Concepts of Low- and Middle-Class Four-Year-Olds Entering Preschool," *The Journal of Educational Research* 88, no. 4 (1995).

[10] Jim Lindsay, "Children's Access to Print Material and Education-Related Outcomes: Findings from a Meta-Analytic Review," (Napierville, IL: Learning Point Associates, 2010), 1.

[11] See, for example, Richard Rothstein, "Class and the Classroom," *American School Board Journal* 191, no. 10 (2004), http://www.asbj.com/MainMenuCategory/Archive/2004/October/Class-and-the-Classroom.aspx; James J. Heckman, *Giving Kids a Fair Chance* (Cambridge, MA: The MIT Press, 2013).

[12] Heckman, *Giving Kids a Fair Chance*, 12-13.

[13] Tim Kautz et al., "Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success," (Paris: OECD, 2014), 25.

[14] Mathilde Almlund et al., "Personality Psychology and Economics," in *Handbook of the Economics of Education*, ed. E. A. Hanushek, S. J. Machin, and L. Woessmann (Amsterdam: Elsevier, 2011), 94.

[15] Kautz et al., "Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success", 23.

[16] Almlund et al., "Personality Psychology and Economics," 90.

[17] Ibid., 98.

[18] Heckman, *Giving Kids a Fair Chance*, 12.

[19] Ibid.

[20] James J. Heckman and Tim D. Kautz, "Hard Evidence on Soft Skills," *Labour Economics* 19, no. 4 (2012): 462.

[21] Ibid., 460.

[22] Samuel Bowles, "A (Science-Based) Poor Kids' Manifesto," *Science* 340(2013): 1044-1045.

[23] William Darity, "Stratification Economics: The Role of Intergroup Inequality," *Journal of Economics and Finance* 29, no. 2 (2005): 144.

[24] Ibid.

Chapter 3

[1] J. S. Coleman et al., "Equality of Educational Opportunity," (Washington, DC: Government Printing Office, 1966).

[2] See Sean F. Reardon, Joseph P. Robinson-Cimpian, and Ericka S. Weathers, "Patterns and Trends in Racial/Ethnic and Socioeconomic Academic Achievement Gaps," in *Handbook of Research in Education Finance and Policy*, ed. Helen A. Ladd and Margaret E. Goertz (New York: Lawrence Erlbaum, 2015), 496.

[3] Coleman et al., "Equality of Educational Opportunity", 319, 322.

[4] Ibid.

[5] W. B. Brookover et al., *Schools, Social Systems and Student Achievement: Schools Can Make a Difference* (New York: Praeger, 1979); W. B. Brookover et al., "Elementary School Social Climate and School Achievement," *American Educational Research Journal* 15(1978); V. Crandall, W. Katkovky, and U. Crandall, "Children's Beliefs in Their Own Control of Reinforcements in Intellectual Academic Achievement Situations," *Child Development* 36(1965); A. D. Kalechstein and S. Nowicki, Jr., "A Meta-Analytic Examination of the Relationship between Control Expectancies and Academic Achievement: An 11-Yr Follow-up to Findley and Cooper," *Genetic, Social, and General Psychology Monographs* 123, no. 1 (1997); T. Z. Keith, S. M. Pottebaum, and S. Eberhart, "Effects of Self Concept and Locus of Control on Academic Achievement: A Large Sample Path Analysis," *Journal of Psychoeducational Assessment* 4, no. 1 (1986); E. A. Skinner, J. G. Wellborn, and J. P. Connell, "What It Takes to Do Well in School and Whether I've Got It: A Process Model of Perceived Control and Children's Engagement and Achievement in School," *Journal of Educational Psychology* 82, no. 1 (1990); C. Teddlie and S. Stringfield, *Schools Do Make a Difference: Lessons Learned from a 10 Year Study of School Effects* (New York: Teachers College Press, 1993); E. A. Skinner et al., "Individual Differences and the Development of Perceived Control," *Monographs of the Society for Research in Child Development* 63, no. 2/3 (1998); M. J. Findley and H. M. Cooper, "Locus of Control and Academic Achievement: A Literature Review," *Journal of Personality and Social Psychology* 44, no. 2 (1983); Kalechstein and Nowicki, "A Meta-Analytic Examination of the Relationship between Control Expectancies and Academic Achievement"; Bernhard Schmitz and Ellen Skinner, "Perceived Control, Effort, and Academic Performance: Interindividual, Intraindividual, and Multivariate Time-Series Analyses," *Journal of Personality and Social Psychology* 64, no. 6 (1993); Allan Wigfield and Michele Karpathian, "Who Am I and What Can I Do? Children's Self-Concepts and Motivation in Achievement Situations," *Educational Psychologist* 26, no. 3 and 4 (1991).

[6] D. H. Schunk, "Self-Efficacy and Achievement Behaviors," *Educational Psychology Review* 1(1989); D. H. Schunk, "Self-Efficacy and Academic Motivation," *Educational Psychologist* 26(1991); D. H. Schunk, "Self-Efficacy Perspective on Achievement Behavior," *Educational Psychologist* 19(1984);

D. H. Schunk, "Self-Efficacy and Classroom Learning," *Psychology in the Schools* 22(1985); A. Bandura, "Self-Efficacy Mechanism in Human Agency," *American Psychologist* 37(1982).

[7] T. J. Crooks, "The Impact of Classroom Evaluation Practices on Students," *Review of Educational Research* 58(1988): 461.

[8] Carol Dweck, "Motivational Processes Affecting Learning," *American Psychologist* 41(1986): 1040–1041.

[9] Ibid.

[10] C. E. Ross and B. A. Broh, "The Roles of Self Esteem and the Sense of Personal Control in the Academic Achievement Process," *Sociology of Education* 73, no. 4 (2000); Schmitz and Skinner, "Perceived Control, Effort, and Academic Performance: Interindividual, Intraindividual, and Multivariate Time-Series Analyses"; Skinner, Wellborn, and Connell, "What It Takes to Do Well in School and Whether I've Got It"; Skinner et al., "Individual Differences and the Development of Perceived Control".

[11] Crooks, "The Impact of Classroom Evaluation Practices on Students", 462.

[12] Dweck, "Motivational Processes Affecting Learning", 1041.

[13] Ross and Broh, "The Roles of Self Esteem and the Sense of Personal Control in the Academic Achievement Process"; Schmitz and Skinner, "Perceived Control, Effort, and Academic Performance: Interindividual, Intraindividual, and Multivariate Time-Series Analyses"; Skinner, Wellborn, and Connell, "What It Takes to Do Well in School and Whether I've Got It"; Skinner et al., "Individual Differences and the Development of Perceived Control".

[14] Crooks, "The Impact of Classroom Evaluation Practices on Students", 462.

[15] Daniel C. Molden and Carol S. Dweck, "Meaning and Motivation," in *Intrinsic and Extrinsic Motivation: The Search for Optimal Motivation and Performance*, ed. Carol Sansone and Judith M. Harackiewicz (San Diego: Academic Press, 2000), 152.

[16] Hart and Risley, *Meaningful Differences in the Everyday Experience of Young American Children.*

[17] See Richard J. Murnane et al., "Understanding Trends in the Black-White Achievement Gaps During the First Years of School," *Brookings-Wharton Papers on Urban Affairs* (2006): 122.

[18] See ibid.

[19] See Fryer and Levitt, "The Black-White Test Score Gap through Third Grade", 273.

[20] See ibid., 252, 278.

[21] Christopher Jencks and Meredith Phillips, eds., *The Black-White Test Score Gap* (Washington, DC: Brookings, 1998), 27.

[22] For evidence of downward spirals, see Laura LoGerfo, Austin Nichols, and Sean Reardon, "Achievement Gains in Elementary and High School," (Washington, DC: Urban Institute, 2006), 26–32, 62–65.

[23] See D. R. Entwisle and K. Alexander, “Summer Setback: Race, Poverty, School Composition, and Mathematics Achievement in the First Two Years of School,” *American Sociological Review* 57(1992); D. R. Entwisle and K. Alexander, “Winter Setback: The Racial Composition of Schools and Learning to Read,” *American Sociological Review* 59(1994); B. Heyns, *Summer Learning and the Effects of Schooling* (New York: Academic Press, 1978).

[24] Fryer and Levitt, “Understanding the Black-White Test Score Gap in the First Two Years of School”, 459.

[25] Ogbu, “Cultural Problems in Minority Education: Their Interpretations and Consequences--Part One: Theoretical Background”; Ogbu, “Cultural Problems in Minority Education: Their Interpretations and Consequences--Part One: Case Studies”; Fordham and Ogbu, “Black Students’ School Success: Coping with the ‘Burden of “Acting White”’”.

[26] Ainsworth-Darnell and Downey, “Assessing the Oppositional Culture Explanation for Racial/Ethnic Differences in School Performance”; Downey and Ainsworth-Darnell, “The Search for Oppositional Culture among Black Students”; Cook and Ludwig, “The ‘Burden of “Acting White”’: Do Black Adolescents Disparage Academic Achievement?.”; Phillip J. Cook and Jens Ludwig, “Weighing the ‘Burden of “Acting White”’: Are There Race Differences in Attitudes toward Education?,” *Journal of Policy Analysis and Management* 16, no. 2 (1997).

[27] Farkas, Lleras, and Maczuga, “Does Oppositional Culture Exist in Minority and Poverty Peer Groups?”.

[28] National Bureau of Economic Research, “Moving to Opportunity (MTO) for Fair Housing Demonstration Program,” http://www.nber.org/mtopublic/.

[29] Ibid.

[30] R. Baron, D. Tom, and H. Cooper, “Social Class, Race, and Teacher Expectations,” in *Teacher Expectancies*, ed. Jerome Dusek (Hillsdale, NJ: Lawrence Erlbaum, 1985); J. B. Dusek, “Do Teachers Bias Children’s Learning?,” *Review of Educational Research* 45(1975); R. F. Ferguson, “Teachers’ Perceptions and Expectations and the Black-White Test Score Gap,” in *The Black-White Test Score Gap*, ed. C. Jencks and M. Phillips (Washington, DC: Brookings Institute, 1998); S. Lightfoot, *Worlds Apart: Relationships between Families and Schools* (New York: Basic Books, 1978).

[31] Fryer and Levitt, “Understanding the Black-White Test Score Gap in the First Two Years of School”, 459–460.

[32] Ibid., 456.

[33] Ibid., 457–458.

[34] Ibid., 458.

[35] See Fryer and Levitt, “The Black-White Test Score Gap through Third Grade”, 250.

[36] Fryer and Levitt, “Understanding the Black-White Test Score Gap in the First Two Years of School”, 458.

[37] See Fryer and Levitt, “The Black-White Test Score Gap through Third Grade”, 271.

[38] See ibid.

[39] See ibid., 251.

[40] See ibid.

[41] Fryer and Levitt, "Understanding the Black-White Test Score Gap in the First Two Years of School", 456.

[42] H.M. Berston, "The School Dropout Problem," *The Clearing House* 35, no. 4 (1960); Ellen Brantlinger, "Low-Income Adolescents' Perceptions of School, Intelligence, and Themselves as Students," *Curriculum Inquiry* 20, no. 3 (1990); William M. Reynolds and Kim L. Miller, "Assessment of Adolescents' Learned Helplessness in Achievement Situations," *Journal of Personality Assessment* 53, no. 2 (1989); C. S. Dweck and N. Dickon Reppucci, "Learned Helplessness and Reinforcement Responsibility in Children," *Journal of Personality and Social Psychology* 25, no. 1 (1973): 109.

[43] Martin E. P. Seligman and Steven F. Maier, "Failure to Escape Traumatic Shock," *Journal of Experimental Psychology* 74, no. 1 (1967).

[44] D.S. Hiroto and M.E.P. Seligman, "Generality of Learned Helplessness in Man," *Journal of Personality and Social Psychology* 31(1975); D.C. Klein, E. Fencil-Morse, and M.E.P. Seligman, "Learned Helplessness, Depression, and the Attribution of Failure," *Journal of Personality and Social Psychology* 33(1976); C. Peterson, S. F. Maier, and M. E. P. Seligman, *Learned Helplessness: A Theory for the Age of Personal Control* (New York: Oxford University Press, 1995); Martin E. P. Seligman, *Helplessness: On Depression, Development, and Death* (New York: WH Freeman, 1992); Dweck and Reppucci, "Learned Helplessness and Reinforcement Responsibility in Children", 109.

[45] Seligman, *Helplessness: On Depression, Development, and Death.*

[46] Berston, "The School Dropout Problem"; Brantlinger, "Low-Income Adolescents' Perceptions of School, Intelligence, and Themselves as Students".

[47] L. Kanevsky and T. Keighley, "To Produce or Not to Produce? Understanding Boredom and the Honor in Underachievement," *Roeper Review* 26, no. 1 (2003).

[48] Berston, "The School Dropout Problem"; Brantlinger, "Low-Income Adolescents' Perceptions of School, Intelligence, and Themselves as Students"; Kanevsky and Keighley, "To Produce or Not to Produce? Understanding Boredom and the Honor in Underachievement".

[49] The achievement gap between black and white children grows from entry into kindergarten through the end of high school. See LoGerfo, Nichols, and Reardon, "Achievement Gains in Elementary and High School".

[50] Berston, "The School Dropout Problem"; Brantlinger, "Low-Income Adolescents' Perceptions of School, Intelligence, and Themselves as Students"; Reynolds and Miller, "Assessment of Adolescents' Learned Helplessness in Achievement Situations".

[51] Gary C. Wehlage and Robert A. Rutter, "Dropping Out: How Much Do Schools Contribute to the Problem?," *Teachers College Record* 87, no. 3 (1986): 376.

[52] Michel Janosz et al., "Disentangling the Weight of School Dropout Predictors: A Test on Two Longitudinal Samples," *Journal of Youth and Adolescence* 26, no. 6 (1997).

[53]J. S. Eccles, A. Wigfield, and U. Schiefele, "Motivation to Succeed," in *Handbook of Child Psychology* ed. N. Eisenberg and W. Damon (New York: John Wiley and Sons, 1998); J. S. Eccles and C. Midgley, "Stage-Environment Fit: Developmentally Appropriate Classrooms for Early Adolescents," in *Research on Motivation in Education, Vol. 3: Goals and Cognitions*, ed. R. E. Ames and C. Ames (New York: Academic, 1989).

[54]Eccles, Wigfield, and Schiefele, "Motivation to Succeed."; Eccles and Midgley, "Stage-Environment Fit: Developmentally Appropriate Classrooms for Early Adolescents."

[55]Paul A. McDermott, Melissa Mordell, and Jill C. Stoltzfus, "The Organization of Student Performance in American Schools: Discipline, Motivation, Verbal Learning, and Nonverbal Learning," *Journal of Educational Psychology* 93, no. 1 (2001); Helen M. Marks, "Student Engagement in Instructional Activity: Patterns in the Elementary, Middle, and High School Years," *American Educational Research Journal* 37, no. 1 (2000).

[56]Gallup, "Gallup Student Poll," http://www.gallup.com/services/180029/gallup-student-poll-2014-overall-report.aspx.

[57]Adena M. Klem and James P. Connell, "Relationships Matter: Linking Teacher Support to Student Engagement and Achievement," *Journal of School Health* 74, no. 7 (2004).

[58]E. M. Anderman and M. L. Maehr, "Motivation and Schooling in the Middle Grades," *Review of Educational Research* 64(1994); Eccles and Midgley, "Stage-Environment Fit: Developmentally Appropriate Classrooms for Early Adolescents."; J. S. Eccles et al., "Development During Adolescence: The Impact of Stage-Environment Fit on Adolescents' Experiences in Schools and Families," *American Psychologist* 48(1993); Eccles, Wigfield, and Schiefele, "Motivation to Succeed."; S. Harter, "Pleasure Derived from Optimal Challenge and the Effects of Extrinsic Rewards on Children's Difficulty Level Choices," *Child Development* 49(1978); S. Harter, "A New Self-Report Scale of Intrinsic Versus Extrinsic Orientation in the Classroom: Motivational and Informational Components," *Developmental Psychology* 17(1981); S. Harter, "A Model of Mastery Motivation in Children: Individual Differences and Developmental Change," in *The Minnesota Symposia on Child Psychology (Vol. 14)*, ed. W. A. Collins (Hillsdale, NJ: Erlbaum, 1981).

[59]Crooks, "The Impact of Classroom Evaluation Practices on Students", 464.

[60]Ross and Broh, "The Roles of Self Esteem and the Sense of Personal Control in the Academic Achievement Process"; Findley and Cooper, "Locus of Control and Academic Achievement"; Kalechstein and Nowicki, "A Meta-Analytic Examination of the Relationship between Control Expectancies and Academic Achievement"; Brookover et al., "Elementary School Social Climate and School Achievement"; S. Hymel, "The Relationship between Motivation Factors and Report Card Evaluation" (paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, CA, April, 1981); Kennedy T. Hill, "The Relation of Evaluative Practices to Test Anxiety

and Achievement Motivation," *UCLA Educator* 19(1977); Kennedy T. Hill and Allan Wigfield, "Test Anxiety: A Major Educational Problem and What Can Be Done About It," *Elementary School Journal* 85, no. 1 (1984); L. J. Fyans, "Test Anxiety, Test Comfort, and Student Achievement Test Performance," (Princeton, NJ: Educational Testing Service, 1979); Kennedy T. Hill, "Eliminating Motivational Testing Error by Developing Optimal Testing Procedures and Teaching Test-Taking Skills," (Princeton, NJ: Educational Testing Service, 1979); Crooks, "The Impact of Classroom Evaluation Practices on Students", 450, 468; C. S. Dweck and E. S. Elliott, "Achievement Motivation," in *Handbook of Child Psychology, Vol. 4: Socialization, Personality, and Social Development*, ed. E. M. Hetherington and P. H. Mussen (New York: Wiley, 1983); Kennedy T. Hill and Seymour B. Sarason, "The Relation of Test Anxiety and Defensiveness to Test and School Performance over the Elementary School Years: A Further Longitudinal Study," *Monographs of the Society for Research in Child Development* 31, no. 2 (1966).

[61] Crooks, "The Impact of Classroom Evaluation Practices on Students", 468.

[62] Ibid., 450.

[63] Dweck, "Motivational Processes Affecting Learning", 1040–1041; Dweck and Elliott, "Achievement Motivation."; Skinner et al., "Individual Differences and the Development of Perceived Control", 10–11.

[64] Skinner et al., "Individual Differences and the Development of Perceived Control", 11.

[65] Dweck, "Motivational Processes Affecting Learning", 1041.

[66] C. Diener and C. Dweck, "Analysis of Learned Helplessness: Continuous Changes in Performance, Strategy, and Achievement Cognitions Following Failure," *Journal of Personality and Social Psychology* 36(1978): 460.

[67] Hill and Sarason, "The Relation of Test Anxiety and Defensiveness to Test and School Performance over the Elementary School Years: A Further Longitudinal Study"; Hill and Wigfield, "Test Anxiety: A Major Educational Problem and What Can Be Done About It".

[68] Hill and Sarason, "The Relation of Test Anxiety and Defensiveness to Test and School Performance over the Elementary School Years: A Further Longitudinal Study"; Hill and Wigfield, "Test Anxiety: A Major Educational Problem and What Can Be Done About It"; Fyans, "Test Anxiety, Test Comfort, and Student Achievement Test Performance"; Hill, "Eliminating Motivational Testing Error by Developing Optimal Testing Procedures and Teaching Test-Taking Skills".

[69] Skinner et al., "Individual Differences and the Development of Perceived Control", v.

[70] Ibid.

[71] Ibid., 11, 146.

[72] Ibid., 11.

[73] Ibid.

[74] Ibid.

[75] Ibid.

[76] Ibid.

[77] Ibid.

[78] Ibid., 146.

[79] Ibid., 154.

[80] Ibid., 139.

[81] Skinner, Wellborn, and Connell, "What It Takes to Do Well in School and Whether I've Got It"; Skinner et al., "Individual Differences and the Development of Perceived Control", 146, 147, 149, 150.

[82] Ross and Broh, "The Roles of Self Esteem and the Sense of Personal Control in the Academic Achievement Process".

[83] Brookover et al., "Elementary School Social Climate and School Achievement".

[84] Findley and Cooper, "Locus of Control and Academic Achievement"; Kalechstein and Nowicki, "A Meta-Analytic Examination of the Relationship between Control Expectancies and Academic Achievement".

[85] Ross and Broh, "The Roles of Self Esteem and the Sense of Personal Control in the Academic Achievement Process"; Findley and Cooper, "Locus of Control and Academic Achievement"; Kalechstein and Nowicki, "A Meta-Analytic Examination of the Relationship between Control Expectancies and Academic Achievement"; Brookover et al., "Elementary School Social Climate and School Achievement"; Skinner, Wellborn, and Connell, "What It Takes to Do Well in School and Whether I've Got It"; Skinner et al., "Individual Differences and the Development of Perceived Control".

[86] Skinner et al., "Individual Differences and the Development of Perceived Control", 176.

[87] Ibid.

[88] Ross and Broh, "The Roles of Self Esteem and the Sense of Personal Control in the Academic Achievement Process"; Findley and Cooper, "Locus of Control and Academic Achievement"; Kalechstein and Nowicki, "A Meta-Analytic Examination of the Relationship between Control Expectancies and Academic Achievement"; Brookover et al., "Elementary School Social Climate and School Achievement"; Skinner, Wellborn, and Connell, "What It Takes to Do Well in School and Whether I've Got It"; Skinner et al., "Individual Differences and the Development of Perceived Control".

[89] Ross and Broh, "The Roles of Self Esteem and the Sense of Personal Control in the Academic Achievement Process"; Findley and Cooper, "Locus of Control and Academic Achievement"; Kalechstein and Nowicki, "A Meta-Analytic Examination of the Relationship between Control Expectancies and Academic Achievement"; Brookover et al., "Elementary School Social Climate and School Achievement"; Skinner, Wellborn, and Connell, "What It Takes to Do Well in School and Whether I've Got It"; Skinner et al., "Individual Differences and the Development of Perceived Control". The achievement gap between black and white children grows from entry into kindergarten through the end of high school. See LoGerfo, Nichols, and Reardon, "Achievement Gains in Elementary and High School".

Chapter 4

[1] National Center for Education Statistics, "Early Childhood Longitudinal Program (ECLS): Kindergarten Class of 1998-99 (ECLS-K)," Institute of Education Sciences, http://nces.ed.gov/ecls/kindergarten.asp.

[2] See Reardon, "Whither Opportunity?," 97.

[3] See ibid., 98.

[4] See ibid.

[5] See ibid., 104.

[6] See ibid., 101-104.

[7] See ibid., 102-103.

[8] See ibid., 108.

[9] See William Deresiewicz, *Excellent Sheep: The Miseducation of the American Elite and the Way to a Meaningful Life* (New York: Free Press, 2014); Madeline Levine, *The Price of Privilege: How Parental Pressure and Material Advantage Are Creating a Generation of Disconnected and Unhappy Kids* (New York: HarperCollins, 2006). Pressure to earn high report card grades and test scores is so intense that many students from middle-to-high income families reportedly spend every available minute studying and suffer high rates of anxiety, depression, and somatic complaints. Psychiatric counselors report cases of students who become severely depressed, even suicidal, when SAT test scores are lower than expected. See ibid. Anxiety is highest among students with low report card grades and low achievement test scores. Test anxiety steadily increases across the elementary school years and is highest among junior high school and high school students. Test anxiety affects children from all major sociocultural groups, including black, white, and Hispanic students. See Hill and Wigfield, "Test Anxiety: A Major Educational Problem and What Can Be Done About It."

[10] See Reardon, "Whither Opportunity?," 102-103.

[11] See Fryer and Levitt, "Understanding the Black-White Test Score Gap in the First Two Years of School", 447.

[12] Ibid., 448.

[13] See Fryer and Levitt, "The Black-White Test Score Gap through Third Grade", 258, 261.

[14] See Murnane et al., "Understanding Trends in the Black-White Achievement Gaps During the First Years of School", 121.

[15] See Fryer and Levitt, "The Black-White Test Score Gap through Third Grade", 258-261.

[16] See Reardon, Robinson-Cimpian, and Weathers, "Patterns and Trends in Racial/Ethnic and Socioeconomic Academic Achievement Gaps," 507. The achievement gap between black and white children grows from entry into kindergarten through the end of high school. See LoGerfo, Nichols, and Reardon, "Achievement Gains in Elementary and High School".

[17] See Fryer and Levitt, "The Black-White Test Score Gap through Third Grade", 258-261.

[18] The achievement gap between black and white children grows from entry into kindergarten through the end of high school. See LoGerfo, Nichols, and Reardon, "Achievement Gains in Elementary and High School".

[19] Alex J. Bowers, "Grades and Graduation: A Longitudinal Risk Perspective to Identify Student Dropouts," *The Journal of Educational Research* 103, no. 3 (2010).

[20] Ibid.

[21] William G. Bowen, Martin A. Kurzweil, and Eugene M. Tobin, *Equity and Excellence in American Higher Education* (Charlottesville, VA: University of Virginia Press, 2005); William G. Bowen and Derek Bok, *The Shape of the River: Long-Term Consequences of Considering Race in College and University Admissions* (Princeton, NJ: Princeton University Press, 1998); William G. Bowen, Matthew Chingos, and Michael S. McPherson, *Crossing the Finish Line: Completing College at America's Public Universities* (Princeton, NJ: Princeton University Press, 2009).

[22] Bowers, "Grades and Graduation: A Longitudinal Risk Perspective to Identify Student Dropouts".

[23] Ibid.

[24] Ibid.

[25] Ibid.

[26] Ibid.

[27] National Center for Education Statistics, "Education Longitudinal Study of 2002," Institute of Education Sciences, http://nces.ed.gov/surveys/els2002/.

Chapter 5

[1] K. J. Kennelly, D. Dietz, and P. Benson, "Reinforcement Schedules, Effort vs. Ability Attributions, and Persistence," *Psychology in the Schools* 22, no. 4 (1985); V. Blankenship, "Individual Differences in Resultant Achievement Motivation and Latency to and Persistence at an Achievement Task," *Motivation and Emotion* 16, no. 1 (1992); R. Nygard, "Personality, Situation, and Persistence: A Study with Emphasis on Achievement Motivation," (Oslo: Universitetsforlaget, 1977); P.M. Drucker et al., "Relation of Task Difficulty to Persistence," *Perceptual and Motor Skills* 86(1998); F. W. Danner and D. Lonsky, "A Cognitive-Developmental Approach to the Effects of Rewards on Intrinsic Motivation," *Child Development* 52(1981); Harter, "Pleasure Derived from Optimal Challenge and the Effects of Extrinsic Rewards on Children's Difficulty Level Choices".

[2] R. J. Vallerand and G. Reid, "On the Causal Effects of Perceived Competence on Intrinsic Motivation: A Test of Cognitive Evaluation Theory," *Journal of Sport Psychology* 6(1984).

[3] Kennelly, Dietz, and Benson, "Reinforcement Schedules, Effort vs. Ability Attributions, and Persistence".

[4] Ibid.

[5] R. M. Ryan and E. L. Deci, "Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being," *American Psychologist* 55(2000).

[6] J. I. Goodlad, *A Place Called School* (New York: McGraw Hill, 1984).

[7] Kanevsky and Keighley, "To Produce or Not to Produce? Understanding Boredom and the Honor in Underachievement"; John M. Bridgeland, John J. DiIulio, Jr., and Karen B. Morison, "The Silent Epidemic: Perspectives of High School Dropouts," (Washington, DC: Civic Enterprises, 2006); Ethan Yazzie-Mintz, "Charting the Path from Engagement to Achievement: A Report on the 2009 High School Survey of Student Engagement," (Bloomington, IN: Center for Evaluation and Education Policy, 2010); Berston, "The School Dropout Problem"; Brantlinger, "Low-Income Adolescents' Perceptions of School, Intelligence, and Themselves as Students"; Reynolds and Miller, "Assessment of Adolescents' Learned Helplessness in Achievement Situations".

[8] D. Mac Iver, "Classroom Environments and the Stratification of Pupils' Ability Perceptions," *Journal of Educational Psychology* 80, no. 4 (1988); Douglas J. Mac Iver and David A. Reuman, "Giving Their Best: Grading and Recognition Practices That Motivate Students to Work Hard," *American Educator* 17, no. 4 (1993/1994); D. J. Mac Iver, D. A. Reuman, and S. R. Main, "Social Structuring of the School: Studying What Is, Illuminating What Could Be," *Annual Review of Psychology* 46(1995).

[9] Mac Iver, "Classroom Environments and the Stratification of Pupils' Ability Perceptions"; Mac Iver and Reuman, "Giving Their Best: Grading and Recognition Practices That Motivate Students to Work Hard"; Mac Iver, Reuman, and Main, "Social Structuring of the School: Studying What Is, Illuminating What Could Be".

[10] D. J. Mac Iver, D. J. Stipek, and D. H. Daniels, "Explaining within-Semester Changes in Student Effort in Junior High School and Senior High School Courses," *Journal of Educational Psychology* 83, no. 2 (1991).

[11] R. M. Ryan, V. Mims, and R. Koestner, "The Relationship of Reward Contingency and Interpersonal Context to Intrinsic Motivation: A Review and Test Using Cognitive Evaluation Theory," *Journal of Personality and Social Psychology* 45(1983).

[12] Donald E. P. Smith, Dale Brethower, and Raymond Cabot, "Increasing Task Behavior in a Language Arts Program by Providing Reinforcement," *Journal of Experimental Child Psychology* 8, no. 1 (1969).

[13] S. L. Robinson, C. DePascale, and F. C. Roberts, "Computer Delivered Feedback in Group Based Instruction: Effects for Learning Disabled Students in Mathematics," *Learning Disabilities Focus* 5, no. 1 (1989).

[14] J. Hattie and H. Timperley, "The Power of Feedback," *Review of Educational Research* 77, no. 1 (2007).

[15] R. L. Bangert-Drowns et al., "The Instructional Effect of Feedback in Test-Like Events," *Review of Educational Research* 61, no. 2 (1991); L. S. Fuchs and D. Fuchs, "Effects of Systematic Formative Evaluation: A Meta-Analysis,"

Exceptional Children 53, no. 3 (1986); A. N. Kluger and A. DeNisi, "The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory," *Psychological Bulletin* 119, no. 2 (1996); P. Black and D. Wiliam, "Assessment and Classroom Learning," *Assessment in Education* 5, no. 1 (1998).

[16] Black and Wiliam, "Assessment and Classroom Learning".

[17] Fuchs and Fuchs, "Effects of Systematic Formative Evaluation: A Meta-Analysis".

[18] Kluger and DeNisi, "The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory", 278.

[19] Hattie and Timperley, "The Power of Feedback".

[20] Ibid.

[21] R. S. Lysakowski and H. J. Walberg, "Instructional Effects of Cues, Participation, and Corrective Feedback: A Quantitative Synthesis," *American Educational Research Journal* 19(1982); H. J. Walberg, "What Makes Schooling Effective?," *Contemporary Education Review* 1(1982); G. Tenenbaum and E. Goldring, "A Meta-Analysis of the Effect of Enhanced Instruction: Cues, Participation, Reinforcement and Feedback and Correctives on Motor Skill Learning," *Journal of Research and Development in Education* 22(1989).

[22] Berston, "The School Dropout Problem"; Brantlinger, "Low-Income Adolescents' Perceptions of School, Intelligence, and Themselves as Students"; Reynolds and Miller, "Assessment of Adolescents' Learned Helplessness in Achievement Situations".

[23] *Reading Assessment* and *Math Assessment* are pseudonyms that are used to avoid the appearance that the author endorses the assessment software. The author is neither affiliated with nor has received any funding from the vendor.

[24] J. Ysseldyke and S. Tardrew, "Use of a Progress-Monitoring System to Enable Teachers to Differentiate Math Instruction," *Journal of Applied School Psychology* 24, no. 1 (2007).

[25] S. M. Ross, J. Nunnery, and E. Goldfeder, "A Randomized Experiment on the Effects of Accelerated Reader/Reading Renaissance in an Urban School District: Final Evaluation Report," (Memphis, TN: Center for Research in Educational Policy, The University of Memphis, 2004).

[26] S. S. Yeh, "High Stakes Testing: Can Rapid Assessment Reduce the Pressure?," *Teachers College Record* 108, no. 4 (2006).

[27] Ibid.

[28] J. A. Nunnery, S. M. Ross, and A. McDonald, "A Randomized Experimental Evaluation of the Impact of Accelerated Reader/Reading Renaissance Implementation on Reading Achievement in Grades 3 to 6," *Journal of Education for Students Placed At Risk* 11, no. 1 (2006); Ross, Nunnery, and Goldfeder, "A Randomized Experiment on the Effects of Accelerated Reader/Reading Renaissance in an Urban School District: Final Evaluation Report".

[29] Ross, Nunnery, and Goldfeder, "A Randomized Experiment on the Effects of Accelerated Reader/Reading Renaissance in an Urban School District: Final Evaluation Report".

[30] Nunnery, Ross, and McDonald, "A Randomized Experimental Evaluation of the Impact of Accelerated Reader/Reading Renaissance Implementation on Reading Achievement in Grades 3 to 6".

[31] J. Ysseldyke and D. M. Bolt, "Effect of Technology-Enhanced Continuous Progress Monitoring on Math Achievement," *School Psychology Review* 36, no. 3 (2007).

[32] Ysseldyke and Tardrew, "Use of a Progress-Monitoring System to Enable Teachers to Differentiate Math Instruction".

[33] S. G. Rivkin, E. A. Hanushek, and J. F. Kain, "Teachers, Schools and Academic Achievement," *Econometrica* 73, no. 2 (2005); B. Rowan, R. Correnti, and R.J. Miller, "What Large-Scale Survey Research Tells Us About Teacher Effects on Student Achievement: Insights from the Prospects Study of Elementary Schools," *Teachers College Record* 104(2002); Douglas O. Staiger and Jonah E. Rockoff, "Searching for Effective Teachers with Imperfect Information," *Journal of Economic Perspectives* 24, no. 3 (2010); S.P. Wright, S.P. Horn, and W.L. Sanders, "Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation," *Journal of Personnel Evaluation in Education* 11(1997); W. L. Sanders and J.C. Rivers, "Cumulative and Residual Effects of Teachers on Future Student Academic Achievement," (Knoxville, TN: University of Tennessee Value-Added Research Center, 1996).

[34] D. Aaronson, L. Barrow, and W. Sander, "Teachers and Student Achievement in the Chicago Public High Schools," *Journal of Labor Economics* 25, no. 1 (2007); Dale Ballou, "Value-Added Assessment: Lessons from Tennessee," in *Value Added Models in Education: Theory and Applications*, ed. R. Lissetz (Maple Grove, MN: JAM Press, 2005); Cory Koedel and Julian R. Betts, "Re-Examining the Role of Teacher Quality in the Educational Production Function," Working Paper No. 2007-03, (Columbia, MO: University of Missouri, 2007); D.F. McCaffrey et al., "The Intertemporal Variability of Teacher Effect Estimates," *Education Finance and Policy* 4, no. 4 (2009).

[35] Dan Goldhaber and Michael Hansen, "Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions," Brief No. 3, (Washington, DC: National Center for Analysis of Longitudinal Data in Education Research, 2008); L. Lefgren and D. Sims, "Using Subject Test Scores Efficiently to Predict Teacher Value-Added," *Educational Evaluation and Policy Analysis* 34, no. 1 (2012).

[36] Lefgren and Sims, "Using Subject Test Scores Efficiently to Predict Teacher Value-Added". When VAM is used to identify and fire the bottom quartile (or bottom quintile) of teachers based on the year t teacher rankings, the results

of the studies cited here imply that this decision is incorrect, according to the year t + 1 teacher rankings, for a minimum of 51 percent of the teachers in that quartile (or quintile), for an overall error rate of 51 percent. Note that reliability is defined here in terms of intertemporal stability. This is the criterion that is examined in the cited studies and is appropriate because the inference when teachers are ranked is that the ranking is stable and reflects a stable trait—i.e., a low-quality teacher this year will remain a low-quality teacher next year. The conclusion that the reliability of VAM is poor is consistent with the conclusion of the National Research Council's Board on Testing and Assessment that VAM is not sufficiently reliable for the purpose of terminating teachers. See Edward H. Haertel, "Letter Report to the U.S. Department of Education on the Race to the Top Fund," The National Academies, https://download.nap.edu/catalog.php?record_id=12780.

The results reported by Lefgren and Sims in Table 4 require additional explanation. For each estimate, Lefgren and Sims reported a second, higher estimate (in brackets) using the method in Jacob and Lefgren. See Brian A. Jacob and Lars Lefgren, "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education," *Journal of Labor Economics* 26, no. 1 (2008). However, the Jacob and Lefgren method assumes "that the true ability of teacher j is distributed normally with a mean equal to the estimated empirical Bayes value added for teacher j"—in other words, Jacob and Lefgren assumed that the true ability of teacher j is centered on the value added estimate for teacher j, which is incorrect if the teacher's true ability fluctuates up and down over time, as indicated by Goldhaber and Hansen's results. See ibid., 132; Dan Goldhaber and Michael Hansen, "Is It Just a Bad Class? Assessing the Long-Term Stability of Estimated Teacher Performance," *Economica* 80, no. 319 (2013). If true ability fluctuates over time, Lefgren and Sims' bracketed estimates in Table 4 are incorrect. The correct estimates are the estimates that are not in brackets; those estimates never exceed 50 percent, indicating that the consistency of the value-added estimates never exceeds 50 percent regardless of whether one, two, three, four, or five years of data are used.

[37] Cory Koedel and Julian R. Betts, "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique," *Education Finance and Policy* 6, no. 1 (2011); Jesse Rothstein, "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables," *Education Finance and Policy* 4, no. 4 (2009); Jesse Rothstein, "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *Quarterly Journal of Economics* 125, no. 1 (2010).

[38] See Henry Braun, Naomi Chudowsky, and Judith Koenig, eds., *Getting Value out of Value-Added: Report of a Workshop*, National Research Council, Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Accountability (Washington, DC: The National Academies Press, 2010), 49.

[39] See ibid., 49-50.

[40] While the effect size for the Reading Assessment intervention is not large, the intervention is relatively inexpensive and could potentially be implemented throughout the K-12 years. In comparison, more expensive interventions would be prohibitively costly to implement throughout the K-12 years. A sustained 12-year intervention throughout the K-12 years may be expected to have a substantial cumulative impact even if the annual impact is relatively small. Furthermore, a sustained intervention would address the phenomenon of fadeout once an intervention is discontinued. Further research is needed to evaluate the impact of implementing the Reading Assessment intervention throughout the K-12 years.

Chapter 6

[1] Darity, “Stratification Economics: The Role of Intergroup Inequality”.

[2] Patrick Bayer and Robert McMillan, “Tiebout Sorting and Neighborhood Stratification,” *Journal of Public Economics* 96, no. 11–12 (2012).

[3] Ibid.

[4] Ibid.

[5] Rivkin, Hanushek, and Kain, “Teachers, Schools and Academic Achievement”; Rowan, Correnti, and Miller, “What Large-Scale Survey Research Tells Us About Teacher Effects on Student Achievement”; Staiger and Rockoff, “Searching for Effective Teachers with Imperfect Information”; Wright, Horn, and Sanders, “Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation”; Sanders and Rivers, “Cumulative and Residual Effects of Teachers on Future Student Academic Achievement”.

[6] Aaronson, Barrow, and Sander, “Teachers and Student Achievement in the Chicago Public High Schools”; Ballou, “Value-Added Assessment: Lessons from Tennessee.”; Koedel and Betts, “Re-Examining the Role of Teacher Quality in the Educational Production Function”; McCaffrey et al., “The Intertemporal Variability of Teacher Effect Estimates”; Goldhaber and Hansen, “Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions”; Lefgren and Sims, “Using Subject Test Scores Efficiently to Predict Teacher Value-Added”.

[7] Aaronson, Barrow, and Sander, “Teachers and Student Achievement in the Chicago Public High Schools”; Ballou, “Value-Added Assessment: Lessons from Tennessee.”; Koedel and Betts, “Re-Examining the Role of Teacher Quality in the Educational Production Function”; McCaffrey et al., “The Intertemporal Variability of Teacher Effect Estimates”.

[8] Goldhaber and Hansen, “Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions”; Lefgren and Sims, “Using Subject Test Scores Efficiently to Predict Teacher Value-Added”.

[9] Lefgren and Sims, “Using Subject Test Scores Efficiently to Predict Teacher Value-Added”.

[10] Aaronson, Barrow, and Sander, "Teachers and Student Achievement in the Chicago Public High Schools"; Ballou, "Value-Added Assessment: Lessons from Tennessee."; Koedel and Betts, "Re-Examining the Role of Teacher Quality in the Educational Production Function"; McCaffrey et al., "The Intertemporal Variability of Teacher Effect Estimates"; Goldhaber and Hansen, "Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions"; Lefgren and Sims, "Using Subject Test Scores Efficiently to Predict Teacher Value-Added".

[11] When VAM is used to categorize teachers into top-quintile and bottom-quintile teachers, the result is highly unstable. If 100 teachers are categorized as "high" and 100 teachers are categorized as "low" at time t, the studies cited here imply that when the VAM ranking procedure is repeated at time t+1, the result is that less than 50 of the 100 teachers who were categorized as "high" at time t are categorized as "high" at time t+1. Less than 50 of the 100 teachers who were categorized as "low" at time t are categorized as "low" at time t+1.

[12] Koedel and Betts, "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness?"; Rothstein, "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables"; Rothstein, "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement".

[13] Braun, Chudowsky, and Koenig, *Getting Value out of Value-Added*, 49; Rothstein, "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables".

[14] Braun, Chudowsky, and Koenig, *Getting Value out of Value-Added*, 49.

[15] Ibid., 49–50.

[16] McCaffrey et al., "The Intertemporal Variability of Teacher Effect Estimates", 577.

[17] Rothstein, "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement".

[18] McCaffrey et al., "The Intertemporal Variability of Teacher Effect Estimates", 578.

[19] Goldhaber and Hansen, "Is It Just a Bad Class?".

[20] See E. A. Hanushek, "Teacher Deselection," in *Creating a New Teaching Profession*, ed. Dan Goldhaber and Jane Hannaway (Washington, DC: Urban Institute Press, 2009); Robert Gordon, Thomas J. Kane, and Douglas O. Staiger, "Identifying Effective Teachers Using Performance on the Job," Discussion Paper No. 2006–01, (Washington, D.C.: The Brookings Institution, 2006); McCaffrey et al., "The Intertemporal Variability of Teacher Effect Estimates"; Staiger and Rockoff, "Searching for Effective Teachers with Imperfect Information".

[21] Koedel and Betts, "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness?"; Rothstein, "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables"; Rothstein, "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement".

[22] See Skinner et al., "Individual Differences and the Development of Perceived Control".

[23] National Board for Professional Teaching Standards, "About Us," http://www.nbpts.org/who-we-are.

[24] Educational Testing Service, "Where We Stand on Teacher Quality," https://www.ets.org/Media/Education_Topics/pdf/teacherquality.pdf; National Board for Professional Teaching Standards, "About Us".

[25] National Board for Professional Teaching Standards, "Guide to National Board Certification," http://boardcertifiedteachers.org/sites/default/files/Guide_to_NB_Certification.pdf.

[26] Dan Goldhaber, David Perry, and Emily Anthony, "NBPTS Certification: Who Applies and What Factors Are Associated with Success?," (Seattle, WA: University of Washington, 2003).

[27] Drew Gitomer, "Reliability and NBPTS Assessments," in *Assessing Teachers for Professional Certification: The First Decade of the National Board for Professional Teaching Standards*, ed. Lawrence Ingvarson and John Hattie, *Advances in Program Evaluation* (Amsterdam: Elsevier, 2008), 241.

[28] Ibid., 231.

[29] Charles T. Clotfelter, Helen F. Ladd, and Jacob L. Vigdor, "Teacher-Student Matching and the Assessment of Teacher Effectiveness," *Journal of Human Resources* 41, no. 4 (2006); Dan Goldhaber and Emily Anthony, "Can Teacher Quality Be Effectively Assessed? National Board Certification as a Signal of Effective Teaching," *Review of Economics and Statistics* 89, no. 1 (2007); L. Cavalluzzo, "Is National Board Certification an Effective Signal of Teacher Quality?," The CNA Corporation, http://www.nbpts.org/sites/default/files/documents/research/Cavalluzzo_IsNBCAnEffectiveSignalofTeachingQuality.pdf; W. L. Sanders, James J. Ashton, and S. Paul Wright, "Comparison of the Effects of NBPTS Certified Teachers with Other Teachers on the Rate of Student Academic Progress," (Cary, NC: SAS Institute, Inc., 2005); D. N. Harris and T. R. Sass, "The Effects of NBPTS-Certified Teachers on Student Achievement," (Madison, WI: University of Wisconsin, 2007); H. F. Ladd, T. R. Sass, and D. N. Harris, "The Impact of National Board Certified Teachers on Student Achievement in Florida and North Carolina: A Summary of the Evidence Prepared for the National Academies Committee on the Evaluation of the Impact of Teacher Certification by NBPTS," Washington, DC: The National Academies, 2007; Charles T. Clotfelter, Helen F. Ladd, and Jacob L. Vigdor, "Teacher Credentials and Student Achievement: Longitudinal Analysis with Student Fixed Effects," *Economics of Education Review* 26, no. 6 (2007). See also National Research Council, "The Impact of Board-Certified Teachers on Student Outcomes," in *Assessing Accomplished Teaching: Advanced-Level Certification Programs*, ed. Milton D. Hakel, Judith Anderson Koenig, and Stuart W. Elliott (Washington, DC: National Academies Press, 2008).

[30] S. S. Yeh, "The Cost-Effectiveness of NBPTS Teacher Certification," *Evaluation Review* 34, no. 3 (2010).

[31] See L. Bond et al., “The Certification System of the National Board for Professional Teaching Standards: A Construct and Consequential Validity Study,” (Greensboro, NC: Center for Educational Research and Evaluation, The University of North Carolina at Greensboro, 2000); L. G. Vandevoort, A. Amrein-Beardsley, and D. C. Berliner, “National Board Certified Teachers and Their Students’ Achievement,” *Education Policy Analysis Archives* 12, no. 46 (2004), http://epaa.asu.edu/ojs/article/view/201; Wendy McColskey et al., “Teacher Effectiveness, Student Achievement, and National Board Certified Teachers: A Comparison of National Board Certified Teachers and Non-National Board Certified Teachers: Is There a Difference in Teacher Effectiveness and Student Achievement?,” National Board for Professional Teaching Standards, http://www.education-consumers.com/articles/W-M%20NBPTS%20certified%20report.pdf; J.E. Stone, “The Value-Added Achievement Gains of NBPTS-Certified Teachers in Tennessee: A Brief Report,” http://www.education-consumers.com/briefs/stoneNBPTS.shtm.

[32] U.S. Department of Education, “New No Child Left Behind Flexibility: Highly Qualified Teachers,” http://www2.ed.gov/nclb/methods/teachers/hqtflexibility.html.

[33] Clotfelter, Ladd, and Vigdor, “Teacher Credentials and Student Achievement: Longitudinal Analysis with Student Fixed Effects”.

[34] Rob De Jong and Klaas J. Westerhof, “The Quality of Student Ratings of Teacher Behavior,” *Learning Environments Research* 4, no. 1 (2001).

[35] Kathy D. Tuck, “Parent Satisfaction and Information (a Customer Satisfaction Survey),” (Washington, DC: District of Columbia Public Schools, Office of Educational Accountability, Assessment and Information, 1995).

[36] Rothstein, “Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables”; Rothstein, “Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement”; Jesse Rothstein, “Revisiting the Impacts of Teachers,” University of California, Berkeley, Goldman School of Public Policy and Department of Economics, October, 2015, http://eml.berkeley.edu/~jrothst/workingpapers/rothstein_cfr_oct2015.pdf.

[37] Gitomer, “Reliability and NBPTS Assessments.”

[38] NCES did not report the reliability of the measures used in the study reported in Chapter Six. However, a sensitivity analysis may be performed that explores the sensitivity of the results to a range of plausible values for measurement reliability. This analysis suggests that the results reported here are not sensitive to measurement reliability. Spearman’s formula for disattenuating sample correlations is:

$$\hat{\rho}_{xy} = r_{xy} \left[r_{xx} + r_{yy}\right]^{-1/2}$$

where:

$\hat{\rho}_{xy}$ = estimated population correlation of true scores

r_{xy} = observed sample correlation

r_{xx} = reliability of variable x

r_{yy} = reliability of variable y

Table F.1 Figure 6.4 Disattenuated Correlations

rxy	*low reliability of x* (*rxx = .69; ryy = .9*)	*high reliability of x* (*rxx = ryy = .9*)
0.16	0.20	0.18
0.18	0.23	0.20
0.14	0.18	0.16
mean	0.20	0.18

Table F.2 Figure 6.5 Disattenuated Correlations

rxy	*low reliability of x* (*rxx = .69; ryy = .9*)	*high reliability of x* (*rxx = ryy = .9*)
0.17	0.22	0.19
0.25	0.32	0.28
0.34	0.43	0.38
mean	0.32	0.28

With regard to parent estimates of school quality, plausible values of item reliabilities are suggested by a survey involving responses from 3,948 District of Columbia public school parents that yielded internal reliabilities of survey items for each section of the survey ranging from .69 to .90. See Tuck, "Parent Satisfaction and Information (a Customer Satisfaction Survey)". Using these figures, Table F.1 reports disattenuated correlations for the path coefficients for school quality on math achievement (see Figure 6.4). Using the low (.69) estimate of reliability for the school quality measure, the values of the disattenuated correlations for school quality on math achievement range from 0.18 to 0.23. Table F.2 reports disattenuated correlations for the path coefficients for self-efficacy on math achievement (see Figure 6.5). Using the high (.9) estimate of reliability, the values of the disattenuated correlations for self-efficacy on math achievement range from 0.19 to 0.38. Even under unfavorable assumptions regarding measurement reliability (low reliability of the school quality measure and high reliability of the self-efficacy measure), the average effect of self-efficacy on achievement remains stronger than the average effect of school quality on achievement, and the effect of self-efficacy on achievement doubles between grade 3 and grade 12.

[39] Sam Dillon, "Formula to Grade Teachers' Skill Gains Acceptance, and Critics," *The New York Times*, September 1, 2010; U.S. Department of Education, "Teacher Incentive Fund," U.S. Department of Education, http://www2.ed.gov/programs/teacherincentive/index.html.

[40] Crooks, "The Impact of Classroom Evaluation Practices on Students", 464; Dweck, "Motivational Processes Affecting Learning", 1041; Dweck and Elliott, "Achievement Motivation."

[41] Kennelly, Dietz, and Benson, "Reinforcement Schedules, Effort vs. Ability Attributions, and Persistence".

[42] Blankenship, "Individual Differences in Resultant Achievement Motivation and Latency to and Persistence at an Achievement Task"; Danner and Lonsky, "A Cognitive-Developmental Approach to the Effects of Rewards on Intrinsic

Motivation"; M. R. Dougherty and J. I. Harbison, "Motivated to Retrieve: How Often Are You Willing to Go Back to the Well When the Well Is Dry?," *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33, no. 6 (2007); Drucker et al., "Relation of Task Difficulty to Persistence"; Harter, "Pleasure Derived from Optimal Challenge and the Effects of Extrinsic Rewards on Children's Difficulty Level Choices"; Kennelly, Dietz, and Benson, "Reinforcement Schedules, Effort vs. Ability Attributions, and Persistence"; Nygard, "Personality, Situation, and Persistence: A Study with Emphasis on Achievement Motivation".

[43] Fuchs and Fuchs, "Effects of Systematic Formative Evaluation: A Meta-Analysis"; Kluger and DeNisi, "The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory"; Bangert-Drowns et al., "The Instructional Effect of Feedback in Test-Like Events"; Mac Iver, Stipek, and Daniels, "Explaining within-Semester Changes in Student Effort in Junior High School and Senior High School Courses"; Ryan, Mims, and Koestner, "The Relationship of Reward Contingency and Interpersonal Context to Intrinsic Motivation: A Review and Test Using Cognitive Evaluation Theory".

[44] S. S. Yeh, "The Cost-Effectiveness of 22 Approaches for Raising Student Achievement," *Journal of Education Finance* 36, no. 1 (2010).

[45] Nunnery, Ross, and McDonald, "A Randomized Experimental Evaluation of the Impact of Accelerated Reader/Reading Renaissance Implementation on Reading Achievement in Grades 3 to 6"; Ross, Nunnery, and Goldfeder, "A Randomized Experiment on the Effects of Accelerated Reader/Reading Renaissance in an Urban School District: Final Evaluation Report"; Ysseldyke and Bolt, "Effect of Technology-Enhanced Continuous Progress Monitoring on Math Achievement".

[46] National Center for Education Statistics, "Early Childhood Longitudinal Program (ECLS): Kindergarten Class of 2010–11 (ECLS-K:2011)," https://nces.ed.gov/ecls/kindergarten2011.asp.

[47] National Center for Education Statistics, "Early Childhood Longitudinal Program (ECLS): Kindergarten Class of 1998–99 (ECLS-K)".

[48] National Center for Education Statistics, "National Education Longitudinal Study of 1988 (NELS: 88)," http://nces.ed.gov/surveys/nels88/index.asp.

[49] For example, the inclusion of race dummies as covariates would cause the coefficient for self-efficacy to drop to zero in a model predicting achievement if the only students who exhibit low self-efficacy and low achievement are black and Hispanic students and the only students who exhibit high self-efficacy and high achievement are white and Asian students. The expectation and presence of collinearity dictates that race be excluded in this model if the guiding theory is that grading practices depress black and Hispanic students' self-efficacy, depressing effort and achievement, and if the purpose of the analysis is to estimate the relationship between levels of self-efficacy and levels of achievement. Note that race dummies are included in the path analyses as exogenous predictors of achievement upon entry at kindergarten.

Similarly, it would not be appropriate to include school quality or teacher quality as a covariate along with self-efficacy in a model predicting achievement if the only students exhibiting low self-efficacy and low achievement are minority students who attend urban schools rated low in quality or are taught by urban teachers rated low in quality, and the only students exhibiting high self-efficacy and high achievement are white students who attend suburban schools rated high in quality or are taught by suburban teachers rated high in quality, and if the purpose of the analysis is to estimate the relationship between levels of self-efficacy and levels of achievement. The expectation and presence of collinearity dictates that school quality and teacher quality be excluded as predictors in this model if the guiding theory is that grading and testing practices depress the self-efficacy, effort, and achievement of minority students who are concentrated in urban schools that are subsequently rated low in quality, while the same practices tend to maintain high levels of self-efficacy, effort, and achievement of white students who are concentrated in suburban schools that are subsequently rated high in quality, and if the purpose of the analysis is to estimate the relationship between levels of self-efficacy and levels of achievement.

The appropriate strategy is the strategy that is adopted here, i.e., to estimate separate models for each of the two theories regarding the persistence of the achievement gap without inserting extraneous predictors that are expected to be highly correlated with the predictors of interest, and then compare the magnitudes of the path coefficients.

[50] It might be argued that prior achievement should be included as a covariate in any model predicting student achievement. For example, a student who performs poorly with regard to addition and subtraction does not have the foundation to advance to multiplication and division, and so forth. But this begs the question about why the student is performing poorly with regard to addition and subtraction. The arrays of assessments that are administered to students starting at entry into kindergarten quickly reveal areas of weakness. Teachers are well-informed about those areas of weakness and have ample opportunities to address those weaknesses.

When students continue to perform poorly despite attention from teachers, one possible explanation is that the quality of teaching is poor. However, significant resources and attention have been devoted to improve the quality of teaching, with disappointing results. Furthermore, as explained in the review of literature (above), value-added measures of teacher quality are highly unreliable and the statistical approach used to calculate teacher rankings appears to be invalid. The lack of reliability suggests that existing VAM models are based on incorrect assumptions and an incorrect understanding of the sources of poor performance. There is a need to reexamine those assumptions and consider alternative views.

When students continue to perform poorly despite attention from teachers, a second possible explanation is that low-achieving students are subjected throughout their academic careers to a steady diet of negative comments, grades, test scores, and other cues that trigger and reinforce declines in self-efficacy, triggering reduced effort and achievement. The current study employs path analysis to compare both theories.

The inclusion of prior achievement as a covariate in any statistical model predicting student achievement is problematic because it dominates the influence of other predictors. One might argue that the data simply indicate the need to drop other, insignificant predictors. However, there are two problems. First, any analysis suggesting that achievement is mainly a function of prior achievement offers little insight into factors that mediate the relationship between prior achievement and later achievement. Second, the effect of prior achievement may simply represent the cumulative effect of a factor (such as erosion of self-efficacy) that is slow and steady but is dominated in any statistical model that includes prior achievement as a covariate.

To draw an analogy, a build-up of corrosion can cause an engine to malfunction. The immediate cause of the malfunction is corrosion. If a statistical analysis were to be performed involving cases of corrosion, an indicator of the level of corrosion would dominate other factors, such as the level of moisture seeping into engine components during periods when the engine is not used. However, in these cases the root cause of engine malfunction is moisture, not corrosion.

The current analysis seeks to compare and contrast two theories regarding the root cause of the persistent achievement gap. Prior achievement is a proximate factor that strongly predicts current achievement, but it is not the root cause of current differences in achievement.

[51] A. Bandura, "Self-Efficacy: Toward a Unifying Theory of Behavior Change," *Psychological Review* 84(1977).

[52] Various indices are available to compare nested models where one parent model includes all of the predictors specified in comparison models that contain a subset of the predictors. However, Figure 6.1 and Figure 6.2 describe two non-nested models. It is not appropriate to compare them to a parent model that includes both sets of predictors because indicators of school and teacher quality are expected to be highly correlated with indicators of student self-efficacy, implying collinearity. Recall that Figure 6.1 describes a model where school and teacher quality are expected to be correlated with race and socioeconomic status. Figure 6.2 describes a model where student self-efficacy is expected to be correlated with race and socioeconomic status. Therefore, Figure 6.1 and Figure 6.2 describe models where the key predictors are expected to be correlated with race and socioeconomic status, implying that the key predictors are expected to be correlated with each other and it would not be appropriate to include them in the same statistical model.

[53] S. S. Yeh, "Two Models of Learning and Achievement: An Explanation for the Achievement Gap?," *Teachers College Record* 117, no. 12 (2015).

[54] Coleman et al., "Equality of Educational Opportunity".

[55] Ibid., 319, 322.

[56] Ibid.

[57] Rivkin, Hanushek, and Kain, "Teachers, Schools and Academic Achievement"; Rowan, Correnti, and Miller, "What Large-Scale Survey

Research Tells Us About Teacher Effects on Student Achievement"; Staiger and Rockoff, "Searching for Effective Teachers with Imperfect Information"; Wright, Horn, and Sanders, "Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation"; Sanders and Rivers, "Cumulative and Residual Effects of Teachers on Future Student Academic Achievement".

[58] Rothstein, "Revisiting the Impacts of Teachers".

[59] Ibid., 32.

[60] See Rivkin, Hanushek, and Kain, "Teachers, Schools and Academic Achievement"; Rowan, Correnti, and Miller, "What Large-Scale Survey Research Tells Us About Teacher Effects on Student Achievement"; Staiger and Rockoff, "Searching for Effective Teachers with Imperfect Information"; Wright, Horn, and Sanders, "Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation"; Sanders and Rivers, "Cumulative and Residual Effects of Teachers on Future Student Academic Achievement".

Chapter 7

[1] See S. S. Yeh, "The Cost-Effectiveness of Five Policies for Improving Student Achievement," *American Journal of Evaluation* 28, no. 4 (2007); S. S. Yeh, "The Cost-Effectiveness of Comprehensive School Reform and Rapid Assessment," *Education Policy Analysis Archives* 16, no. 13 (2008), http://epaa.asu.edu/ojs/article/view/38/164; S. S. Yeh, "Class Size Reduction or Rapid Formative Assessment? A Comparison of Cost-Effectiveness," *Educational Research Review* 4, no. 1 (2009); S. S. Yeh, "The Cost-Effectiveness of Raising Teacher Quality," *Educational Research Review* 4, no. 3 (2009); Yeh, "The Cost-Effectiveness of NBPTS Teacher Certification"; Yeh, "The Cost-Effectiveness of 22 Approaches for Raising Student Achievement"; S. S. Yeh, *The Cost-Effectiveness of 22 Approaches for Raising Student Achievement* (Charlotte, NC: Information Age Publishing, 2011); S. S. Yeh, "The Reliability, Impact and Cost-Effectiveness of Value-Added Teacher Assessment Methods," *Journal of Education Finance* 37, no. 4 (2012); S. S. Yeh, "A Re-Analysis of the Effects of Teacher Replacement Using Value-Added Modeling," *Teachers College Record* 115, no. 12 (2013); S. S. Yeh and J. Ritter, "The Cost-Effectiveness of Replacing the Bottom Quartile of Novice Teachers through Value-Added Teacher Assessment," *Journal of Education Finance* 34, no. 4 (2009).

[2] See, for example Bowen, Chingos, and McPherson, *Crossing the Finish Line: Completing College at America's Public Universities*; Saul Geiser and Maria Veronica Santelices, "Validity of High-School Grades in Predicting Student Success Beyond the Freshman Year: High-School Record vs. Standardized Tests as Indicators of Four-Year College Outcomes," Research and Occasional Paper Series No. CSHE.6.07: University of California, Berkeley, 2007); Saul Geiser and Roger Studley, "UC and the SAT: Predictive Validity and Differential Impact of the SAT I and SAT II at the University of California," *Educational Assessment* 8, no. 1 (2002); Sunny X. Niu and Marta Tienda, "Testing, Ranking

and College Performance: Does High School Matter?," Princeton University, September, 2009.

[3] Jencks and Phillips, *The Black-White Test Score Gap*; Christopher Jencks, and others, *Who Gets Ahead? The Determinants of Economic Success in America* (Basic Books, 1979); Otis D. Duncan, David L. Featherman, and Beverly Duncan, *Socioeconomic Background and Achievement* (New York: Seminar Press, 1972); R. J. Herrnstein and C. Murray, *The Bell Curve: Intelligence and Class Structure in American Life* (New York: Simon and Schuster, 1994); C. Winship and S.D. Korenman, "Economic Success and the Evolution of Schooling and Mental Ability," in *Earning and Learning: How Schools Matter*, ed. S. E. Mayer and P. E. Peterson (Washington, DC: Brookings Institution Press, 1999).

[4] Leonard Ramist, Charles Lewis, and Laura McCamley-Jenkins, "Student Group Differences in Predicting College Grades: Sex, Language, and Ethnic Groups," College Board Report No. 93–1, (New York: College Entrance Examination Board, 1994); Wayne J. Camara and Gary Echternacht, "The SAT I and High School Grades: Utility in Predicting Success in College," Research Note No. RN-10, (New York: The College Entrance Examination Board, 2000).

[5] Bowen and Bok, *The Shape of the River: Long-Term Consequences of Considering Race in College and University Admissions.*

[6] Bowen, Kurzweil, and Tobin, *Equity and Excellence in American Higher Education.*

[7] Ibid.

[8] Ibid., 224.

[9] Bowen, Chingos, and McPherson, *Crossing the Finish Line: Completing College at America's Public Universities.*

[10] For a review, see Nancy W. Burton and Leonard Ramist, "Predicting Success in College: SAT Studies of Classes Graduating since 1980," College Board Research Report No. 2001–2, (New York: College Entrance Examination Board, 2001).

[11] Heinrich Stumpf and Julian C. Stanley, "Group Data on High School Grade Point Averages and Scores on Academic Aptitude Tests as Predictors of Institutional Graduation Rates," *Educational and Psychological Measurement* 62, no. 6 (2002).

[12] Alexander W. Astin and Leticia Oseguera, *Degree Attainment Rates at American Colleges and Universities* (Los Angeles, CA: Higher Education Research Institute, Graduate School of Education, University of California, Los Angeles, 2005).

[13] Leslie Scott, Steven J. Ingels, and Jeffrey A. Owings, "Interpreting 12th-Graders' NAEP-Scaled Mathematics Performance Using High School Predictors and Postsecondary Outcomes from the National Education Longitudinal Study of 1988 (NELS:88)," Statistical Analysis Report No. NCES 2007–328, (Washington, DC: National Center for Education Statistics, 2007).

[14] Clifford Adelman, "Answers in the Tool Box: Academic Intensity, Attendance Patterns, and Bachelor's Degree Attainment," (Washington, DC: U.S. Department of Education, 1999).

[15] Charles F. Manski and David A. Wise, *College Choice in America* (Cambridge, MA: Harvard University Press, 1983).

[16] Ibid.

[17] Brent Bridgeman, Judy Pollack, and Nancy Burton, "Understanding What SAT Reasoning Test Scores Add to High School Grades: A Straightforward Approach," Research Report No. 2004–4, (New York: College Entrance Examination Board, 2004).

[18] Ibid.

[19] College Board, "Net Prices by Income over Time: Public Sector," http://trends.collegeboard.org/college-pricing/figures-tables/net-prices-income-over-time-public-sector.

[20] Ibid.

[21] Ibid.

[22] Ibid.

[23] Calculated from Bowen, Chingos, and McPherson, *Crossing the Finish Line: Completing College at America's Public Universities*, 169, Figure 9.2.

[24] Calculated from ibid., 171, Figure 9.4.

[25] David Leonhardt, "Measuring Colleges' Success in Enrolling the Less Affluent," *The New York Times*, September 9, 2014.

[26] Committee on Governmental Affairs, *Testimony by Caroline M. Hoxby, "the Rising Cost of College Tuition and the Effectiveness of Government Financial Aid"*, 106th Congress, 2d, February 9, 2000.

[27] Ibid.

[28] C. Avery and C. M. Hoxby, "Do and Should Financial Aid Packages Affect Students' College Choices?," in *College Choices: The Economics of Where to Go, When to Go, and How to Pay for It*, ed. C. M. Hoxby (Chicago: University of Chicago Press, 2004); Pedro Carneiro and James J. Heckman, "Human Capital Policy," in *Inequality in America: What Role for Human Capital Policies?*, ed. James J. Heckman and Alan Krueger (Cambridge, MA: MIT Press, 2003).

[29] T.J. Kane, *The Price of Admission: Rethinking How Americans Pay for College* (Washington, D.C.: Brookings Institution Press, 1999).

[30] Bowen, Kurzweil, and Tobin, *Equity and Excellence in American Higher Education*, 91.

[31] R. Spies, "The Effect of Rising Costs on College Choice," Research Report No. 117, (Princeton, N.J.: Princeton University, 2001), 17.

[32] The National Commission on Excellence in Education, "A Nation at Risk: The Imperative for Educational Reform," Report No. 065-000-00177-2, (Washington, DC: The National Commission on Excellence in Education, 1983).

[33] U.S. Department of Education, "Department of Education Budget Tables," http://www2.ed.gov/about/overview/budget/tables.html?src=ct.

[34] National Center for Education Statistics, "NAEP 2008 Trends in Academic Progress," NCES Report No. 2009–479, (Washington, DC: U.S. Department of Education, 2008).

[35] Hattie and Timperley, "The Power of Feedback".

[36] Bangert-Drowns et al., "The Instructional Effect of Feedback in Test-Like Events"; Fuchs and Fuchs, "Effects of Systematic Formative Evaluation: A Meta-Analysis"; Kluger and DeNisi, "The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory"; Black and Wiliam, "Assessment and Classroom Learning".

[37] Black and Wiliam, "Assessment and Classroom Learning".

[38] Fuchs and Fuchs, "Effects of Systematic Formative Evaluation: A Meta-Analysis".

[39] Kluger and DeNisi, "The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory", 278.

[40] Lysakowski and Walberg, "Instructional Effects of Cues, Participation, and Corrective Feedback: A Quantitative Synthesis".

[41] Walberg, "What Makes Schooling Effective?".

[42] Tenenbaum and Goldring, "A Meta-Analysis of the Effect of Enhanced Instruction: Cues, Participation, Reinforcement and Feedback and Correctives on Motor Skill Learning".

Chapter 8

[1] Will Dobbie and Roland G. Fryer, Jr., "Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone," *American Economic Journal: Applied Economics* 3, no. 3 (2011): 158.

[2] Christina Clark Tuttle et al., "Student Characteristics and Achievement in 22 KIPP Middle Schools," (Washington, DC: Mathematica Policy Research, 2010), xi.

[3] Vilsa E. Curto, Roland G. Fryer, Jr., and Meghan L. Howard, "It May Not Take a Village: Increasing Achievement among the Poor," in *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances*, ed. Greg J. Duncan and Richard J. Murnane (New York: Russell Sage Foundation, 2011), 500.

[4] Katrina R. Woodworth et al., "San Francisco Bay Area KIPP Schools: A Study of Early Implementation and Achievement: Final Report," (Menlo Park, CA: SRI International, 2008); Jeffrey R. Henig, "What Do We Know About the Outcomes of KIPP Schools?," (Tempe, AZ: Arizona State University, 2008); Curto, Fryer, and Howard, "It May Not Take a Village."

[5] See D. T. Campbell and J. C. Stanley, "Experimental and Quasi-Experimental Designs for Research on Teaching," in *Handbook of Research on Teaching*, ed. N. L. Gage (Chicago: Rand McNally, 1963).

[6] Curto, Fryer, and Howard, "It May Not Take a Village," 500.

[7] When researchers examine teacher attrition, they typically compare sample statistics regarding observable characteristics such as gender, teaching experience, and credentials. If both groups of teachers are comparable on these observable measures, then researchers tend to conclude that differential attri-

tion was not a problem. However, the problem that I have identified is different. It is invisible to researchers because it is typically unmeasured and unobserved. It is the difference in the level of dedication between teachers who stay, versus teachers who quit. To measure and observe this, it would be necessary for researchers to design and administer a measure of "dedication" to both groups of teachers and then use the results to investigate the possibility of differential attrition. This test was not performed in either of the two studies of HCZ and KIPP that provide the strongest evidence of effectiveness. See Dobbie and Fryer, "Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone"; Tuttle et al., "Student Characteristics and Achievement in 22 KIPP Middle Schools".

[8] Dobbie and Fryer, "Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone"; Tuttle et al., "Student Characteristics and Achievement in 22 KIPP Middle Schools"; Woodworth et al., "San Francisco Bay Area KIPP Schools: A Study of Early Implementation and Achievement: Final Report"; Joshua D. Angrist et al., "Who Benefits from KIPP?," *Journal of Policy Analysis and Management* 31, no. 4 (2012).

[9] Dobbie and Fryer, "Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone", 171.

[10] Ibid., 131. 1.146 SD/6 years = 0.191 SD per year; 0.570 SD/6 years = 0.095 SD per year. Annualization of the effect sizes permits comparisons with the annualized effect sizes reported in Table 5.1.

[11] Grover J. Whitehurst and Michelle Croft, "The Harlem Children's Zone, Promise Neighborhoods, and the Broader, Bolder Approach to Education," (Washington, DC: Brookings Institution, 2010), 9.

[12] Dobbie and Fryer, "Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone", 162.

[13] Anderson Cooper, "Harlem's Education Experiment Gone Right," CBS News, http://www.cbsnews.com/stories/2009/12/04/60minutes/main5889558_page2.shtml?tag=contentMain;contentBody.

[14] Gail Collins, "Waiting for Somebody," *The New York Times*, September 30, 2010, A35.

[15] Karl Weber, ed. *Waiting for "Superman": How We Can Save America's Failing Public Schools* (New York: Public Affairs, 2010), 194.

[16] Dobbie and Fryer, "Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone", 162.

[17] Paul Tough, *Whatever It Takes: Geoffrey Canada's Quest to Change Harlem and America* (Boston: Houghton Mifflin Harcourt, 2008).

[18] Dobbie and Fryer, "Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone", 162.

[19] Tough, *Whatever It Takes: Geoffrey Canada's Quest to Change Harlem and America.*

[20] Ibid.

[21] Ibid., 165.

[22] Ibid., 172.

[23] Ibid., 251.

[24] Ibid., 252.

[25] Tuttle et al., "Student Characteristics and Achievement in 22 KIPP Middle Schools", 28. The evaluation report does not indicate the average gain in math achievement.

[26] Ibid., 30. The evaluation report does not indicate the average gain in reading achievement.

[27] Woodworth et al., "San Francisco Bay Area KIPP Schools: A Study of Early Implementation and Achievement: Final Report", 32.

[28] Ibid.

[29] Curto, Fryer, and Howard, "It May Not Take a Village," 500.

[30] Woodworth et al., "San Francisco Bay Area KIPP Schools: A Study of Early Implementation and Achievement: Final Report", 33.

[31] Ibid., 35.

[32] Ibid., 34.

[33] Ibid., 35.

[34] Ibid., 34.

[35] Ibid.

[36] Ibid.

[37] Ibid., 35.

[38] Samuel Casey Carter, "No Excuses: Seven Principals of Low-Income Schools Who Set the Standard for High Achievement," (Washington, DC: Heritage Foundation, 1999), 19.

[39] Ibid., 17.

[40] Ibid., 19.

[41] Ibid., 19, italics in original.

[42] Ibid., 19.

[43] Ibid.

[44] Ibid., 20.

[45] Ibid., 20.

[46] If f[x] is nonlinear, it would be possible to obtain an increase in aggregate student achievement if teachers are systematically redistributed in a purposeful manner. For example, if f[x] is concave with respect to each school, aggregate gains in student achievement could be obtained by shifting highly-dedicated teachers from a high-b school to a low-b school. When f[x] is concave, the gain obtained by the low-b school is larger than the loss to the high-b school. The opposite would be true if f[x] is convex. Thus, it would be possible to obtain an increase in aggregate achievement even if c=0 in the aggregate. However, any gains from global implementation would have to come via systematic redistribution, and the magnitude of those gains would be limited by the available

pool of highly-dedicated teachers. There is no obvious reason to expect this type of systematic redistribution of teachers.

[47]It is possible that performance might remain high even if the unusually dedicated staff goes away. However, that result would undermine the basic assumption underlying the KIPP/HCZ models, which emphasizes the importance of recruiting dedicated teachers. It is unlikely that the high performance of the KIPP/HCZ schools is due to the extended school day. Previous research indicates that lengthening the school day by 60 minutes improves student achievement by only 0.03 SD in math and 0.07 SD in reading. See H. M. Levin, G.V. Glass, and G. Meister, "A Cost-Effectiveness Analysis of Computer-Assisted Instruction," *Evaluation Review* 11, no. 1 (1987). This accounts for only a small portion of the effects attributed to the KIPP and HCZ schools. However, the performance of the KIPP/HCZ schools might be attributable to the use of an extended school year. Meta-analytic results of summer school programs estimate a median effect size of 0.19 SD. See H. Cooper et al., "Making the Most of Summer School: A Meta-Analytic and Narrative Review," *Monographs of the Society for Research in Child Development* 65, no. 1 (2000). The results of the Tennessee class size experiment suggest that a portion of the effects of KIPP and HCZ may be due to small class sizes. See J.D. Finn et al., "The Enduring Effects of Small Classes," *Teachers College Record* 103, no. 2 (2001); B. Nye, L. V. Hedges, and S. Konstantopoulos, "The Long-Term Effects of Small Classes: A Five-Year Follow-up of the Tennessee Class Size Experiment," *Educational Evaluation and Policy Analysis* 21, no. 2 (1999); B. Nye, L. V. Hedges, and S. Konstantopoulos, "Are Effects of Small Classes Cumulative? Evidence from a Tennessee Experiment," *Journal of Educational Research* 94, no. 6 (2001). However, evidence from natural experiments suggests that the Tennessee results are likely due to Hawthorne effects. See J. Levin, "For Whom the Reductions Count: A Quantile Regression Analysis of Class Size and Peer Effects on Scholastic Achievement," *Empirical Economics* 26(2001); C. M. Hoxby, "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *The Quarterly Journal of Economics* 115, no. 4 (2000); L. Wößmann, "International Evidence on Expenditures and Class Size: A Review," in *Brookings Papers on Education Policy: 2006/2007*, ed. T. Loveless and F. Hess (Washington, D.C.: Brookings Institution Press, 2007). Thus, the small class sizes that are characteristic of the KIPP and HCZ schools are unlikely to explain the putative effects of KIPP and HCZ on student achievement. Finally, it is unlikely that the high performance of the KIPP/HCZ schools is due to the enhanced expenditure per pupil that is associated with KIPP and HCZ schools. Based on meta-analytic results, Greenwald, Hedges, and Laine estimated that a 10 percent increase in expenditure per pupil increases student achievement in math and reading by only 0.083 SD per year. See Rob Greenwald, Larry V. Hedges, and Richard D. Laine, "The Effect of School Resources on Student Achievement," *Review of Educational Research* 66, no. 3 (1996). In contrast, Hanushek found that student achievement may be improved by an entire grade-level if high-performing teachers are

substituted for low-performing teachers, suggesting that the most likely source of the high performance of the KIPP and HCZ schools is the recruitment of highly-dedicated, high-performing teachers. See E. A. Hanushek, "The Trade-Off between Child Quantity and Quality," *Journal of Political Economy* 100, no. 1 (1992).

[48] Hanushek, "Teacher Deselection."

[49] Ibid., 12.

[50] Curto, Fryer, and Howard, "It May Not Take a Village," 500.

[51] Carter, "No Excuses: Seven Principals of Low-Income Schools Who Set the Standard for High Achievement", 20.

[52] R. M. Ingersoll, "Teacher Turnover and Teacher Shortages: An Organizational Analysis," *American Education Research Journal* 38, no. 3 (2001).

[53] Curto, Fryer, and Howard, "It May Not Take a Village," 500.

[54] See, for example, ibid., 500.

[55] Gordon, Kane, and Staiger, "Identifying Effective Teachers Using Performance on the Job".

[56] Dobbie and Fryer, "Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone", 158.

Chapter 9

[1] Braun, Chudowsky, and Koenig, *Getting Value out of Value-Added*.

[2] E. A. Hanushek, "The Difference Is Great Teachers," in *Waiting for 'Superman': How We Can Save America's Failing Public Schools*, ed. Karl Weber (New York: Public Affairs, 2010); Staiger and Rockoff, "Searching for Effective Teachers with Imperfect Information"; Hanushek, "Teacher Deselection."; Gordon, Kane, and Staiger, "Identifying Effective Teachers Using Performance on the Job"; S. Glazerman et al., "America's Teacher Corps," (Washington, DC: Brown Center on Education Policy at Brookings, 2010); Frederick M. Hess, "Beyond School Choice," *National Review*, October 18, 2010.

[3] Dillon, "Formula to Grade Teachers' Skill Gains Acceptance, and Critics"; The Center for Greater Philadelphia, "Value-Added Assessment," University of Pennsylvania, http://www.cgp.upenn.edu/ope_value.html#9.

[4] Linda Wesson, Kim Potts, and Kelly Hill, "Use of Value-Added in Teacher Evaluations: Key Concepts and State Profiles," (Nashville, TN: Tennessee Comptroller of the Treasury, Offices of Research and Education Accountability, 2015), 2.

[5] Ibid.

[6] Bill Turque, "D.C. Teacher Evaluation Formula Could Change," The Washington Post, http://www.washingtonpost.com/local/education/dc-teacher-evaluation-formula-could-change/2012/05/24/gJQACOyRoU_story.html.

[7] Wesson, Potts, and Hill, "Use of Value-Added in Teacher Evaluations: Key Concepts and State Profiles", 2.

[8] Ibid.

[9] Matthew Johnson et al., "Value-Added Models for the Pittsburgh Public Schools," (Cambridge, MA: Mathematica Policy Research, 2012).

[10] New York State Department of Education, "New York State Teacher and Principal Evaluation: Summary of Provisions in Draft Regulations," http://usny.nysed.gov/rttt/docs/summary.pdf; State Council for Educator Effectiveness, "Report and Recommendations," Colorado State Department of Education, https://www.cde.state.co.us/sites/default/files/documents/educatoreffectiveness/downloads/report%20%26%20appendices/scee_final_report.pdf; Teresa Watanabe, "'Value-Added' Teacher Evaluations: L.A. Unified Tackles a Tough Formula," Los Angeles Times, http://articles.latimes.com/2011/mar/28/local/la-me-adv-value-add-20110328; Louisiana Department of Education, "ESEA Flexibility Renewal Form: Louisiana," Board of Elementary and Secondary Education, (Baton Rouge, LA: Louisiana Department of Education, 2015).

[11] Wesson, Potts, and Hill, "Use of Value-Added in Teacher Evaluations: Key Concepts and State Profiles", 2.

[12] Dillon, "Formula to Grade Teachers' Skill Gains Acceptance, and Critics"; U.S. Department of Education, "Teacher Incentive Fund".

[13] Raj Chetty, John N. Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review* 104, no. 9 (2014); Annie Lowrey, "Big Study Links Good Teachers to Lasting Gain," *The New York Times*, January 6, 2012.

[14] Chetty, Friedman, and Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood", 2636, footnote 3.

[15] Nicholas D. Kristof, "The Value of Teachers," The New York Times, http://www.nytimes.com/2012/01/12/opinion/kristof-the-value-of-teachers.html; Nicholas D. Kristof, "The Value of Teachers," *The International Herald Tribune*, January 13, 2012.

[16] Rothstein, "Revisiting the Impacts of Teachers", 32.

[17] Ibid.

[18] Sanders and Rivers, "Cumulative and Residual Effects of Teachers on Future Student Academic Achievement".

[19] Wright, Horn, and Sanders, "Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation".

[20] Rivkin, Hanushek, and Kain, "Teachers, Schools and Academic Achievement"; Rowan, Correnti, and Miller, "What Large-Scale Survey Research Tells Us About Teacher Effects on Student Achievement"; Staiger and Rockoff, "Searching for Effective Teachers with Imperfect Information".

[21] A. Milanowski, "The Relationship between Teacher Performance Evaluation Scores and Student Achievement: Evidence from Cincinnati," *Peabody Journal of Education* 79, no. 4 (2004); H. C. Hill, L. Kapitula, and K. Umland, "A Validity Argument Approach to Evaluating Teacher Value-Added Scores," *American Educational Research Journal* 48, no. 3 (2011);

J. Schacter and Y. M. Thum, “Paying for High- and Low-Quality Teaching,” *Economics of Education Review* 23(2004); See also Jacob and Lefgren, “Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education”.

[22] Hanushek, “The Difference Is Great Teachers.”; Staiger and Rockoff, “Searching for Effective Teachers with Imperfect Information”; Hanushek, “Teacher Deselection.”; Gordon, Kane, and Staiger, “Identifying Effective Teachers Using Performance on the Job”.

[23] Staiger and Rockoff, “Searching for Effective Teachers with Imperfect Information”, 108.

[24] S. Glazerman et al., “Evaluating Teachers: The Important Role of Value-Added,” (Washington, DC: Brookings, 2010), 6.

[25] Braun, Chudowsky, and Koenig, *Getting Value out of Value-Added*, 49.

[26] Ibid.

[27] Rothstein, “Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables”; Rothstein, “Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement”; Rothstein, “Revisiting the Impacts of Teachers”.

[28] Braun, Chudowsky, and Koenig, *Getting Value out of Value-Added*, 49; Rothstein, “Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables”.

[29] Braun, Chudowsky, and Koenig, *Getting Value out of Value-Added*, 49–50.

[30] Rothstein, “Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables”, 537.

[31] Ibid., 539.

[32] Ibid., 566; Koedel and Betts, “Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness?”.

[33] Koedel and Betts, “Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness?”.

[34] Ibid.

[35] Xiaoxia A. Newton et al., “Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts,” *Education Policy Analysis Archives* 18, no. 23 (2010), http://epaa.asu.edu/ojs/article/view/810.

[36] D. B. Rubin, E. A. Stuart, and E. L. Zanutto, “A Potential Outcomes View of Value-Added Assessment in Education,” *Journal of Behavioral and Educational Statistics* 29, no. 1 (2004).

[37] Jun Ishii and Steven G. Rivkin, “Impediments to the Estimation of Teacher Value Added,” *Education Finance and Policy* 4, no. 4 (2009).

[38] Ibid.

[39] Derek Briggs and Ben Domingue, “A Review of the Value-Added Analysis Underlying the Effectiveness Rankings of Los Angeles Unified School District Teachers by the Los Angeles Times,” (Boulder, CO: University of Colorado at Boulder, 2011).

[40] Ibid.

[41] Ibid.

[42] Dale Ballou, W. Sanders, and P. Wright, "Controlling for Student Background in Value-Added Assessment of Teachers," *Journal of Behavioral and Educational Statistics* 29, no. 1 (2004).

[43] D.F. McCaffrey et al., "Models for Value-Added Modeling of Teacher Effects," *Journal of Behavioral and Educational Statistics* 29, no. 1 (2004).

[44] H. Kupermintz, "Teacher Effects and Teacher Effectiveness: A Validity Investigation of the Tennessee Value Added Assessment System," *Educational Evaluation and Policy Analysis* 25, no. 3 (2003).

[45] S. W. Raudenbush, "What Are Value-Added Models Estimating and What Does This Imply for Statistical Practice?," *Journal of Behavioral and Educational Statistics* 29, no. 1 (2004).

[46] Goldhaber and Hansen, "Is It Just a Bad Class?".

[47] Ibid., 591.

[48] Ibid., 605.

[49] Hanushek, "Teacher Deselection."

[50] Gordon, Kane, and Staiger, "Identifying Effective Teachers Using Performance on the Job".

[51] McCaffrey et al., "The Intertemporal Variability of Teacher Effect Estimates".

[52] Staiger and Rockoff, "Searching for Effective Teachers with Imperfect Information".

[53] See Goldhaber and Hansen, "Is It Just a Bad Class?", 597.

[54] See Dan Goldhaber and Michael Hansen, "Is It Just a Bad Class? Assessing the Long-Term Stability of Estimated Teacher Performance," Working Paper No. 73, (Washington, DC: National Center for Analysis of Longitudinal Data in Education Research, 2012), 31.

[55] M. D. Reckase, "The Real World Is More Complicated Than We Would Like," *Journal of Behavioral and Educational Statistics* 29, no. 1 (2004).

[56] J. A. Martineau, "Distorting Value-Added: The Use of Longitudinal, Vertically-Scaled Student Achievement Data for Growth-Based, Value-Added Accountability," *Journal of Educational and Behavioral Statistics* 31, no. 1 (2006).

[57] Ibid., 57–58.

[58] J. P. Papay, "Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates across Outcome Measures," *American Educational Research Journal* 48, no. 1 (2011).

[59] Chetty, Friedman, and Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood", 2637.

[60] Ibid., 2638.

[61] Ibid.

[62] Raj Chetty, John N. Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review* 104, no. 9 (2014): 2594.

[63] Ibid., 2600.

[64] Chetty, Friedman, and Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood", 2673.

[65] Ibid.

[66] Goldhaber and Hansen, "Is It Just a Bad Class?".

[67] Aaronson, Barrow, and Sander, "Teachers and Student Achievement in the Chicago Public High Schools".

[68] Ballou, "Value-Added Assessment: Lessons from Tennessee."

[69] Koedel and Betts, "Re-Examining the Role of Teacher Quality in the Educational Production Function".

[70] McCaffrey et al., "The Intertemporal Variability of Teacher Effect Estimates".

[71] Suppose that half of a sample of teachers is fired, using a coin flip to determine the fate of each teacher. From Table 9.1, a minimum of 60 percent of all teachers deserve to be retained, while 40 percent do not, according to the year t + 1 teacher rankings, even when the sample is drawn from the bottom quartile, as determined by year t VAM rankings. The coin flip results in firing half of those who deserve retention (30 percent of all teachers) and retention of half of those who deserve firing (20 percent of all teachers) for an overall error rate of 50 percent. In comparison, when VAM is used to identify and fire the bottom quartile (or bottom quintile) of teachers, the results of Tables 9.1 and 9.2 imply that this decision is incorrect, according to the year t + 1 teacher rankings, for a minimum of 59 percent of the teachers in that quartile (or quintile).

[72] The validity of using test scores for a particular purpose depends on "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores." See American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing* (Washington, DC: American Psychological Association, 1985), 9. Validity "refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests." See American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing* (Washington, DC: American Educational Research Association, 1999), 9. In the case of teacher rankings based on value-added test scores, the inference that the results reliably categorize teachers as either high-performing or low-performing teachers is not appropriate, nor does the available evidence support the use of value-added teacher rankings for the purpose of high-stakes decisions regarding hiring, firing, promotion, or compensation.

[73] McCaffrey et al., "The Intertemporal Variability of Teacher Effect Estimates".

[74] Newton et al., "Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts".

[75] Goldhaber and Hansen, "Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions".

[76] Ibid.

[77] Ibid.

[78] Lefgren and Sims, “Using Subject Test Scores Efficiently to Predict Teacher Value-Added”. For each estimate, Lefgren and Sims reported a second, higher estimate (in brackets) using the method in Jacob and Lefgren. See Jacob and Lefgren, “Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education”. However, the Jacob and Lefgren method assumes “that the true ability of teacher j is distributed normally with a mean equal to the estimated empirical Bayes value added for teacher j”—in other words, Jacob and Lefgren assumed that the true ability of teacher j is centered on the value-added estimate for teacher j, which is incorrect if the teacher’s true ability fluctuates up and down over time, as indicated by Goldhaber and Hansen’s results. See ibid., 132; Goldhaber and Hansen, “Is It Just a Bad Class?”. If true ability fluctuates over time, Lefgren and Sims’ bracketed estimates in Table 4 are incorrect. The correct estimates are the estimates that are not in brackets; those estimates never exceed 50 percent, indicating that the consistency of the value-added estimates never exceeds 50 percent regardless of whether one, two, three, four, or five years of data are used.

[79] Chetty, Friedman, and Rockoff, “Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood”, 2646.

[80] Ibid., 2654.

[81] Ibid.

[82] Ibid., 2671.

[83] Ibid., 2655.

[84] Chetty et al. do not explain the discrepancy between this figure and their assertion that undiscounted gains for a classroom of 28.2 students equal $1.4 million. See ibid., 2636 footnote 3.

[85] Ibid., 2671.

[86] Chetty, Friedman, and Rockoff, “Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates”, 2630.

[87] Rothstein, “Revisiting the Impacts of Teachers”, 3. After receiving Rothstein’s preliminary results, Chetty et al. argued that they reflect a problem with his test for bias, rather than with Chetty et al.’s research design: “Because teacher value-added is estimated using data from students in the same schools in previous years, teachers will tend to have high value-added estimates when their students happened to do well in prior years. Regressing changes in prior test scores on changes in teacher value-added effectively puts the same data on the left- and right-hand side of the regression, mechanically yielding a positive coefficient.” See ibid., 18. As noted by Rothstein, Chetty et al. point to two potential sources of such “mechanical” effects. “First, some teachers who teach grade-g students in t or t – 1 might have taught the same cohorts of students previously, in grade g–1 in t–1 or t–2 (or in grade g–2 in t–2 or t–3). This could induce a positive correlation between the teachers’ effectiveness and the students’ g – 1 scores—in effect, these prior-year scores are intermediate outcomes of the effectiveness of the grade g teacher. Second, even when teachers do not follow students across grades, a mechanical effect could arise from the

fact that data from t – 2 is used both to measure the prior-year achievement of t–1 students and to forecast the t–1 teachers' value-added. Any shock that is common across grades in the school-year cell could create a positive correlation between the measured value-added of the t – 1 teachers and the t – 2 scores of the t – 1 students, biasing the placebo coefficient upward." See ibid., 18–19. Rothstein, however, conducted alternative tests for bias that excluded the possibility of any mechanical effects. Rothstein formed a predicted score for each student using only demographic variables: the student's age and indicators for gender, ethnicity, free-lunch status, special education, limited English, and grade repetition, along with class and school-year means. It would be impossible to generate the type of mechanical correlation described by Chetty et al. between these predicted student scores and the teachers' value-added scores. Rothstein, however, found that changes in predicted student scores based on these demographic characteristics are significantly associated with changes in the teachers' value-added scores: "This conclusively establishes that the placebo result cannot be attributed to the mechanical explanations proposed by Chetty et al." See ibid., 19. Rothstein pointed out that Chetty et al. dropped teachers from their sample who were only observed at time t or time t-1. Rothstein reported results suggesting that this introduced sorting bias that resulted in biased estimates of teacher quality. It would not be uncommon for this type of sorting bias to occur in routine applications of VAM. For large numbers of teachers, it may be expected that student test score data would only be available at time t, or time t-1, or would be missing. See Newton et al., "Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts". The bottom line is that estimates of teacher quality using VAM are not unbiased.

[88] Rothstein, "Revisiting the Impacts of Teachers", 8.

[89] Ibid., 2.

[90] Ibid., 32.

[91] See Rivkin, Hanushek, and Kain, "Teachers, Schools and Academic Achievement"; Rowan, Correnti, and Miller, "What Large-Scale Survey Research Tells Us About Teacher Effects on Student Achievement"; Staiger and Rockoff, "Searching for Effective Teachers with Imperfect Information"; Wright, Horn, and Sanders, "Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation"; Sanders and Rivers, "Cumulative and Residual Effects of Teachers on Future Student Academic Achievement".

[92] Chetty, Friedman, and Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood", 2674.

[93] Thomas J. Kane and Douglas O. Staiger, "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," Working Paper No. 14607, (Cambridge, MA: National Bureau of Economic Research, 2008).

[94] See also Scott E. Carrell and James E. West, "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors," *Journal of Political Economy* 118, no. 3 (2010); Brian A. Jacob, Lars Lefgren,

and David P. Sims, "The Persistence of Teacher-Induced Learning Gains," *Journal of Human Resources* 45, no. 4 (2010); Rothstein, "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement"; Goldhaber and Hansen, "Is It Just a Bad Class? Assessing the Long-Term Stability of Estimated Teacher Performance"; Spyros Konstantopoulos, "Teacher Effects in Early Grades: Evidence from a Randomized Study," *Teachers College Record* 113, no. 7 (2011); McCaffrey et al., "The Intertemporal Variability of Teacher Effect Estimates".

[95] While Chetty et al. assume that the estimated gain in earnings at age 28 is an accurate indicator of the gain in earnings throughout each student's working lifetime, they acknowledge that "Teachers' impacts on test scores fade out rapidly in subsequent years and appear to stabilize at approximately 25 percent of the initial impact after three to four years." It is unclear why it would make sense to assume that earning gains do not fade in the same way that test score gains fade. See Chetty, Friedman, and Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood", 2658.

[96] Ibid., Appendix Table 10.

[97] Chetty, Friedman, and Rockoff, "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates", 2610.

[98] 0.993 × 0.116 = 0.115 SD; 0.221 × 0.116 = 0.026 SD; 0.993 × 0.042 = 0.042 SD; 0.221 × 0.042= 0.009 SD. Chetty et al. state the implied relationship in an early version of their article. See Raj Chetty, John N. Friedman, and Jonah E. Rockoff, "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood," Working Paper 17699, (Cambridge, MA: National Bureau of Economic Research, 2011), 23–24.

[99] Calculation of the average gain in student achievement across all students enables comparisons with alternative interventions that may be applied across all students, not merely a small fraction of all students. See, for example, Table 5.1 in Chapter Five, which compares the effect sizes for 23 interventions for raising student achievement. It would be misleading to display, in a league table, an unadjusted effect size for an intervention that can only be applied to a small fraction of all students. A league table comparison suggests that all effect sizes in the table are comparable and may be compared to the effect sizes for alternative interventions that may be applied across all students.

[100] See Chetty, Friedman, and Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood", 2655.

[101] Ibid., 2672.

[102] Ibid.

[103] Ibid., 2673.

[104] Ibid., 2672.

[105] Ibid.

[106] Newton et al., "Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts".

[107] Kristof, "The Value of Teachers".

[108] Chetty, Friedman, and Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood", 2672–2673.

[109] See ibid., 2672.

[110] Chetty, Friedman, and Rockoff, "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood", 49. The statement was deleted from the published version of Chetty et al.'s article.

[111] See McCaffrey et al., "The Intertemporal Variability of Teacher Effect Estimates".

[112] See E. A. Hanushek et al., "The Market for Teacher Quality," Working Paper 11154, (Cambridge, MA: National Bureau of Economic Research, 2005); H. Wenglinsky, "Teacher Classroom Practices and Student Performance: How Schools Can Make a Difference," Research Report No. RR-01-19, (Princeton, NJ: Educational Testing Service, 2001); D. W. Grissmer et al., *Improving Student Achievement: What State NAEP Test Scores Tell Us* (Santa Monica, CA: RAND Corporation, 2000); Gordon, Kane, and Staiger, "Identifying Effective Teachers Using Performance on the Job".

[113] U.S. Department of Education, "Teacher Shortage Areas: Nationwide Listing, 1990–91 Thru 2011–12," (Washington, DC: U.S. Department of Education, 2011).

[114] Gordon, Kane, and Staiger, "Identifying Effective Teachers Using Performance on the Job".

[115] U.S. Department of Education, "Teacher Shortage Areas: Nationwide Listing, 1990–91 Thru 2011-12". An oversupply of teachers in large urban districts that are reducing their teaching forces may permit experienced teachers to be hired to replace low-performing teachers who are terminated.

[116] McCaffrey et al., "The Intertemporal Variability of Teacher Effect Estimates".

[117] Hanushek et al., "The Market for Teacher Quality"; Wenglinsky, "Teacher Classroom Practices and Student Performance: How Schools Can Make a Difference"; Grissmer et al., *Improving Student Achievement: What State NAEP Test Scores Tell Us*; Gordon, Kane, and Staiger, "Identifying Effective Teachers Using Performance on the Job".

[118] Yeh, "The Reliability, Impact and Cost-Effectiveness of Value-Added Teacher Assessment Methods".

[119] Hanushek et al., "The Market for Teacher Quality"; Wenglinsky, "Teacher Classroom Practices and Student Performance: How Schools Can Make a Difference"; Grissmer et al., *Improving Student Achievement: What State NAEP Test Scores Tell Us*; Gordon, Kane, and Staiger, "Identifying Effective Teachers Using Performance on the Job".

[120] Chetty, Friedman, and Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood", 2672, footnote 37.

[121] Gordon, Kane, and Staiger, "Identifying Effective Teachers Using Performance on the Job". See the authors' footnote 7 for conversion of percentile point scores into SD units.

[122] Yeh, "The Reliability, Impact and Cost-Effectiveness of Value-Added Teacher Assessment Methods"; Yeh and Ritter, "The Cost-Effectiveness of Replacing the Bottom Quartile of Novice Teachers through Value-Added Teacher Assessment".

[123] It may be argued that low-performing teachers are not well-matched to the occupation of teaching and, therefore, there would be a gain to society if those individuals are redirected to other occupations. However, the numerator of the benefit-cost ratio accounts for this gain to society, measured in terms of the increase in the lifetime earnings of students taught by teacher replacements who are presumably better suited to the occupation of teaching than the terminated teachers. In addition, the hypothesis that fired teachers are better suited to other occupations is not supported by the available evidence. Only 3.8 percent of new female elementary teachers and 5.4 percent of new female high school teachers who left full-time teaching during the 1994–2001 time period took a noneducation sector job in Georgia that paid more than the state minimum teaching salary in Georgia. See Benjamin Scafidi, David L. Sjoquist, and Todd R. Stinebrickner, "Do Teachers Really Leave for Higher Paying Jobs in Alternative Occupations?," *The B.E. Journal of Economic Analysis and Policy* 6, no. 1 (2006). Since these figures include all exiting teachers, including teachers who left voluntarily and, therefore, were likely to be considered more productive by potential employers than teachers who were fired, it is likely that the percentage of fired teachers who took noneducation sector jobs paying more than the state minimum teaching salary is even lower. This implies that well over 94 percent of fired teachers are unable to earn more in their new occupations. Fired teachers are not more productive in new occupations.

[124] U.S. Department of Education, "Digest of Education Statistics," National Center for Education Statistics, http://nces.ed.gov/programs/digest/d05/tables/dt05_075.asp.

[125] Note that the 3 percent discount rate used by Chetty, et al. to discount earnings is based on their assumption of 2 percent wage growth adjusted for a 5 percent discount rate. See Chetty, Friedman, and Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood", 2654. For consistency, I use the same assumptions for wage growth and the discount rate as Chetty, et al. These assumptions are slightly different than the assumptions used in my previous publications. See Yeh and Ritter, "The Cost-Effectiveness of Replacing the Bottom Quartile of Novice Teachers through Value-Added Teacher Assessment"; Yeh, "The Reliability, Impact and Cost-Effectiveness of Value-Added Teacher Assessment Methods".

The best available estimate of the career duration of the average teacher was derived using proportional hazards modeling, which accounts for the difficulty of estimating career duration when some members of the research sample have not exited the teaching profession by the end of the research study period. See Richard J. Murnane, Judith D. Singer, and John B. Willett, "The Career Paths of Teachers: Implications for Teacher Supply and Methodological Lessons for

Research," *Educational Researcher* 17, no. 6 (1988). Proportional hazards modeling incorporates information about the pattern of teacher attrition during the study period to predict the median length of each spell of teaching. Using data from Michigan covering a 12-year time period, Murnane, Singer, and Willett provided separate estimates, for six subject area specialties, of the duration of the average teacher's first two spells of teaching. See ibid. The authors reported the percentage distribution of teachers across the six subject area specialties, as well as the percentage of teachers in each of the six subject areas who returned to teaching after a career interruption. I used this information to calculate the average career duration (9.11 years) for an average teacher, weighted by the percentage distribution of teachers across the six subject area specialties and including the expected length of a second spell of teaching based on the probability of a second spell.

[126] Congressional Budget Office, "An Analysis of the Administration's Proposed Program for Displaced Workers," (Washington, D.C.: Congressional Budget Office, 1994).

[127] Alfred O. Gottschalck, "Dynamics of Economic Well-Being: Spells of Unemployment 2001–2003," Report No. P70-105, (Washington, D.C.: U.S. Census Bureau, 2006).

[128] Gordon, Kane, and Staiger, "Identifying Effective Teachers Using Performance on the Job".

[129] Suppose that half of a sample of teachers is fired, using a coin flip to determine the fate of each teacher. From Table 9.1, a minimum of 60 percent of all teachers deserve to be retained, while 40 percent do not, according to the year t + 1 teacher rankings, even when the sample is drawn from the bottom quartile, as determined by year t VAM rankings. The coin flip results in firing half of those who deserve retention (30 percent of all teachers) and retention of half of those who deserve firing (20 percent of all teachers) for an overall error rate of 50 percent. In comparison, when VAM is used to identify and fire the bottom quartile (or bottom quintile) of teachers, the results of Tables 9.1 and 9.2 imply that this decision is incorrect, according to the year t + 1 teacher rankings, for a minimum of 59 percent of the teachers in that quartile (or quintile).

[130] NY Daily News, "The Tenure Trap," http://www.nydailynews.com/opinion/tenure-trap-careful-hiring-teachers-stuck-life-article-.443471.

[131] Associated Press, "Superintendent: Bad Teachers Hard to Fire: Some Say Teacher Tenure Rules Need to Be Overhauled to Address Problem," http://www.nbcnews.com/id/25430476/ns/us_news-education/t/superintendent-bad-teachers-hard-fire/#.V362f-srLIU.

[132] Ibid.

[133] Daniel Weisberg et al., "The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness," (Brooklyn, NY: The New Teacher Project, 2009).

[134] Ibid.

[135] Haertel, "Letter Report to the U.S. Department of Education on the Race to the Top Fund", 10.

[136] Michael Blacher, "K-12 Teacher Termination Hearings: Are They Worth the Cost?," *CPER Journal*, no. 180 (October, 2006).

[137] Ibid.; Andy Nixon, Abbot Packard, and Gus Douvanis, "Non-Renewal of Probationary Teachers: Negative Retention," *Education* 131, no. 1 (2010): 43.

[138] Blacher, "K-12 Teacher Termination Hearings: Are They Worth the Cost?". Terminating a teacher using VAM would likely be even more litigious and costly if a court agrees with the judgment of the National Research Council's Board on Testing and Assessment, which concluded that VAM is not sufficiently reliable for the purpose of terminating teachers. See Haertel, "Letter Report to the U.S. Department of Education on the Race to the Top Fund". The NRC's report represents an independent judgment that VAM is unreliable.

[139] Blacher, "K-12 Teacher Termination Hearings: Are They Worth the Cost?".

[140] Ibid.

[141] Patricia Gould, "3020-a Process Remains Slow, Costly," New York State School Boards Association, http://www.nyssba.org/news/2009/05/11/on-board-online-may-11-2009/3020-a-process-remains-slow-costly/.

[142] S. E. Bratton, Jr., S. P. Horn, and S. P. Wright, "Using and Interpreting Tennessee's Value-Added Assessment System: A Primer for Teachers and Principals,"http://www.shearonforschools.com/documents/TVAAS.HTML#Let's.

[143] Gordon, Kane, and Staiger, "Identifying Effective Teachers Using Performance on the Job"; U.S. Department of Education, "Teacher Shortage Areas: Nationwide Listing, 1990–91 Thru 2011-12".

[144] C. F. Manski, "Academic Ability, Earnings, and the Decision to Become a Teacher: Evidence from the National Longitudinal Study of the High School Class of 1972," in *Public Sector Payrolls*, ed. D. Wise (Chicago: University of Chicago Press, 1987).

[145] National Center for Education Statistics, "Digest of Education Statistics," U.S. Department of Education, http://nces.ed.gov/programs/digest/d05/.

[146] U.S. Department of Labor, "Employer Costs for Employee Compensation: December 2007," Bureau of Labor Statistics, http://www.bls.gov/news.release/archives/ecec_03122008.pdf.

[147] Henry M. Levin, "Cost-Effectiveness and Educational Policy," *Educational Evaluation and Policy Analysis* 10, no. 1 (1988).

[148] Effectiveness-cost ratios for rapid performance assessment were previously calculated by the author. See Yeh, "The Cost-Effectiveness of 22 Approaches for Raising Student Achievement"; Yeh, "The Cost-Effectiveness of Comprehensive School Reform and Rapid Assessment".

[149] See Chetty, Friedman, and Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood", 2655, footnote 27.

[150] Bureau of Labor Statistics, "Consumer Price Index," U.S. Department of Labor, http://www.bls.gov/cpi/cpid1509.pdf.

[151]These calculations assume that Chetty et al.'s proposal is implemented on an ongoing annual basis. The benefits and costs are calculated per student in each cohort that would benefit from the intervention. However, it might be argued that each replacement teacher generates a "legacy" stream of student cohorts that benefit from the increased productivity of that teacher and, therefore, the benefit of replacing each teacher should be multiplied by the number of cohorts taught by each teacher. This might be accurate if no fired teacher was rehired, if the composition of the teaching force was frozen and there were no retirements or exits by any teacher in the entire teaching force, and annual culling of the teaching force reliably eliminated low-performing teachers. To clarify:

a.) If all fired teachers were rehired by other schools, the benefit of Chetty et al.'s proposal would drop to zero, but most of the costs would remain.

b.) If some fired teachers were rehired by other schools, the benefit of Chetty et al.'s proposal would diminish in proportion to the degree of rehiring.

c.) In principle, the ability of the federal government to regulate policies regarding the rehiring of fired teachers is limited because the U.S. Constitution effectively delegates this role to the states.

d.) While the federal government might be able to require, as a condition of receiving federal education funds, that each state must implement a policy forbidding the rehiring of a fired teacher, it is likely that such a regulation would be widely opposed for two reasons: i.) If VAM is used to identify and fire the bottom quartile (or quintile) of teachers, the results in Tables 9.1 and 9.2 indicate that this decision is incorrect, according to the year t + 1 teacher rankings, between 59 and 70 percent of the time, ii.) there is a teacher shortage and many classrooms could not be staffed under such a requirement.

e.) In the absence of a federal requirement forbidding the rehiring of fired teachers, each state would establish its own policy. The benefit of Chetty et al.'s proposal would depend on each state and/or locality establishing such a requirement. It is likely that such a regulation would be widely opposed for both of the reasons specified in (d), above.

f.) However, even under the strong assumption that no state permits the rehiring of fired teachers, the dynamics of the teacher labor market would cause the legacy benefits of Chetty et al.'s proposal to fade every year, unless the culling process envisioned in Chetty et al.'s proposal is implemented on an ongoing annual basis. The reason is that a significant portion of the entire teaching force exits every year and is replaced by a new set of teachers with heterogeneous abilities. This "waters down" the composition of the upper 95 percent that was retained under Chetty et al.'s policy with novices representing the full distribution—the full bell curve—of teacher ability. To understand the issue, consider an extreme example: assume that attrition is 100 percent after one year and all teachers are replaced with novices. Clearly, there would be no legacy benefit of Chetty et al.'s proposal because the distribution of teacher performance would reflect the entire bell curve from that point forward. If attrition is 50 percent, then the benefits of the proposal are cut in half. Benefits are reduced even if attrition is limited entirely to the upper 95 percent of all teachers because attrition and replacement causes that group of teachers to take on the characteristics of the full distribution of teachers—the entire bell curve, not just the upper 95 percent of the bell curve. The only way to maintain the upper 95 percent advantage that is presumably conferred by Chetty et al.'s proposal is through ongoing culling. Data from the national Schools and Staffing Survey indicate teacher retention rates of 76 percent after two years, 67 percent after three years, 60 percent after four years, and 54 percent after five years. See K.H. Quartz et al., "Urban Teacher Retention Policy: A Research Brief," Report No. rrs-rr006-0704, (Los Angeles: University of California, Los Angeles, 2004). These figures include retention in all roles within the field of education, not only teaching, implying that the teacher retention rate is lower and attrition is a significant problem.

g.) Would annual culling gradually improve the stock of teachers over time, overcoming slippage due to attrition? This depends on the validity of the assumptions underlying Chetty et al.'s analysis. In particular, if VAM is used to identify and fire the bottom quartile (or quintile) of teachers, the results in Tables 9.1 and 9.2 indicate that this decision is incorrect, according to the year t + 1 teacher rankings, for 59 to 70 percent of the teachers. These results suggest that productive teachers would be culled more frequently than unproductive bottom quartile (or bottom quintile) teachers. The problem is illustrated by data from six large urban school districts indicating that an English language arts teacher who is predicted, based on VAM, to score at the 25th percentile is actually more likely to fall in the top half of the distribution than in the bottom quarter. See Jesse Rothstein, "Review of *Learning About Teaching*," University of California at Berkeley, 2011), 8–9. Depending on the distribution of teachers, this implies the possibility that a VAM-based decision rule to fire the bottom quartile of teachers could actually reduce the quality of the teaching force! A second implication is that it would be extremely difficult to justify any policy prohibiting the rehiring of fired teachers, if in fact teachers who are predicted to score at the 25th percentile are actually more likely to score above average. Third, it raises serious doubts, not only about the validity of Chetty et al.'s analysis, but all proposals to use VAM-based decision rules to fire low-performing teachers. The evidence affirms the conclusion of the NRC's expert panel that VAM is not sufficiently reliable to make operational decisions about firing teachers. See Haertel, "Letter Report to the U.S. Department of Education on the Race to the Top Fund".

h.) Given the evidence presented in this chapter, plus the strong likelihood that many (perhaps most) fired teachers would be rehired (because prohibitions against rehiring would be difficult to justify), plus the slippage of gains due to attrition and, most importantly, the evidence that VAM-based teacher rankings fluctuate up and down and are poor predictors of future performance, both the short- and long-term benefits of Chetty et al.'s policy are questionable. Using VAM, a teacher who is ranked "poor" this year is more likely to be classified as a productive teacher next year than to remain a poor teacher. Therefore, the gains from replacing "poor" teachers are questionable and it would be questionable to multiply those presumed gains by the number of cohorts taught by each teacher. In any case, the average career duration of a teacher is 9.11 years; multiplying the effectiveness-cost ratio for Chetty et al.'s intervention by 9.11 gives a ratio equal to 0.00009 (9.11 X 0.00001). However, rapid performance assessment remains 191 times more cost-effective than Chetty et al.'s intervention. Even after accounting for the legacy benefits of Chetty et al.'s policy, it is far less cost-effective than performance feedback. While additional research is needed to clarify each of these issues, the burden is on advocates to demonstrate that VAM-based teacher replacement is a cost-effective strategy, compared to rapid performance feedback and other leading alternatives.

[152] Haertel, "Letter Report to the U.S. Department of Education on the Race to the Top Fund".

[153] Chetty, Friedman, and Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood"; Gordon, Kane, and Staiger, "Identifying Effective Teachers Using Performance on the Job"; McCaffrey et al., "The Intertemporal Variability of Teacher Effect Estimates".

[154] Yeh, "The Cost-Effectiveness of 22 Approaches for Raising Student Achievement"; Yeh, "The Reliability, Impact and Cost-Effectiveness of Value-Added Teacher Assessment Methods"; Yeh and Ritter, "The Cost-Effectiveness of Replacing the Bottom Quartile of Novice Teachers through Value-Added Teacher Assessment".

[155] Goldhaber and Hansen, "Is It Just a Bad Class?", 591.

[156] Ibid., 605.

[157] Glazerman et al., "Evaluating Teachers: The Important Role of Value-Added", 6.

[158] Yeh, "The Cost-Effectiveness of 22 Approaches for Raising Student Achievement".

[159] Yeh, "High Stakes Testing".

[160] Ibid.; S. S. Yeh, "Understanding and Addressing the Achievement Gap through Individualized Instruction and Formative Assessment," *Assessment in Education* 17, no. 2 (2010).

Chapter 10

[1] Yeh, *The Cost-Effectiveness of 22 Approaches for Raising Student Achievement.*

[2] Ibid., xix-xx.

[3] Peter Schrag, *Final Test: The Battle for Adequacy in America's Schools* (New York: New Press, 2005); National Access Network, "School Funding Litigation Overview," Teachers College, Columbia University, http://schoolfunding.info/.

[4] U.S. Department of Education, "Approved Evidence-Based, Whole-School Reform Models," U.S. Department of Education, www.ed.gov.

Chapter 11

[1] Almlund et al., "Personality Psychology and Economics."

[2] Ibid., 89.

[3] Ibid. Table 1.8.

[4] See Clancy Blair and Rachel Peters Razza, "Relating Effortful Control, Executive Function, and False Belief Understanding to Emerging Math and Literacy Ability in Kindergarten," *Child Development* 78, no. 2 (2007).

[5] Megan M. McClelland et al., "Links between Behavioral Regulation and Preschoolers' Literacy, Vocabulary, and Math Skills," *Developmental Psychology* 43, no. 4 (2007).

[6] Walter Mischel, Yuichi Shoda, and Monica L. Rodriguez, "Delay of Gratification in Children," *Science* 244, no. 4907 (1989).

[7] Greg J. Duncan et al., "School Readiness and Later Achievement," *Developmental Psychology* 43, no. 6 (2007).

[8] Ibid., 1443.

[9] As noted by Heckman and his co-authors, the Perry preschool program did not generate lasting effects on student achievement, consistent with the thesis of this book that the effects of grading and testing practices over the K-12 years swamp the influences of personality traits acquired at an early age prior to entry into school. Barnett and Masse, however, note that the Abecedarian preschool program generated lasting effects on student achievement. See W.S. Barnett and Leonard N. Masse, "Comparative Benefit-Cost Analysis of the Abecedarian Program and Its Policy Implications," *Economics of Education Review* 26(2007). Regardless, the cost-effectiveness results presented in Chapter Ten indicate that the rapid assessment intervention is dramatically more cost-effective than either the Perry preschool program or the Abecedarian preschool program when the outcome of interest is student achievement.

Much has been made of the finding that both the Perry preschool program and the Abecedarian preschool program generated benefits that extend beyond student achievement. Barnett and Masse indicate that the vast majority of the benefits of the Perry preschool program were obtained from reductions in the value of crime committed by treatment group members. Barnett and Masse indicate that the Abecedarian preschool program failed to reduce crime committed by treatment group members. Instead, most of the benefits from the Abecedarian program were obtained because the program provided child care, released mothers of participants from child care duties, and enabled them to secure employment and earn more money. While the Perry and Abecedarian programs demonstrated benefits that extend beyond student achievement, it is not clear that these programs are efficient strategies for addressing either the achievement gap or inequalities in participant earnings. A benefit-cost evaluation of the rapid assessment program suggests that this program would have much larger effects on participant earnings and the benefits of the program would greatly exceed costs. It is estimated that benefits would exceed costs by a ratio of 28 to 1. See S. S. Yeh, "Shifting the Bell Curve: The Benefits and Costs of Raising Student Achievement," *Evaluation and Program Planning* 32, no. 1 (2009). In contrast, Barnett and Masse estimated that the Perry preschool program generated benefits exceeding costs by a ratio of 9 to 1, while the Abecedarian preschool program generated benefits exceeding costs by a ratio of 2.5 to 1. This comparison suggests that neither the Perry preschool program nor the Abecedarian preschool program is an efficient strategy for raising earnings.

Appendix D

[1] Heather Woltmann et al., "An Introduction to Hierarchical Linear Modeling," *Tutorials in quantitative methods for psychology* 8, no. 1 (2012).

[2] S. W. Raudenbush, "Educational Applications of Hierarchical Linear Models: A Review," *Journal of Educational and Behavioral Statistics* 13, no. 2 (1988).

[3] Ibid.

[4] David Afshartous and Jan de Leeuw, "Prediction in Multilevel Models," *Journal of Educational and Behavioral Statistics* 30, no. 2 (2005): 119; S. W. Raudenbush, "Many Small Groups," in *Handbook of Multilevel Analysis*, ed. Jan De Leeuw and Erik Meijer (New York: Springer, 2008), 213, 229, 230, 231, 234; Frank M. T. A. Busing, *Distribution Characteristics of Variance Estimates in Two-Level Models: A Monte Carlo Study* (Leiden, The Netherlands: University of Leiden, 1993).

Bibliography

Aaronson, D., L. Barrow, and W. Sander. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics* 25, no. 1 (2007): 95–135.

Adelman, Clifford. "Answers in the Tool Box: Academic Intensity, Attendance Patterns, and Bachelor's Degree Attainment." Washington, DC: U.S. Department of Education, 1999.

Afshartous, David, and Jan de Leeuw. "Prediction in Multilevel Models." *Journal of Educational and Behavioral Statistics* 30, no. 2 (2005): 109–139.

Ainsworth-Darnell, James W., and Douglas B. Downey. "Assessing the Oppositional Culture Explanation for Racial/Ethnic Differences in School Performance." *American Sociological Review* 63, no. 4 (1998): 536–553.

Allington, R., S. Guice, K. Baker, N. Michaelson, and S. Li. "Access to Books: Variations in Schools and Classrooms." *The Language and Literacy Spectrum* 5 (1995): 23–25.

Almlund, Mathilde, Angela Lee Duckworth, James J. Heckman, and Tim Kautz. "Personality Psychology and Economics." Chap. 1 In *Handbook of the Economics of Education*, edited by E. A. Hanushek, S. J. Machin and L. Woessmann, 1–182. Amsterdam: Elsevier, 2011.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association, 1985.

———. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association, 1999.

Anderman, E. M., and M. L. Maehr. "Motivation and Schooling in the Middle Grades." *Review of Educational Research* 64 (1994): 287–309.

Anderson, A. B., and S. J. Stokes. "Social and Institutional Influences on the Development and Practice of Literacy." In *Awakening to Literacy*, edited by H. Goelman, A. Oberg and F. Smith, 24–37. Portsmouth, NH: Heinemann, 1984.

Angrist, Joshua D., Susan M. Dynarski, Thomas J. Kane, Parag A. Pathak, and Christopher R. Walters. "Who Benefits from KIPP?" *Journal of Policy Analysis and Management* 31, no. 4 (2012): 837–860.

Astin, Alexander W., and Leticia Oseguera. *Degree Attainment Rates at American Colleges and Universities.* Los Angeles, CA: Higher Education Research Institute, Graduate School of Education, University of California, Los Angeles, 2005.

S.S. Yeh, *Solving the Achievement Gap*,
DOI 10.1057/978-1-137-58767-1

Associated Press. "Superintendent: Bad Teachers Hard to Fire: Some Say Teacher Tenure Rules Need to Be Overhauled to Address Problem." Accessed July 7, 2016. http://www.nbcnews.com/id/25430476/ns/us_news-education/t/superintendent-bad-teachers-hard-fire/#.V362f-srLIU.

Avery, C., and C. M. Hoxby. "Do and Should Financial Aid Packages Affect Students' College Choices?". In *College Choices: The Economics of Where to Go, When to Go, and How to Pay for It*, edited by C. M. Hoxby, 239–301. Chicago: University of Chicago Press, 2004.

Ballou, Dale. "Value-Added Assessment: Lessons from Tennessee." In *Value Added Models in Education: Theory and Applications*, edited by R. Lissetz, 1–26. Maple Grove, MN: JAM Press, 2005.

Ballou, Dale, W. Sanders, and P. Wright. "Controlling for Student Background in Value-Added Assessment of Teachers." *Journal of Behavioral and Educational Statistics* 29, no. 1 (2004): 37–65.

Bandura, A. "Self-Efficacy Mechanism in Human Agency." *American Psychologist* 37 (1982): 122–147.

———. "Self-Efficacy: Toward a Unifying Theory of Behavior Change." *Psychological Review* 84 (1977): 191–215.

Bangert-Drowns, R. L., C. C. Kulik, J. A. Kulik, and M. Morgan. "The Instructional Effect of Feedback in Test-Like Events." *Review of Educational Research* 61, no. 2 (1991): 213–238.

Barnett, W.S., and Leonard N. Masse. "Comparative Benefit-Cost Analysis of the Abecedarian Program and Its Policy Implications." *Economics of Education Review* 26 (2007): 113–125.

Baron, R., D. Tom, and H. Cooper. "Social Class, Race, and Teacher Expectations." In *Teacher Expectancies*, edited by Jerome Dusek, 251–269. Hillsdale, NJ: Lawrence Erlbaum, 1985.

Bayer, Patrick, and Robert McMillan. "Tiebout Sorting and Neighborhood Stratification." *Journal of Public Economics* 96, no. 11–12 (2012): 1129–1143.

Berston, H.M. "The School Dropout Problem." *The Clearing House* 35, no. 4 (1960): 207–210.

Blacher, Michael. "K-12 Teacher Termination Hearings: Are They Worth the Cost?" *CPER Journal* no. 180 (October, 2006): 13–19.

Black, P., and D. Wiliam. "Assessment and Classroom Learning." *Assessment in Education* 5, no. 1 (1998): 7–74.

Blair, Clancy, and Rachel Peters Razza. "Relating Effortful Control, Executive Function, and False Belief Understanding to Emerging Math and Literacy Ability in Kindergarten." *Child Development* 78, no. 2 (2007): 647–663.

Blankenship, V. "Individual Differences in Resultant Achievement Motivation and Latency to and Persistence at an Achievement Task." *Motivation and Emotion* 16, no. 1 (1992): 35–63.

Bond, L., T. Smith, W. K. Baker, and J. A. Hattie. "The Certification System of the National Board for Professional Teaching Standards: A Construct and Consequential Validity Study." Greensboro, NC: Center for Educational Research and Evaluation, The University of North Carolina at Greensboro, 2000.

Bowen, William G., and Derek Bok. *The Shape of the River: Long-Term Consequences of Considering Race in College and University Admissions.* Princeton, NJ: Princeton University Press, 1998.

Bowen, William G., Matthew Chingos, and Michael S. McPherson. *Crossing the Finish Line: Completing College at America's Public Universities*. Princeton, NJ: Princeton University Press, 2009.

Bowen, William G., Martin A. Kurzweil, and Eugene M. Tobin. *Equity and Excellence in American Higher Education*. Charlottesville, VA: University of Virginia Press, 2005.

Bowers, Alex J. "Grades and Graduation: A Longitudinal Risk Perspective to Identify Student Dropouts." *The Journal of Educational Research* 103, no. 3 (2010): 191–207.

Bowles, Samuel. "A (Science-Based) Poor Kids' Manifesto." *Science* 340 (2013): 1044–1045.

Brantlinger, Ellen. "Low-Income Adolescents' Perceptions of School, Intelligence, and Themselves as Students." *Curriculum Inquiry* 20, no. 3 (1990): 305–324.

Bratton, S. E., Jr., S. P. Horn, and S. P. Wright. "Using and Interpreting Tennessee's Value-Added Assessment System: A Primer for Teachers and Principals." Accessed November 13, 2015. http://www.shearonforschools.com/documents/TVAAS.HTML#Let's.

Braun, Henry, Naomi Chudowsky, and Judith Koenig, eds. *Getting Value out of Value-Added: Report of a Workshop*. Washington, DC: The National Academies Press, 2010.

Bridgeland, John M., John J. DiIulio, Jr., and Karen B. Morison. "The Silent Epidemic: Perspectives of High School Dropouts." Washington, DC: Civic Enterprises, 2006.

Bridgeman, Brent, Judy Pollack, and Nancy Burton. "Understanding What SAT Reasoning Test Scores Add to High School Grades: A Straightforward Approach." Research Report No. 2004-4. New York: College Entrance Examination Board, 2004.

Briggs, Derek, and Ben Domingue. "A Review of the Value-Added Analysis Underlying the Effectiveness Rankings of Los Angeles Unified School District Teachers by the Los Angeles Times." Boulder, CO: University of Colorado at Boulder, 2011.

Brookover, W. B., C. H. Beady, P. K. Flood, J. G. Schweitzer, and J. M. Wisenbaker. *Schools, Social Systems and Student Achievement: Schools Can Make a Difference*. New York: Praeger, 1979.

Brookover, W. B., J. G. Schweitzer, J. M. Schneider, C. H. Beady, P. K. Flood, and J. M. Wisenbaker. "Elementary School Social Climate and School Achievement." *American Educational Research Journal* 15 (1978): 301–318.

Bureau of Labor Statistics. "Consumer Price Index." U.S. Department of Labor. Accessed November 13, 2015. http://www.bls.gov/cpi/cpid1509.pdf.

Burton, Nancy W., and Leonard Ramist. "Predicting Success in College: SAT Studies of Classes Graduating since 1980." College Board Research Report No. 2001-2. New York: College Entrance Examination Board, 2001.

Bus, A. G., M. H. van IJzendoorn, and A. D. Pellegrini. "Joint Book Reading Makes for Success in Learning to Read: A Meta-Analysis on Intergenerational Transmission of Literacy." *Review of Educational Research* 65, no. 1 (1995): 1–21.

Busing, Frank M. T. A. *Distribution Characteristics of Variance Estimates in Two-Level Models: A Monte Carlo Study*. Leiden, The Netherlands: University of Leiden, 1993.

Camara, Wayne J., and Gary Echternacht. "The SAT I and High School Grades: Utility in Predicting Success in College." Research Note No. RN-10. New York: The College Entrance Examination Board, 2000.

Campbell, D. T., and J. C. Stanley. "Experimental and Quasi-Experimental Designs for Research on Teaching." In *Handbook of Research on Teaching*, edited by N. L. Gage, 171–246. Chicago: Rand McNally, 1963.

Carneiro, Pedro, and James J. Heckman. “Human Capital Policy.” In *Inequality in America: What Role for Human Capital Policies?*, edited by James J. Heckman and Alan Krueger, 77–239. Cambridge, MA: MIT Press, 2003.

Carrell, Scott E., and James E. West. “Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors.” *Journal of Political Economy* 118, no. 3 (2010): 409–432.

Carter, Samuel Casey. “No Excuses: Seven Principals of Low-Income Schools Who Set the Standard for High Achievement.” Washington, DC: Heritage Foundation, 1999.

Cavalluzzo, L. “Is National Board Certification an Effective Signal of Teacher Quality?” The CNA Corporation. Accessed November 13, 2015. http://www.nbpts.org/sites/default/files/documents/research/Cavalluzzo_IsNBCAnEffectiveSignalofTeachingQuality.pdf.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. “The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood.” Working Paper 17699 Cambridge, MA: National Bureau of Economic Research, 2011.

———. “Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates.” *American Economic Review* 104, no. 9 (2014): 2593–2632.

———. “Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood.” *American Economic Review* 104, no. 9 (2014): 2633–2679.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. “Teacher-Student Matching and the Assessment of Teacher Effectiveness.” *Journal of Human Resources* 41, no. 4 (2006): 778–820.

———. “Teacher Credentials and Student Achievement: Longitudinal Analysis with Student Fixed Effects.” *Economics of Education Review* 26, no. 6 (2007): 673–682.

Coleman, J. S., E. Campbell, C. Hobson, J. McPartland, A. Mood, R. Weinfeld, and R. York. “Equality of Educational Opportunity.” Washington, DC: Government Printing Office, 1966.

College Board. “Net Prices by Income over Time: Public Sector.” Accessed November 13,2015.http://trends.collegeboard.org/college-pricing/figures-tables/net-prices-income-over-time-public-sector.

Collins, Gail. “Waiting for Somebody.” *The New York Times*, September 30, 2010, A35.

Congressional Budget Office. “An Analysis of the Administration’s Proposed Program for Displaced Workers.” CBO Papers. Washington, D.C.: Congressional Budget Office, 1994.

Cook, Phillip J., and Jens Ludwig. “The ‘Burden of “Acting White”’: Do Black Adolescents Disparage Academic Achievement?”. In *The Black-White Test Score Gap*, edited by C. Jencks and M. Phillips, 375-400. Washington, DC: The Brookings Institution, 1998.

———. “Weighing the ‘Burden of “Acting White”’: Are There Race Differences in Attitudes toward Education?” *Journal of Policy Analysis and Management* 16, no. 2 (1997): 256–278.

Cooper, Anderson. “Harlem’s Education Experiment Gone Right.” CBS News. Accessed November 13, 2015. http://www.cbsnews.com/stories/2009/12/04/60minutes/main5889558_page2.shtml?tag=contentMain;contentBody.

Cooper, H., K. Charlton, J. C. Valentine, and L. Muhlenbruck. “Making the Most of Summer School: A Meta-Analytic and Narrative Review.” *Monographs of the Society for Research in Child Development* 65, no. 1 (2000).

Crandall, V., W. Katkovky, and U. Crandall. "Children's Beliefs in Their Own Control of Reinforcements in Intellectual Academic Achievement Situations." *Child Development* 36 (1965): 91–109.

Crooks, T. J. "The Impact of Classroom Evaluation Practices on Students." *Review of Educational Research* 58 (1988): 438–481.

Curto, Vilsa E. , Roland G. Fryer, Jr., and Meghan L. Howard. "It May Not Take a Village: Increasing Achievement among the Poor." In *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances*, edited by Greg J. Duncan and Richard J. Murnane, 483–505. New York: Russell Sage Foundation, 2011.

Danner, F. W., and D. Lonsky. "A Cognitive-Developmental Approach to the Effects of Rewards on Intrinsic Motivation." *Child Development* 52 (1981): 1043–1052.

Darity, William. "Stratification Economics: The Role of Intergroup Inequality." *Journal of Economics and Finance* 29, no. 2 (2005): 144–153.

De Jong, Rob, and Klaas J. Westerhof. "The Quality of Student Ratings of Teacher Behavior." *Learning Environments Research* 4, no. 1 (2001): 51–85.

Deresiewicz, William. *Excellent Sheep: The Miseducation of the American Elite and the Way to a Meaningful Life*. New York: Free Press, 2014.

Di Loreto, C., and L. Tse. "Seeing Is Believing: Disparity in Books in Two Los Angeles Area Public Libraries." *School Library Quarterly* 17, no. 3 (1999): 31–36.

Diener, C., and C. Dweck. "Analysis of Learned Helplessness: Continuous Changes in Performance, Strategy, and Achievement Cognitions Following Failure." *Journal of Personality and Social Psychology* 36 (1978): 461–482.

Dillon, Sam. "Formula to Grade Teachers' Skill Gains Acceptance, and Critics." *The New York Times*, September 1, 2010, A1, A3.

Dobbie, Will, and Roland G. Fryer, Jr. "Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone." *American Economic Journal: Applied Economics* 3, no. 3 (2011): 158–187.

Dougherty, M. R., and J. I. Harbison. "Motivated to Retrieve: How Often Are You Willing to Go Back to the Well When the Well Is Dry?" *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33, no. 6 (2007): 1108–1117.

Downey, Douglas B., and James W. Ainsworth-Darnell. "The Search for Oppositional Culture among Black Students." *American Sociological Review* 67, no. 1 (2002): 156–164.

Drucker, P.M., D. B. Drucker, T. Litto, and R. Stevens. "Relation of Task Difficulty to Persistence." *Perceptual and Motor Skills* 86 (1998): 787–794.

Duncan, Greg J., Chantelle J. Dowsett, Amy Claessens, Mugnuson Katherine, Aletha C. Huston, Pamela Klebanov, Linda S. Pagani, *et al.* "School Readiness and Later Achievement." *Developmental Psychology* 43, no. 6 (2007): 1428–1446.

Duncan, Otis D., David L. Featherman, and Beverly Duncan. *Socioeconomic Background and Achievement*. New York: Seminar Press, 1972.

Dusek, J. B. "Do Teachers Bias Children's Learning?" *Review of Educational Research* 45 (1975): 661–684.

Dweck, C. S., and E. S. Elliott. "Achievement Motivation." In *Handbook of Child Psychology, Vol. 4: Socialization, Personality, and Social Development*, edited by E. M. Hetherington and P. H. Mussen, 643–691. New York: Wiley, 1983.

Dweck, C. S., and N. Dickon Reppucci. "Learned Helplessness and Reinforcement Responsibility in Children." *Journal of Personality and Social Psychology* 25, no. 1 (1973): 109–116.

Dweck, Carol. "Motivational Processes Affecting Learning." *American Psychologist* 41 (1986): 1040–1048.

Eccles, J. S., and C. Midgley. "Stage-Environment Fit: Developmentally Appropriate Classrooms for Early Adolescents." In *Research on Motivation in Education, Vol. 3: Goals and Cognitions*, edited by R. E. Ames and C. Ames, 139–186. New York: Academic, 1989.

Eccles, J. S., C. Midgley, A. Wigfield, C. M. Buchanan, D. Reuman, C. Flanagan, and D. MacIver. "Development During Adolescence: The Impact of Stage-Environment Fit on Adolescents' Experiences in Schools and Families." *American Psychologist* 48 (1993): 90–101.

Eccles, J. S., A. Wigfield, and U. Schiefele. "Motivation to Succeed." In *Handbook of Child Psychology* edited by N. Eisenberg and W. Damon, 1017–1095. New York: John Wiley and Sons, 1998.

Educational Testing Service. "Where We Stand on Teacher Quality." Accessed November 13, 2015. https://www.ets.org/Media/Education_Topics/pdf/teacherquality.pdf.

Entwisle, D. R., and K. Alexander. "Summer Setback: Race, Poverty, School Composition, and Mathematics Achievement in the First Two Years of School." *American Sociological Review* 57 (1992): 72–84.

———. "Winter Setback: The Racial Composition of Schools and Learning to Read." *American Sociological Review* 59 (1994): 446–460.

Farkas, George, Christy Lleras, and Steve Maczuga. "Does Oppositional Culture Exist in Minority and Poverty Peer Groups?" *American Sociological Review* 67, no. 1 (2002): 148–155.

Feitelson, D., and Z. Goldstein. "Patterns of Book Ownership and Reading to Young Children in Israeli School-Oriented and Nonschool-Oriented Families." *The Reading Teacher* 39 (1986): 924–930.

Ferguson, R. F. "Teachers' Perceptions and Expectations and the Black-White Test Score Gap." In *The Black-White Test Score Gap*, edited by C. Jencks and M. Phillips, 273–317. Washington, DC: Brookings Institute, 1998.

Findley, M. J., and H. M. Cooper. "Locus of Control and Academic Achievement: A Literature Review." *Journal of Personality and Social Psychology* 44, no. 2 (1983): 419–427.

Finn, J.D., S.B. Gerber, C.M. Achilles, and J. Boyd-Zaharias. "The Enduring Effects of Small Classes." *Teachers College Record* 103, no. 2 (2001): 145–183.

Fordham, Signithia, and John Ogbu. "Black Students' School Success: Coping with the 'Burden of "Acting White"'." *Urban Review* 18 (1986): 176–206.

Fryer, Roland G., Jr., and Steven D. Levitt. "The Black-White Test Score Gap through Third Grade." *American Law and Economics Review* 8, no. 2 (2006): 249–281.

———. "Understanding the Black-White Test Score Gap in the First Two Years of School." *The Review of Economics and Statistics* 86, no. 2 (2004): 447–464.

Fuchs, L. S., and D. Fuchs. "Effects of Systematic Formative Evaluation: A Meta-Analysis." *Exceptional Children* 53, no. 3 (1986): 199–208.

Fyans, L. J. "Test Anxiety, Test Comfort, and Student Achievement Test Performance." Princeton, NJ: Educational Testing Service, 1979.

Gallup. "Gallup Student Poll." Accessed November 13, 2015. http://www.gallup.com/services/180029/gallup-student-poll-2014-overall-report.aspx.

Geiser, Saul, and Maria Veronica Santelices. "Validity of High-School Grades in Predicting Student Success Beyond the Freshman Year: High-School Record vs.

Standardized Tests as Indicators of Four-Year College Outcomes." Research and Occasional Paper Series No. CSHE.6.07. University of California, Berkeley, 2007.

Geiser, Saul, and Roger Studley. "UC and the SAT: Predictive Validity and Differential Impact of the SAT I and SAT II at the University of California." *Educational Assessment* 8, no. 1 (2002): 1–26.

Gitomer, Drew. "Reliability and NBPTS Assessments." In *Assessing Teachers for Professional Certification: The First Decade of the National Board for Professional Teaching Standards*, edited by Lawrence Ingvarson and John Hattie. Advances in Program Evaluation, 231–253. Amsterdam: Elsevier, 2008.

Glazerman, S., D. Goldhaber, S. Loeb, D. O. Staiger, and G. J. Whitehurst. "America's Teacher Corps." Washington, DC: Brown Center on Education Policy at Brookings, 2010.

Glazerman, S., S. Loeb, D. Goldhaber, D. O. Staiger, S. W. Raudenbush, and G. J. Whitehurst. "Evaluating Teachers: The Important Role of Value-Added." Washington, DC: Brookings, 2010.

Goldhaber, Dan, and Emily Anthony. "Can Teacher Quality Be Effectively Assessed? National Board Certification as a Signal of Effective Teaching." *Review of Economics and Statistics* 89, no. 1 (2007): 134–150.

Goldhaber, Dan, and Michael Hansen. "Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions." Brief No. 3. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research, 2008.

———. "Is It Just a Bad Class? Assessing the Long-Term Stability of Estimated Teacher Performance." *Economica* 80, no. 319 (2013): 589–612.

———. "Is It Just a Bad Class? Assessing the Long-Term Stability of Estimated Teacher Performance." Working Paper No. 73. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research, 2012.

Goldhaber, Dan, David Perry, and Emily Anthony. "NBPTS Certification: Who Applies and What Factors Are Associated with Success?" Seattle, WA: University of Washington, 2003.

Goodlad, J. I. *A Place Called School*. New York: McGraw Hill, 1984.

Gordon, Robert , Thomas J. Kane, and Douglas O. Staiger. "Identifying Effective Teachers Using Performance on the Job." Discussion Paper No. 2006-01. Washington, D.C.: The Brookings Institution, 2006.

Gottschalck, Alfred O. "Dynamics of Economic Well-Being: Spells of Unemployment 2001–2003." Report No. P70-105. Washington, D.C.: U.S. Census Bureau, 2006.

Gould, Patricia. "3020-a Process Remains Slow, Costly." New York State School Boards Association. Accessed July 7, 2016. http://www.nyssba.org/news/2009/05/11/on-board-online-may-11-2009/3020-a-process-remains-slow-costly/.

Greenwald, Rob, Larry V. Hedges, and Richard D. Laine. "The Effect of School Resources on Student Achievement." *Review of Educational Research* 66, no. 3 (1996): 361-396.

Grissmer, D. W., A. Flanagan, J. Kawata, and S. Williamson. *Improving Student Achievement: What State NAEP Test Scores Tell Us*. Santa Monica, CA: RAND Corporation, 2000.

Haertel, Edward H. "Letter Report to the U.S. Department of Education on the Race to the Top Fund." The National Academies. Accessed July 8, 2016. https://download.nap.edu/catalog.php?record_id=12780.

Hanushek, E. A. “The Difference Is Great Teachers.” In *Waiting for ‘Superman’: How We Can Save America’s Failing Public Schools*, edited by Karl Weber, 81–100. New York: Public Affairs, 2010.

———. “Teacher Deselection.” In *Creating a New Teaching Profession*, edited by Dan Goldhaber and Jane Hannaway, 165–180. Washington, DC: Urban Institute Press, 2009.

———. “The Trade-Off between Child Quantity and Quality.” *Journal of Political Economy* 100, no. 1 (1992): 84–117.

Hanushek, E. A., J. F. Kain, D. M. O’Brien, and S. G. Rivkin. “The Market for Teacher Quality.” Working Paper 11154 Cambridge, MA: National Bureau of Economic Research, 2005.

Harris, D. N., and T. R. Sass. “The Effects of NBPTS-Certified Teachers on Student Achievement.” Madison, WI: University of Wisconsin, 2007.

Harris, M. M., and N. J. Smith. “Literacy Assessment of Chapter 1 and Non-Chapter 1 Homes.” *Reading Improvement* 24 (1987): 137–142.

Hart, Betty, and Todd R. Risley. *Meaningful Differences in the Everyday Experience of Young American Children*. Baltimore, MD: Paul H. Brookes Publishing Co., 1995.

Harter, S. “A Model of Mastery Motivation in Children: Individual Differences and Developmental Change.” In *The Minnesota Symposia on Child Psychology (Vol. 14)*, edited by W. A. Collins, 215–255. Hillsdale, NJ: Erlbaum, 1981.

———. “A New Self-Report Scale of Intrinsic Versus Extrinsic Orientation in the Classroom: Motivational and Informational Components.” *Developmental Psychology* 17 (1981): 300–312.

———. “Pleasure Derived from Optimal Challenge and the Effects of Extrinsic Rewards on Children’s Difficulty Level Choices.” *Child Development* 49 (1978): 788–799.

Hattie, J., and H. Timperley. “The Power of Feedback.” *Review of Educational Research* 77, no. 1 (2007): 81–112.

Heath, S. B. *Ways with Words: Language, Life, and Work in Communities and Classrooms.* Cambridge: Cambridge University Press, 1983.

———. “What No Bedtime Story Means: Narrative Skills at Home and School.” *Language in Society* 11 (1982): 49–76.

Heckman, James J. *Giving Kids a Fair Chance.* Cambridge, MA: The MIT Press, 2013.

Heckman, James J., and Tim D. Kautz. “Hard Evidence on Soft Skills.” *Labour Economics* 19, no. 4 (2012): 451–464.

Henig, Jeffrey R. “What Do We Know About the Outcomes of KIPP Schools?” Tempe, AZ: Arizona State University, 2008.

Herrnstein, R. J., and C. Murray. *The Bell Curve: Intelligence and Class Structure in American Life.* New York: Simon and Schuster, 1994.

Hess, Frederick M. “Beyond School Choice.” *National Review*, October 18, 2010, 41–42.

Heyns, B. *Summer Learning and the Effects of Schooling.* New York: Academic Press, 1978.

Hill, H. C., L. Kapitula, and K. Umland. “A Validity Argument Approach to Evaluating Teacher Value-Added Scores.” *American Educational Research Journal* 48, no. 3 (2011): 794–831.

Hill, Kennedy T. “Eliminating Motivational Testing Error by Developing Optimal Testing Procedures and Teaching Test-Taking Skills.” Princeton, NJ: Educational Testing Service, 1979.

———. “The Relation of Evaluative Practices to Test Anxiety and Achievement Motivation.” *UCLA Educator* 19 (1977): 15–21.

Hill, Kennedy T., and Seymour B. Sarason. "The Relation of Test Anxiety and Defensiveness to Test and School Performance over the Elementary School Years: A Further Longitudinal Study." *Monographs of the Society for Research in Child Development* 31, no. 2 (1966): 1–76.

Hill, Kennedy T., and Allan Wigfield. "Test Anxiety: A Major Educational Problem and What Can Be Done About It." *Elementary School Journal* 85, no. 1 (1984): 105–126.

Hiroto, D.S., and M.E.P. Seligman. "Generality of Learned Helplessness in Man." *Journal of Personality and Social Psychology* 31 (1975): 311–327.

Hoff-Ginsberg, Erika. "Mother-Child Conversation in Different Social Classes and Communicative Settings." *Child Development* 62, no. 4 (1991): 782–796.

Hoxby, C. M. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *The Quarterly Journal of Economics* 115, no. 4 (November 2000): 1239–1285.

Committee on Governmental Affairs. *Testimony by Caroline M. Hoxby, "the Rising Cost of College Tuition and the Effectiveness of Government Financial Aid"*, 106th Congress, 2d, February 9, 2000.

Hymel, S. "The Relationship between Motivation Factors and Report Card Evaluation." Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, CA, April, 1981.

Ingersoll, R. M. "Teacher Turnover and Teacher Shortages: An Organizational Analysis." *American Education Research Journal* 38, no. 3 (2001): 499–534.

Ishii, Jun, and Steven G. Rivkin. "Impediments to the Estimation of Teacher Value Added." *Education Finance and Policy* 4, no. 4 (2009): 520–536.

Jacob, Brian A., and Lars Lefgren. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education." *Journal of Labor Economics* 26, no. 1 (2008): 101-136.

Jacob, Brian A., Lars Lefgren, and David P. Sims. "The Persistence of Teacher-Induced Learning Gains." *Journal of Human Resources* 45, no. 4 (2010): 915–943.

Janosz, Michel, Marc LeBlanc, Bernard Boulerice, and Richard E. Tremblay. "Disentangling the Weight of School Dropout Predictors: A Test on Two Longitudinal Samples." *Journal of Youth and Adolescence* 26, no. 6 (1997).

Jencks, Christopher, and others. *Who Gets Ahead? The Determinants of Economic Success in America*. Basic Books, 1979.

Jencks, Christopher, and Meredith Phillips, eds. *The Black-White Test Score Gap*. Washington, DC: Brookings, 1998.

Johnson, Matthew, Stephen Lipscomb, Brian Gill, Kevin Booker, and Julie Bruch. "Value-Added Models for the Pittsburgh Public Schools." Cambridge, MA: Mathematica Policy Research, 2012.

Kalechstein, A. D., and S. Nowicki, Jr. "A Meta-Analytic Examination of the Relationship between Control Expectancies and Academic Achievement: An 11-Yr Follow-up to Findley and Cooper." *Genetic, Social, and General Psychology Monographs* 123, no. 1 (1997): 27–56.

Kane, T.J. *The Price of Admission: Rethinking How Americans Pay for College*. Washington, D.C.: Brookings Institution Press, 1999.

Kane, Thomas J., and Douglas O. Staiger. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Working Paper No. 14607. Cambridge, MA: National Bureau of Economic Research, 2008.

Kanevsky, L., and T. Keighley. "To Produce or Not to Produce? Understanding Boredom and the Honor in Underachievement." *Roeper Review* 26, no. 1 (2003): 20–28.

Kautz, Tim, James J. Heckman, Ron Diris, Bas ter Weel, and Lex Borghans. "Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success." Paris: OECD, 2014.

Keith, T. Z., S. M. Pottebaum, and S. Eberhart. "Effects of Self Concept and Locus of Control on Academic Achievement: A Large Sample Path Analysis." *Journal of Psychoeducational Assessment* 4, no. 1 (1986): 61–72.

Kennelly, K. J., D. Dietz, and P. Benson. "Reinforcement Schedules, Effort vs. Ability Attributions, and Persistence." *Psychology in the Schools* 22, no. 4 (1985): 459–464.

Klein, D.C., E. Fencil-Morse, and M.E.P. Seligman. "Learned Helplessness, Depression, and the Attribution of Failure." *Journal of Personality and Social Psychology* 33 (1976): 508–516.

Klem, Adena M., and James P. Connell. "Relationships Matter: Linking Teacher Support to Student Engagement and Achievement." *Journal of School Health* 74, no. 7 (2004): 262–273.

Kluger, A. N., and A. DeNisi. "The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory." *Psychological Bulletin* 119, no. 2 (1996): 254–284.

Koedel, Cory, and Julian R. Betts. "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique." *Education Finance and Policy* 6, no. 1 (2011): 18–42.

Koedel, Cory, and Julian R. Betts. "Re-Examining the Role of Teacher Quality in the Educational Production Function." Working Paper No. 2007-03. Columbia, MO: University of Missouri, 2007.

Konstantopoulos, Spyros. "Teacher Effects in Early Grades: Evidence from a Randomized Study." *Teachers College Record* 113, no. 7 (2011): 1541–1565.

Kristof, Nicholas D. "The Value of Teachers." The New York Times. Accessed July 8, 2016. http://www.nytimes.com/2012/01/12/opinion/kristof-the-value-of-teachers.html.

———. "The Value of Teachers." *The International Herald Tribune*, January 13, 2012, 7.

Kupermintz, H. "Teacher Effects and Teacher Effectiveness: A Validity Investigation of the Tennessee Value Added Assessment System." *Educational Evaluation and Policy Analysis* 25, no. 3 (Fall 2003): 287–298.

Ladd, H. F., T. R. Sass, and D. N. Harris. "The Impact of National Board Certified Teachers on Student Achievement in Florida and North Carolina: A Summary of the Evidence Prepared for the National Academies Committee on the Evaluation of the Impact of Teacher Certification by NBPTS." Washington DC: The National Academies, 2007.

Lefgren, L., and D. Sims. "Using Subject Test Scores Efficiently to Predict Teacher Value-Added." *Educational Evaluation and Policy Analysis* 34, no. 1 (2012): 109–121.

Leonhardt, David. "Measuring Colleges' Success in Enrolling the Less Affluent." *The New York Times*, September 9, 2014, A3.

Levin, H. M., G.V. Glass, and G. Meister. "A Cost-Effectiveness Analysis of Computer-Assisted Instruction." *Evaluation Review* 11, no. 1 (1987): 50–72.

Levin, Henry M. "Cost-Effectiveness and Educational Policy." *Educational Evaluation and Policy Analysis* 10, no. 1 (1988): 51–69.

Levin, J. "For Whom the Reductions Count: A Quantile Regression Analysis of Class Size and Peer Effects on Scholastic Achievement." *Empirical Economics* 26 (2001): 221–246.

Levine, Madeline. *The Price of Privilege: How Parental Pressure and Material Advantage Are Creating a Generation of Disconnected and Unhappy Kids.* New York: HarperCollins, 2006.

Lightfoot, S. *Worlds Apart: Relationships between Families and Schools.* New York: Basic Books, 1978.

Lindsay, Jim. "Children's Access to Print Material and Education-Related Outcomes: Findings from a Meta-Analytic Review." Napierville, IL: Learning Point Associates, 2010.

LoGerfo, Laura, Austin Nichols, and Sean Reardon. "Achievement Gains in Elementary and High School." Washington, DC: Urban Institute, 2006.

Louisiana Department of Education. "ESEA Flexibility Renewal Form: Louisiana." Board of Elementary and Secondary Education. Baton Rouge, LA: Louisiana Department of Education, 2015.

Lowrey, Annie. "Big Study Links Good Teachers to Lasting Gain." *The New York Times*, January 6, 2012, A1.

Lysakowski, R. S., and H. J. Walberg. "Instructional Effects of Cues, Participation, and Corrective Feedback: A Quantitative Synthesis." *American Educational Research Journal* 19 (1982): 559–578.

Mac Iver, D. "Classroom Environments and the Stratification of Pupils' Ability Perceptions." *Journal of Educational Psychology* 80, no. 4 (1988): 495–505.

Mac Iver, D. J., D. A. Reuman, and S. R. Main. "Social Structuring of the School: Studying What Is, Illuminating What Could Be." *Annual Review of Psychology* 46 (1995): 375–400.

Mac Iver, D. J., D. J. Stipek, and D. H. Daniels. "Explaining within-Semester Changes in Student Effort in Junior High School and Senior High School Courses." *Journal of Educational Psychology* 83, no. 2 (1991): 201–211.

Mac Iver, Douglas J., and David A. Reuman. "Giving Their Best: Grading and Recognition Practices That Motivate Students to Work Hard." *American Educator* 17, no. 4 (1993/1994): 24–31.

Manski, C. F. "Academic Ability, Earnings, and the Decision to Become a Teacher: Evidence from the National Longitudinal Study of the High School Class of 1972." In *Public Sector Payrolls*, edited by D. Wise, 291–312. Chicago: University of Chicago Press, 1987.

Manski, Charles F., and David A. Wise. *College Choice in America*. Cambridge, MA: Harvard University Press, 1983.

Marks, Helen M. "Student Engagement in Instructional Activity: Patterns in the Elementary, Middle, and High School Years." *American Educational Research Journal* 37, no. 1 (2000): 153–184.

Martineau, J. A. "Distorting Value-Added: The Use of Longitudinal, Vertically-Scaled Student Achievement Data for Growth-Based, Value-Added Accountability." *Journal of Educational and Behavioral Statistics* 31, no. 1 (2006): 35–62.

Mason, J., and C. McCormick. "An Investigation of Prereading Instruction from a Developmental Perspective: Foundations for Literacy." Technical Report No. 224. Urbana-Champaign, IL: Center for the Study of Reading, 1981.

McCaffrey, D.F., J. R. Lockwood, D. Koretz, T. A. Louis, and L. Hamilton. "Models for Value-Added Modeling of Teacher Effects." *Journal of Behavioral and Educational Statistics* 29, no. 1 (2004): 67–101.

McCaffrey, D.F., Tim R. Sass, J.R. Lockwood, and Kata Mihaly. "The Intertemporal Variability of Teacher Effect Estimates." *Education Finance and Policy* 4, no. 4 (2009): 572–606.

McClelland, Megan M., Claire E. Cameron, Carol McDonald Connor, Carrie L. Farris, Abigail M. Jewkes, and Frederick J. Morrison. "Links between Behavioral Regulation and Preschoolers' Literacy, Vocabulary, and Math Skills." *Developmental Psychology* 43, no. 4 (2007): 947–959.

McColskey, Wendy, James H. Stronge, Thomas J. Ward, Pamela D. Tucker, Barbara Howard, Karla Lewis, and Jennifer L. Hindman. "Teacher Effectiveness, Student Achievement, and National Board Certified Teachers: A Comparison of National Board Certified Teachers and Non-National Board Certified Teachers: Is There a Difference in Teacher Effectiveness and Student Achievement?" National Board for Professional Teaching Standards. Accessed July 8, 2016. http://www.education-consumers.com/articles/W-M%20NBPTS%20certified%20report.pdf.

McDermott, Paul A., Melissa Mordell, and Jill C. Stoltzfus. "The Organization of Student Performance in American Schools: Discipline, Motivation, Verbal Learning, and Nonverbal Learning." *Journal of Educational Psychology* 93, no. 1 (2001): 65–76.

Milanowski, A. "The Relationship between Teacher Performance Evaluation Scores and Student Achievement: Evidence from Cincinnati." *Peabody Journal of Education* 79, no. 4 (2004): 33–53.

Miller, W. H. "Home Prereading Experiences and First-Grade Achievement." *The Reading Teacher* 22 (1969): 641–645.

Mischel, Walter, Yuichi Shoda, and Monica L. Rodriguez. "Delay of Gratification in Children." *Science* 244, no. 4907 (1989): 933–938.

Molden, Daniel C., and Carol S. Dweck. "Meaning and Motivation." In *Intrinsic and Extrinsic Motivation: The Search for Optimal Motivation and Performance*, edited by Carol Sansone and Judith M. Harackiewicz, 131–159. San Diego: Academic Press, 2000.

Murnane, Richard J., Judith D. Singer, and John B. Willett. "The Career Paths of Teachers: Implications for Teacher Supply and Methodological Lessons for Research." *Educational Researcher* 17, no. 6 (1988): 22–30.

Murnane, Richard J., John B. Willett, Kristen L. Bub, and Kathleen McCartney. "Understanding Trends in the Black-White Achievement Gaps During the First Years of School." *Brookings-Wharton Papers on Urban Affairs* (2006): 97–135.

National Access Network. "School Funding Litigation Overview." Teachers College, Columbia University. Accessed July 8, 2016. http://www.schoolfunding.info/litigation/overview.php3.

National Board for Professional Teaching Standards. "About Us." Accessed November 13, 2015. http://www.nbpts.org/who-we-are.

———. "Guide to National Board Certification." Accessed July 8, 2016. http://boardcertifiedteachers.org/sites/default/files/Guide_to_NB_Certification.pdf.

National Bureau of Economic Research. "Moving to Opportunity (MTO) for Fair Housing Demonstration Program." Accessed July 8, 2016. http://www.nber.org/mtopublic/.

National Center for Education Statistics. "Digest of Education Statistics." U.S. Department of Education. Accessed July 8, 2016. http://nces.ed.gov/programs/digest/d05/tables/dt05_076.asp.

———. "Early Childhood Longitudinal Program (ECLS): Kindergarten Class of 1998-99 (ECLS-K)." Institute of Education Sciences. Accessed July 8, 2016. http://nces.ed.gov/ecls/kindergarten.asp.

———. "Early Childhood Longitudinal Program (ECLS): Kindergarten Class of 2010–11 (ECLS-K:2011)." Accessed July 8, 2016. https://nces.ed.gov/ecls/kindergarten2011.asp.

———. “Education Longitudinal Study of 2002.” Institute of Education Sciences. Accessed July 8, 2016. http://nces.ed.gov/surveys/els2002/.

———. “NAEP 2008 Trends in Academic Progress.” NCES Report No. 2009–479. Washington, DC: U.S. Department of Education, 2008.

———. “National Education Longitudinal Study of 1988 (NELS: 88).” Accessed July 8, 2016. http://nces.ed.gov/surveys/nels88/index.asp.

National Research Council. “The Impact of Board-Certified Teachers on Student Outcomes.” In *Assessing Accomplished Teaching: Advanced-Level Certification Programs*, edited by Milton D. Hakel, Judith Anderson Koenig and Stuart W. Elliott, 154–181. Washington, DC: National Academies Press, 2008.

Neuman, S., and D. Celano. “Access to Print in Low-Income and Middle-Income Communities.” *Reading Research Quarterly* 36, no. 1 (2001): 8–26.

New York State Department of Education. “New York State Teacher and Principal Evaluation: Summary of Provisions in Draft Regulations.” Accessed July 8, 2016. http://usny.nysed.gov/rttt/docs/summary.pdf.

Newton, Xiaoxia A., Linda Darling-Hammond, Edward Haertel, and Ewart Thomas. “Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts.” *Education Policy Analysis Archives* 18, no. 23 (2010). Accessed July 8, 2016. http://epaa.asu.edu/ojs/article/view/810.

Ninio, A. “Picture-Book Reading in Mother-Infant Dyads Belonging to Two Subgroups in Israel.” *Child Development* 51 (1980): 587–590.

Niu, Sunny X., and Marta Tienda. “Testing, Ranking and College Performance: Does High School Matter?” Princeton University, September, 2009.

Nixon, Andy, Abbot Packard, and Gus Douvanis. “Non-Renewal of Probationary Teachers: Negative Retention.” *Education* 131, no. 1 (2010): 43–53.

Nunnery, J. A. , S. M. Ross, and A. McDonald. “A Randomized Experimental Evaluation of the Impact of Accelerated Reader/Reading Renaissance Implementation on Reading Achievement in Grades 3 to 6.” *Journal of Education for Students Placed At Risk* 11, no. 1 (2006): 1–18.

NY Daily News. “The Tenure Trap.” Accessed July 7, 2016. http://www.nydailynews.com/opinion/tenure-trap-careful-hiring-teachers-stuck-life-article-1.443471.

Nye, B., L. V. Hedges, and S. Konstantopoulos. “Are Effects of Small Classes Cumulative? Evidence from a Tennessee Experiment.” *Journal of Educational Research* 94, no. 6 (2001): 336–345.

———. “The Long-Term Effects of Small Classes: A Five-Year Follow-up of the Tennessee Class Size Experiment.” *Educational Evaluation and Policy Analysis* 21, no. 2 (1999): 127–142.

Nygard, R. “Personality, Situation, and Persistence: A Study with Emphasis on Achievement Motivation.” Oslo: Universitetsforlaget, 1977.

Ogbu, John U. “Cultural Problems in Minority Education: Their Interpretations and Consequences--Part One: Case Studies.” *The Urban Review* 27, no. 4 (1995): 271–297.

———. “Cultural Problems in Minority Education: Their Interpretations and Consequences--Part One: Theoretical Background.” *The Urban Review* 27, no. 3 (1995): 189–205.

Papay, J. P. “Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates across Outcome Measures.” *American Educational Research Journal* 48, no. 1 (2011): 163–193.

Peterson, C., S. F. Maier, and M. E. P. Seligman. *Learned Helplessness: A Theory for the Age of Personal Control*. New York: Oxford University Press, 1995.

Quartz, K.H., K.B. Lyons, K. Masyn, B. Olsen, L. Anderson, A. Thomas, J. Goode, and E.L. Horng. “Urban Teacher Retention Policy: A Research Brief.” Retention report series: A longitudinal study of career urban educators. Report No. rrs-rr006-0704. Los Angeles: University of California, Los Angeles, 2004.

Ramist, Leonard, Charles Lewis, and Laura McCamley-Jenkins. “Student Group Differences in Predicting College Grades: Sex, Language, and Ethnic Groups.” College Board Report No. 93–1. New York: College Entrance Examination Board, 1994.

Raudenbush, S. W. “Educational Applications of Hierarchical Linear Models: A Review.” *Journal of Educational and Behavioral Statistics* 13, no. 2 (1988): 85–116.

———. “Many Small Groups.” In *Handbook of Multilevel Analysis*, edited by Jan De Leeuw and Erik Meijer, 207–236. New York: Springer, 2008.

———. “What Are Value-Added Models Estimating and What Does This Imply for Statistical Practice?” *Journal of Behavioral and Educational Statistics* 29, no. 1 (2004): 121–129.

Raz, I. S., and P. Bryant. “Social Background, Phonological Awareness and Children’s Reading.” *British Journal of Developmental Psychology* 8 (1990): 209–225.

Reardon, Sean. “The Widening Academic Achievement Gap between the Rich and the Poor: New Evidence and Possible Explanations.” In *Whither Opportunity? Rising Inequality, Schools, and Children’s Life Chances*, edited by Greg J. Duncan and Richard J. Murnane, 91–116. New York: Russell Sage Foundation, 2011.

Reardon, Sean F., Joseph P. Robinson-Cimpian, and Ericka S. Weathers. “Patterns and Trends in Racial/Ethnic and Socioeconomic Academic Achievement Gaps.” In *Handbook of Research in Education Finance and Policy*, edited by Helen A. Ladd and Margaret E. Goertz, 491–509. New York: Lawrence Erlbaum, 2015.

Reckase, M. D. “The Real World Is More Complicated Than We Would Like.” *Journal of Behavioral and Educational Statistics* 29, no. 1 (2004): 117–120.

Reynolds, William M., and Kim L. Miller. “Assessment of Adolescents’ Learned Helplessness in Achievement Situations.” *Journal of Personality Assessment* 53, no. 2 (1989): 211–228.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain. “Teachers, Schools and Academic Achievement.” *Econometrica* 73, no. 2 (2005): 417–458.

Robinson, S. L., C. DePascale, and F. C. Roberts. “Computer Delivered Feedback in Group Based Instruction: Effects for Learning Disabled Students in Mathematics.” *Learning Disabilities Focus* 5, no. 1 (1989): 28–35.

Ross, C. E., and B. A. Broh. “The Roles of Self Esteem and the Sense of Personal Control in the Academic Achievement Process.” *Sociology of Education* 73, no. 4 (2000): 270–284.

Ross, S. M. , J. Nunnery, and E. Goldfeder. “A Randomized Experiment on the Effects of Accelerated Reader/Reading Renaissance in an Urban School District: Final Evaluation Report.” Memphis, TN: Center for Research in Educational Policy, The University of Memphis, 2004.

Rothstein, Jesse. “Review of *Learning About Teaching*.” University of California at Berkeley, 2011.

———. “Revisiting the Impacts of Teachers.” University of California, Berkeley, Goldman School of Public Policy and Department of Economics, October, 2015. http://eml.berkeley.edu/~jrothst/workingpapers/rothstein_cfr_oct2015.pdf.

———. “Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables.” *Education Finance and Policy* 4, no. 4 (2009): 537–571.

———. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125, no. 1 (2010): 175–214.

Rothstein, Richard. "Class and the Classroom." *American School Board Journal* 191, no. 10 (2004). Accessed July 8, 2016. http://www.asbj.com/MainMenuCategory/Archive/2004/October/Class-and-the-Classroom.aspx.

Rowan, B., R. Correnti, and R.J. Miller. "What Large-Scale Survey Research Tells Us About Teacher Effects on Student Achievement: Insights from the Prospects Study of Elementary Schools." *Teachers College Record* 104 (2002): 1525–1567.

Rubin, D. B., E. A. Stuart, and E. L. Zanutto. "A Potential Outcomes View of Value-Added Assessment in Education." *Journal of Behavioral and Educational Statistics* 29, no. 1 (2004): 103–116.

Ryan, R. M., and E. L. Deci. "Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being." *American Psychologist* 55 (2000): 68–78.

Ryan, R. M., V. Mims, and R. Koestner. "The Relationship of Reward Contingency and Interpersonal Context to Intrinsic Motivation: A Review and Test Using Cognitive Evaluation Theory." *Journal of Personality and Social Psychology* 45 (1983): 736–750.

Sanders, W. L., James J. Ashton, and S. Paul Wright. "Comparison of the Effects of NBPTS Certified Teachers with Other Teachers on the Rate of Student Academic Progress." Cary, NC: SAS Institute, Inc., 2005.

Sanders, W. L., and J.C. Rivers. "Cumulative and Residual Effects of Teachers on Future Student Academic Achievement." Knoxville, TN: University of Tennessee Value-Added Research Center, 1996.

Scafidi, Benjamin, David L. Sjoquist, and Todd R. Stinebrickner. "Do Teachers Really Leave for Higher Paying Jobs in Alternative Occupations?" *The B.E. Journal of Economic Analysis and Policy* 6, no. 1 (2006): 1–42.

Scarborough, Hollis S., and Wanda Dobrich. "On the Efficacy of Reading to Preschoolers." *Developmental Review* 14 (1994): 245–302.

Schacter, J., and Y. M. Thum. "Paying for High- and Low-Quality Teaching." *Economics of Education Review* 23 (2004): 411–430.

Schmitz, Bernhard, and Ellen Skinner. "Perceived Control, Effort, and Academic Performance: Interindividual, Intraindividual, and Multivariate Time-Series Analyses." *Journal of Personality and Social Psychology* 64, no. 6 (1993): 1010–1028.

Schrag, Peter. *Final Test: The Battle for Adequacy in America's Schools*. New York: New Press, 2005.

Schunk, D. H. "Self-Efficacy and Academic Motivation." *Educational Psychologist* 26 (1991): 207–231.

———. "Self-Efficacy and Achievement Behaviors." *Educational Psychology Review* 1 (1989): 173–182.

———. "Self-Efficacy and Classroom Learning." *Psychology in the Schools* 22 (1985): 208–223.

———. "Self-Efficacy Perspective on Achievement Behavior." *Educational Psychologist* 19 (1984): 48–58.

Scott, Leslie, Steven J. Ingels, and Jeffrey A. Owings. "Interpreting 12th-Graders' NAEP-Scaled Mathematics Performance Using High School Predictors and Postsecondary Outcomes from the National Education Longitudinal Study of 1988 (NELS:88)." Statistical Analysis Report No. NCES 2007–328. Washington, DC: National Center for Education Statistics, 2007.

Seligman, Martin E. P. *Helplessness: On Depression, Development, and Death*. New York: WH Freeman, 1992.

Seligman, Martin E. P., and Steven F. Maier. "Failure to Escape Traumatic Shock." *Journal of Experimental Psychology* 74, no. 1 (1967): 1–9.

Skinner, E. A. , M. J. Zimmer-Gembeck, J. P. Connell, J. S. Eccles, and J. G. Wellborn. "Individual Differences and the Development of Perceived Control." *Monographs of the Society for Research in Child Development* 63, no. 2/3 (1998): 1–231.

Skinner, E. A., J. G. Wellborn, and J. P. Connell. "What It Takes to Do Well in School and Whether I've Got It: A Process Model of Perceived Control and Children's Engagement and Achievement in School." *Journal of Educational Psychology* 82, no. 1 (1990): 22–32.

Smith, C., R. Constantino, and S. Krashen. "Differences in Print Environment for Children in Beverly Hills, Compton, and Watts." *Emergency Librarian* 24, no. 2 (1996): 4–5.

Smith, Donald E. P., Dale Brethower, and Raymond Cabot. "Increasing Task Behavior in a Language Arts Program by Providing Reinforcement." *Journal of Experimental Child Psychology* 8, no. 1 (1969): 45–62.

Smith, Susan Sidney, and Rhonda G. Dixon. "Literacy Concepts of Low- and Middle-Class Four-Year-Olds Entering Preschool." *The Journal of Educational Research* 88, no. 4 (1995): 243–253.

Spies, R. "The Effect of Rising Costs on College Choice." Research Report No. 117. Princeton, N.J.: Princeton University, 2001.

Staiger, Douglas O., and Jonah E. Rockoff. "Searching for Effective Teachers with Imperfect Information." *Journal of Economic Perspectives* 24, no. 3 (2010): 97–118.

State Council for Educator Effectiveness. "Report and Recommendations." Colorado State Department of Education. Accessed July 8, 2016. https://www.cde.state.co.us/sites/default/files/documents/educatoreffectiveness/downloads/report%20%26%20appendices/scee_final_report.pdf.

Stone, J.E. "The Value-Added Achievement Gains of NBPTS-Certified Teachers in Tennessee: A Brief Report." Accessed July 8, 2016. http://www.education-consumers.com/briefs/stoneNBPTS.shtm.

Stumpf, Heinrich, and Julian C. Stanley. "Group Data on High School Grade Point Averages and Scores on Academic Aptitude Tests as Predictors of Institutional Graduation Rates." *Educational and Psychological Measurement* 62, no. 6 (2002): 1042–1052.

Swinson, J. "A Parental Involvement Project in a Nursery School." *Educational Psychology in Practice* 1 (1985): 19–22.

Teddlie, C., and S. Stringfield. *Schools Do Make a Difference: Lessons Learned from a 10 Year Study of School Effects*. New York: Teachers College Press, 1993.

Tenenbaum, G., and E. Goldring. "A Meta-Analysis of the Effect of Enhanced Instruction: Cues, Participation, Reinforcement and Feedback and Correctives on Motor Skill Learning." *Journal of Research and Development in Education* 22 (1989): 53–64.

The Center for Greater Philadelphia. "Value-Added Assessment." University of Pennsylvania. Accessed July 8, 2016. http://www.cgp.upenn.edu/ope_value.html#9.

The National Commission on Excellence in Education. "A Nation at Risk: The Imperative for Educational Reform." Report No. 065-000-00177-2. Washington, DC: The National Commission on Excellence in Education, 1983.

Tough, Paul. *Whatever It Takes: Geoffrey Canada's Quest to Change Harlem and America*. Boston: Houghton Mifflin Harcourt, 2008.

Tuck, Kathy D. "Parent Satisfaction and Information (a Customer Satisfaction Survey)." Washington, DC: District of Columbia Public Schools, Office of Educational Accountability, Assessment and Information, 1995.

Turque, Bill. "D.C. Teacher Evaluation Formula Could Change." The Washington Post. Accessed July 8, 2016. http://www.washingtonpost.com/local/education/dc-teacher-evaluation-formula-could-change/2012/05/24/gJQACOyRoU_story.html.

Tuttle, Christina Clark, Bingru Teh, Ira Nichols-Barrer, Brian P. Gill, and Philip Gleason. "Student Characteristics and Achievement in 22 KIPP Middle Schools." Washington, DC: Mathematica Policy Research, 2010.

U.S. Department of Education. "Approved Evidence-Based, Whole-School Reform Models." U.S. Department of Education. Accessed July 8, 2016. http://www2.ed.gov/programs/sif/sigevidencebased/index.html.

———. "Department of Education Budget Tables." Accessed July 8, 2016. http://www2.ed.gov/about/overview/budget/tables.html?src=ct.

———. "Digest of Education Statistics." National Center for Education Statistics. Accessed July 8, 2016. http://nces.ed.gov/programs/digest/d05/tables/dt05_075.asp.

———. "New No Child Left Behind Flexibility: Highly Qualified Teachers." Accessed July 8, 2016. http://www2.ed.gov/nclb/methods/teachers/hqtflexibility.html.

———. "Teacher Incentive Fund." U.S. Department of Education. Accessed July 8, 2016. http://www2.ed.gov/programs/teacherincentive/index.html.

———. "Teacher Shortage Areas: Nationwide Listing, 1990-91 Thru 2011-12." Washington, DC: U.S. Department of Education, 2011.

U.S. Department of Labor. "Employer Costs for Employee Compensation: December 2007." Bureau of Labor Statistics. Accessed July 8, 2016. http://www.bls.gov/news.release/archives/ecec_03122008.pdf.

Vallerand, R. J., and G. Reid. "On the Causal Effects of Perceived Competence on Intrinsic Motivation: A Test of Cognitive Evaluation Theory." *Journal of Sport Psychology* 6 (1984): 94–102.

Vandevoort, L. G., A. Amrein-Beardsley, and D. C. Berliner. "National Board Certified Teachers and Their Students' Achievement." *Education Policy Analysis Archives* 12, no. 46 (2004). Accessed July 8, 2016. http://epaa.asu.edu/ojs/article/view/201.

Walberg, H. J. "What Makes Schooling Effective?" *Contemporary Education Review* 1 (1982): 1–34.

Watanabe, Teresa. "'Value-Added' Teacher Evaluations: L.A. Unified Tackles a Tough Formula." Los Angeles Times. Accessed July 8, 2016. http://articles.latimes.com/2011/mar/28/local/la-me-adv-value-add-20110328.

Weber, Karl, ed. *Waiting for "Superman": How We Can Save America's Failing Public Schools*. New York: Public Affairs, 2010.

Wehlage, Gary C., and Robert A. Rutter. "Dropping Out: How Much Do Schools Contribute to the Problem?" *Teachers College Record* 87, no. 3 (1986): 374–392.

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, NY: The New Teacher Project, 2009.

Wells, G. "Preschool Literacy-Related Activities and Success in School." In *Literacy, Language, and Learning*, edited by D. R. Olson, N. Torrance and A. Hildyard, 229-255. Cambridge: Cambridge University Press, 1985.

Wenglinsky, H. “Teacher Classroom Practices and Student Performance: How Schools Can Make a Difference.” Research Report No. RR-01-19. Princeton, NJ: Educational Testing Service, 2001.

Wesson, Linda, Kim Potts, and Kelly Hill. “Use of Value-Added in Teacher Evaluations: Key Concepts and State Profiles.” Nashville, TN: Tennessee Comptroller of the Treasury, Offices of Research and Education Accountability, 2015.

Whitehurst, Grover J., and Michelle Croft. “The Harlem Children’s Zone, Promise Neighborhoods, and the Broader, Bolder Approach to Education.” Washington, DC: Brookings Institution, 2010.

Wigfield, Allan, and Michele Karpathian. “Who Am I and What Can I Do? Children’s Self-Concepts and Motivation in Achievement Situations.” *Educational Psychologist* 26, no. 3 and 4 (1991): 233–261.

Winship, C., and S.D. Korenman. “Economic Success and the Evolution of Schooling and Mental Ability.” In *Earning and Learning: How Schools Matter*, edited by S. E. Mayer and P. E. Peterson, 49–78. Washington, DC: Brookings Institution Press, 1999.

Woltmann, Heather, Andrea Feldstain, J. Christine MacKay, and Meredith Rocchi. “An Introduction to Hierarchical Linear Modeling.” *Tutorials in quantitative methods for psychology* 8, no. 1 (2012): 52–69.

Woodworth, Katrina R., Jane L. David, Roneeta Guha, Haiwen Wang, and Alejandra Lopez-Torkos. “San Francisco Bay Area KIPP Schools: A Study of Early Implementation and Achievement: Final Report.” Menlo Park, CA: SRI International, 2008.

Wößmann, L. “International Evidence on Expenditures and Class Size: A Review.” In *Brookings Papers on Education Policy: 2006/2007*, edited by T. Loveless and F. Hess, 245–272. Washington, D.C.: Brookings Institution Press, 2007.

Wright, S.P., S.P. Horn, and W.L. Sanders. “Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation.” *Journal of Personnel Evaluation in Education* 11 (1997): 57–67.

Yazzie-Mintz, Ethan. “Charting the Path from Engagement to Achievement: A Report on the 2009 High School Survey of Student Engagement.” Bloomington, IN: Center for Evaluation and Education Policy, 2010.

Yeh, S. S. “Class Size Reduction or Rapid Formative Assessment? A Comparison of Cost-Effectiveness.” *Educational Research Review* 4, no. 1 (2009): 7–15.

———. *The Cost-Effectiveness of 22 Approaches for Raising Student Achievement.* Charlotte, NC: Information Age Publishing, 2011.

———. “The Cost-Effectiveness of 22 Approaches for Raising Student Achievement.” *Journal of Education Finance* 36, no. 1 (2010): 38–75.

———. “The Cost-Effectiveness of Comprehensive School Reform and Rapid Assessment.” *Education Policy Analysis Archives* 16, no. 13 (2008). Accessed July 8, 2016. http://epaa.asu.edu/ojs/article/view/38/164.

———. “The Cost-Effectiveness of Five Policies for Improving Student Achievement.” *American Journal of Evaluation* 28, no. 4 (2007): 416–436.

———. “The Cost-Effectiveness of NBPTS Teacher Certification.” *Evaluation Review* 34, no. 3 (2010): 220–241.

———. “The Cost-Effectiveness of Raising Teacher Quality.” *Educational Research Review* 4, no. 3 (2009): 220–232.

———. “High Stakes Testing: Can Rapid Assessment Reduce the Pressure?” *Teachers College Record* 108, no. 4 (2006): 621–661.

———. “A Re-Analysis of the Effects of Teacher Replacement Using Value-Added Modeling.” *Teachers College Record* 115, no. 12 (2013).

———. “The Reliability, Impact and Cost-Effectiveness of Value-Added Teacher Assessment Methods.” *Journal of Education Finance* 37, no. 4 (2012): 374–399.

———. “Shifting the Bell Curve: The Benefits and Costs of Raising Student Achievement.” *Evaluation and Program Planning* 32, no. 1 (2009): 74–82.

———. “Two Models of Learning and Achievement: An Explanation for the Achievement Gap?” *Teachers College Record* 117, no. 12 (2015).

———. “Understanding and Addressing the Achievement Gap through Individualized Instruction and Formative Assessment.” *Assessment in Education* 17, no. 2 (2010): 169–182.

Yeh, S. S., and J. Ritter. “The Cost-Effectiveness of Replacing the Bottom Quartile of Novice Teachers through Value-Added Teacher Assessment.” *Journal of Education Finance* 34, no. 4 (2009): 426–451.

Ysseldyke, J., and D. M. Bolt. “Effect of Technology-Enhanced Continuous Progress Monitoring on Math Achievement.” *School Psychology Review* 36, no. 3 (2007): 453–467.

Ysseldyke, J., and S. Tardrew. “Use of a Progress-Monitoring System to Enable Teachers to Differentiate Math Instruction.” *Journal of Applied School Psychology* 24, no. 1 (2007): 1–28.

About the Author

Stuart Yeh is Associate Professor of Evaluation Studies at the University of Minnesota. Over the past 18 years, he has pursued studies regarding the source of the achievement gap and promising strategies for addressing the gap. He has published numerous scholarly articles and two books, including a book-length evaluation of the cost-effectiveness of 22 approaches for raising student achievement.

Dr. Yeh has served as an invited expert at the Organisation for Economic Co-operation and Development (OECD). He previously served as Senior Research Associate at the Center for the Study of Testing, Evaluation, and Educational Policy at Boston College, where he worked on a national study of educational testing. He has evaluated Head Start literacy programs in the Dorchester, Jamaica Plain, and Allston-Brighton neighborhoods of Boston and evaluated critical thinking and analytical writing programs in East Palo Alto and San Jose, California. At the Manpower Demonstration Research Corporation (MDRC), he worked on a randomized evaluation of high school career academy programs.

Dr. Yeh was awarded a BA in Economics and a Master of Public Policy degree at the University of Michigan, a PhD in Evaluation Studies at Stanford University, and a post-doctoral fellowship at Harvard University.

S.S. Yeh, *Solving the Achievement Gap*,
DOI 10.1057/978-1-137-58767-1

Index

Note: Page numbers followed by “n” refers to endnotes.

S.S. Yeh, *Solving the Achievement Gap*,
DOI 10.1057/978-1-137-58767-1