TEACHER MATTERS:
RE-EXAMINING THE EFFECTS OF GRADE-3 TEST-BASED RETENTION POLICY

by

Yihua Hong

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Department of Curriculum, Teaching and Learning
Ontario Institute for Studies in Education of the
University of Toronto

© Copyright by Yihua Hong (2012)
ABSTRACT

This study is aimed to unpack the ‘black box’ that connects the grade-3 test-based retention policy with students’ academic outcomes. I theorized that the policy effects on teaching and learning may be modified by instructional capacity, but are unlikely to occur through enhancing teachers’ capability to teach. Analyzing the Early Childhood Longitudinal Study Kindergarten cohort (ECLS-K) dataset, I first explored the relationship between the test-based retention policy and instructional capacity as indicated by teacher expectations of students’ learning capability and then investigated whether and how the expectations moderated the policy effects on instructional time reallocation, student academic performance, and student self-perceived academic competence and interests. To remove the selection bias associated with the non-experimental data, I applied a novel propensity score-based causal inference method, the marginal mean weighting through stratification (MMW-S) method and extended it to a causal analysis that approximates a randomization of schools to the test-based retention policy followed by a randomization of classes to teachers with different levels of expectations. Consistent with my theory, I found that the test-based retention policy had no effects on teacher expectations. Although the policy uniformly increased the time allocated to math instruction, it produced no significant changes in students’ overall performance and overall self-perception in math. In addition, I found that students responded differently to the test-based retention policy depending on the expectations they received from the grade-3 teachers. The results suggested some benefits
of positive expectations over negative and indifferent expectations in moderating the policy effects, including more access to advanced content, higher learning gains of average-ability students, and more resilient student learning over a long term. However, the results also showed that having positive expectations alone is not sufficient for academic improvement under the high-stakes policy. If implemented by a positive-expectation teacher, the policy could be detrimental to students’ learning in the nontested subject or to their learning of basic reading/math skills. It would as well place the bottom-ability students at a disadvantage. The findings have significant implications for the ongoing high-stakes testing debate, for school improvement under the current accountability reform, and for research of teacher effectiveness.
To My Son, Hongze
ACKNOWLEDGEMENTS

My heartfelt appreciation goes to the many people who have contributed to the completion of my dissertation. First and foremost, I would like to express my sincerest gratitude to my supervisor, Dr. Guanglei Hong, whose unconditional support and constant encouragement have helped me survive the toughest time of my life and have made this dissertation possible. Guanglei is for me an exceptional mentor, and a role model of rigorous scholarship, professionalism, and kindness. No words can begin to describe how much I have benefited from her careful guidance.

I am greatly indebted to my committee members, Dr. Kenneth Leithwood and Dr. Eunice Jang for their invaluable advice and help. Dr. Leithwood’s insightful words on the importance of policy settings have reshaped my view of high-stakes testing issues and steered me to this exciting research. Dr. Jang’s detailed feedback on my comprehensive exam paper and on the subsequent thesis work has been very helpful and has always pointed out the blind spots in my writing. Each of the members has unreservedly shared with me his/her unique expertise throughout the thesis process, for which I am truly grateful.

I am also very grateful to my external examiner, Dr. Brian Jacob from the University of Michigan, for his thorough reading of my dissertation. With Dr. Jacob’s constructive feedback, I was able to further improve my work and to identify a new direction for my future research.

I would like to acknowledge my thanks to Dr. Yongyun Shin from the Virginia Commonwealth University for his advice on the multilevel imputation procedure. My thanks also go to Richard Congdon from the Harvard University whose timely help with the HLM program has greatly eased my analysis process.

I appreciate the opportunities provided by the William T. Grant Foundation and the
AERA Grants program, which enabled me to communicate with other researchers on my thesis problems. I thank Dr. Nonie Lesaux from the Harvard University for her discussion on my work-in-progress during the 2011 William T. Grant Mentoring Workshop. I also thank Dr. Stephen Raudenbush and Dr. Elaine Allensworth from the University of Chicago. Despite their busy schedule, each of them took the time to listen to my research findings before my final defense and raised some good questions that helped refine the thesis.

I owe thanks to many great professors and staff at OISE/UT who have not been directly involved in my thesis process, but have had positive influence on my academic life. I have always been thankful to Dr. Tony Lam for his teaching and guidance in my early graduate studies. I appreciate Dr. Alister Cummings for his generous offer of help during my search for directions. I would like to express my appreciation as well to Dr. Carl Corter and Dr. Janette Pelletier, whom Guanglei and I have coauthored with on a series of papers that applied the MMW method to evaluations of instructional practices. For the last five years, I have benefited a lot from working with them and from basking in their intellectual discussion. In addition, I want to acknowledge the support of the CTL Department, especially of Terry Louisy for my further study of mathematics and statistics outside OISE in the 4th and 5th years of the doctoral program.

I have been very fortunate to be accompanied by many friends throughout my academic journey. I extend my thanks to my thesis buddies, Emiko and Bing, and my lab mates, Marija, Natasha, and Sara, for their intellectual and moral support. My same thanks go to all my friends from the OISE Chinese Community, Jingping, Jingshun, Naxin, Jian, Shijing, Yang, Zhimei, Lichun, Lingqing, Mingzhu, ... and many more. I give special thanks to my friend Su for always being there to listen, to offer advice, and to help me out.
This dissertation could not have been written without the love and support from my family. I wish to thank each of the family members for everything he/she has done for me. My son, Hongze, has always been very curious about my research and writing and very understanding of my busy schedule. He has been eagerly awaiting my “book” to come out for many years. I dedicate this dissertation to him.

This research was supported by a grant from the American Educational Research Association which receives funds for its “AERA Grants Program” from the National Science Foundation under Grant #DRL-0941014. Opinions reflect those of the author and do not necessarily reflect those of the granting agencies.
TABLE OF CONTENTS

ABSTRACT ................................................................................................................................... ii
ACKNOWLEDGEMENTS .......................................................................................................... v
LIST OF TABLES ...................................................................................................................... x
LIST OF FIGURES ................................................................................................................... xii
LIST OF APPENDICES ........................................................................................................... xiii

CHAPTER 1
INTRODUCTION......................................................................................................................... 1
  Research Scope and Questions ............................................................................................ 2
  Conceptual Clarification ...................................................................................................... 5
  Organization of the Dissertation ......................................................................................... 8

CHAPTER 2
CONCEPTUAL FRAMEWORK AND PAST EVIDENCE ..................................................... 9
  Effects of High-stakes Testing and Test-based Retention Policy: A Retrospect .............. 10
  Variations in Instructional Practices and Capacity: A Supplementary Account .......... 18
  Grade-3 Test-based Retention Policy and Teacher Expectations: A Situated View ....... 27

CHAPTER 3
METHODOLOGICAL CHALLENGES IN STUDIES OF HIGH-STAKES TESTING ....... 42
  Measures of Student Academic Performance .............................................................. 42
  Problem of Selection Bias ................................................................................................. 47
  Generalizability of Research Findings .............................................................................. 51

CHAPTER 4
RESEARCH METHODS ......................................................................................................... 54
  Data .................................................................................................................................... 54
  Measures ............................................................................................................................ 54
  Causal Inference Strategies .............................................................................................. 65
  Analytic Procedure ............................................................................................................ 74
  Notes to Chapter 4 .............................................................................................................. 82

CHAPTER 5
GRADE-3 TEST-BASED RETENTION POLICY AND TEACHER EXPECTATIONS .. 85
  Association with Pretreatment Characteristics ............................................................. 85
  Relationship between Grade-3 Test-based Retention Policy and Teacher Expectations .... 86
LIST OF TABLES

Table 4.1  Summary of K-5 $\theta$ and OLDS Scores .............................................................. 61
Table 4.2  Student Distribution over the Proficiency Levels ...................................................... 62
Table 4.3  Potential Outcomes and Causal Estimands ............................................................. 68
Table 4.4  Class Distribution across Treatment Conditions in the Final Analytic Sample ....... 78
Table 4.5  Student Distribution across Ability Levels and Treatment Conditions in the Final Analytic Sample .......................................................................................................... 78
Table 4.6  Comparison between ECLS-K K-5 Full Sample and the Final Analytic Sample ..... 79
Table 6.1  Weighted Analysis of the Policy-by-Teacher Effects on Instructional Time Allocation .................................................................................................................... 90
Table 7.1  Weighted Analysis of the Policy-by-Teacher Effects on Student Overall Academic Performance .................................................................................................................. 100
Table 7.2  Estimated Odds of Skill Mastery Corresponding to the Policy-by-Teacher Treatments Conditions.................................................................................................................. 106
Table 7.3  Weighted Analysis of the Policy-by-Teacher Effects on Student Mastery of Each Cognitive Skills ........................................................................................................ 109
Table 7.4  Estimated Mean (Standard Error) of Student Academic Performance by Prior Ability and Treatment Conditions ......................................................................................... 116
Table 7.5  Estimated Effects of the Test-based Retention Policy on Student Academic Performance by Student Prior Ability and Teacher Expectations............................. 118
Table 8.1  Estimated Mean (Standard Error) of Student Academic Self-perception by Prior Ability and Treatment Conditions............................................................................. 134
Table 8.2  Estimated Policy Effects on Student Academic Self-perception by Teacher Expectations ........................................................................................................ 135
Table D1  List of Confounders for the School-level Propensity Model ..................................... 172
Table D2  List of Confounders for the Class-level Propensity Models .................................... 173
Table F1  School-level Marginal Mean Weights for the Final Analytic Sample .......................... 180
Table F2  Class-level Marginal Mean Weights for the Final Analytic Sample .......................... 181
Table F3  Student-level Marginal Mean Weights for the Final Analytic Sample ........................ 184

Table G1  Between-treatment Group Differences in Logit Propensity Scores before and after Weighting .......................................................................................................................... 187

Table H1  Prognostic Variables in the Final Outcome Models ....................................................... 190
LIST OF FIGURES

Figure 2.1 Conceptual Framework for Evaluating the Effects of Grade-3 Test-based Retention Policy ................................................................. 9

Figure 6.1 Estimated Effects of the Test-based Retention Policy and Teacher Expectations on Instructional Time by Subjects ................................................. 91

Figure 6.2 Between-subjects Comparison of the Estimated Policy-by-teacher Effects on Instructional Time Allocation ................................................................. 95

Figure 7.1 Estimated Effects of the Test-based Retention Policy and Teacher Expectations on Student Overall Academic Performance by Subjects ......................................... 99

Figure 7.2 Between-subjects Comparison of the Estimated Policy-by-teacher Effects on Student Overall Academic Performance ......................................................... 103

Figure 7.3 Across-skills Comparison of the Estimated Policy Effects by Teacher Expectations on Student Mastery of Cognitive Skills .................................................... 113

Figure 7.4 Across-ability Levels Comparison of the Estimated Policy-by-Teacher Effects on Student Academic Performance ......................................................... 123

Figure E1 Initial Distribution of the School-level Logit Scores ......................................................... 175

Figure E2 Distribution of the Class-level Logit Probability Scores in Cluster 1 ............................. 179
LIST OF APPENDICES

APPENDIX A. Empirical Identification of Student Prior Academic Ability..........................161
APPENDIX B. List of Pretreatment Variables........................................................................163
APPENDIX C. Imputation of Multilevel Structured Data .........................................................168
APPENDIX D. Estimating Multilevel Propensity Scores..........................................................170
APPENDIX E. Strategy for Identifying Common Support.........................................................175
APPENDIX F. Multilevel Marginal Mean Weights (MMW).....................................................180
APPENDIX G. Weighted Multilevel Balance Checking........................................................... 186
APPENDIX H. List of Prognostic Variables in the Final Outcome Models.............................. 190
CHAPTER 1
INTRODUCTION

Throughout the evolution of test-based accountability reforms (see review by Haertel & Herman, 2005; Shepard, 2008), policymakers have been promoting high-stakes testing as a panacea for improving American public schools. They believe that high stakes incentives, when tied to test results, will motivate students and teachers, and ultimately will improve student learning (Hamilton, Stecher, & Klein, 2002; Madaus, Russell, & Higgins, 2009). This belief (also known as the working theory of high-stakes testing), though appealing, has not been substantiated by empirical research. Past evaluations of high-stakes testing (see review by Harris & Herrington, 2006; Lee, 2008a; Loveless, 2005; NRC, 2011) suggest that student academic performance may vary across different cognitive skills, differ between tested and nontested subjects, and may change over time. In addition, students may respond to high-stakes incentives differently according to their ability levels. However, findings in accumulation still remain inconclusive. Previous studies have attempted to use economic incentive theories (i.e. Holmstrom, 1979; Holmstrom & Milgrom, 1991; Stiglitz, 1987) and/or educational motivation theories (i.e. Ames, 1992; Deci & Ryan, 1985) in explaining the variations in the observed policy effects. The theories, however, seem unable to reconcile the inconsistent findings.

In this dissertation, I argue that the confusion in the past research may partly arise from the lack of attention to the role of teachers and to their instructional capacity. Broader literature on school reform and teacher change (Cohen & Hill, 2001; EEPA, 1990; McLaughlin, 1987) told us that teachers are a key factor for understanding the differences in academic performance under a high-stakes regime. Teachers’ instructional capacity, or crudely speaking, their skill and will to teach (Cohen, 1996a), which may vary even within the same school, shapes teachers’ interpretation of high-stakes testing and their instructional responses to the policy. As high-
stakes incentives do not provide teachers with opportunities to improve their teaching, I theorize that teachers’ instructional capacity moderates the effects of high-stakes testing on student academic outcomes.

**Research Scope and Questions**

In view of the diversity of high-stakes testing policies and the variations in teachers’ instructional capacity, as a first step to testing my theory, this study first evaluated the relationship between grade-3 test-based retention policy and teachers’ expectations of their students’ learning capability, and then examined their joint effects on three outcomes: student academic performance, instructional time reallocation, and student self-perceived competence and interests. A review of literature (e.g. Hughes, Gleason, & Zhang, 2005; Ready & Wright, 2011; Timperly & Philips, 2003; van den Bergh, Hornstra, Voeten, & Holland, 2010) suggests that although teacher expectations are most strongly determined by students’ actual ability, they can also be shaped by classroom and school contexts and may be associated with some important attributes of individual teachers such as their knowledge and skills, their belief of gender/ethnicity differences, and their self-efficacy. Therefore, I view teacher expectations as partly reflecting teachers’ instructional capacity though the latter is a far more complicated and comprehensive construct than the former. Since teachers tend to calibrate instruction to the self-perception of student ability (Cohen, Raudenbush, & Ball, 2003), teachers’ expectations may exert important influence on their own and the students’ responses to high-stakes incentives.

At present the test-based retention policy has been formerly implemented in several states (e.g. Texas, Florida, and Louisiana) and large school districts (e.g., New York and Chicago) (Penfield, 2010). Judging from the current high demand for school accountability, chances are that more schools will abandon social promotion in favor of test-based promotion and will
expand the policy to early grades in order to push students to exert more study efforts and to improve school overall performance rating. Therefore it is necessary to evaluate whether and how the policy facilitates the academic learning of students in grade 3 – a crucial mid-point in elementary education often chosen as the promotional gate year; and it is also necessary to examine how instructional capacity determines student performance during the policy process.

Previous studies (e.g. Bryk, 2003; Jacob, 2005; Roderick, Jacob, & Bryk, 2002), though having generated some interesting findings regarding the effects of test-based retention policy in the promotional gate grade, were mostly limited in the Chicago context. They were also constrained by the choice of high/ moderate-stakes tests to measure student performance and in their statistical approach to removing selection bias. Although most of them attended to school and student variations, none has taken into account the moderating effects of the instructional capacity of individual teachers.

To overcome the limitations of the existing research, my dissertation study is situated in the U.S. national context and uses low-stakes tests to measure student performance. In order to remove selection bias in estimating the policy effects, I apply a propensity score-based causal inference procedure, the marginal mean weighting through stratification (MMW-S) method (Hong, 2010) and extend it to a causal analysis of multilevel treatments. The primary goal is to unpack the ‘black box’ that connects the grade-3 test-based retention policy and student academic performance. I analyze four different aspects of student performance under the test-based policy, including students’ learning in tested versus nontested subjects, their mastery of different cognitive skills in tested subjects, their differential learning by prior ability, and the long-term academic trend. The working theory of high-stakes testing implies that the policy effects operate through teaching practices and student motivation (Hamilton et al., 2002; Madaus
et al., 2009). I therefore also examine teachers’ instructional time allocation and students’ self-perceived academic competence and interests as two extra policy outcomes and evaluate the assumed mechanism in the grade-3 policy context. Seeking an in-depth understanding of the role of instructional capacity, I first explore the relationship between teacher expectations and the grade-3 test-based retention policy and confirm that teachers’ perception of student academic competence cannot be manipulated by the provision of the high-stakes incentive. I then investigate how the expectations modify the policy effects on each of the three outcomes. Specifically my investigation involves the following four sets of research questions.

1. **Relationship between the policy and teacher expectations:**

   Does the grade-3 test-based retention policy have any influence on teachers’ expectations of student learning ability?

2. **Policy-by-teacher effects on instructional time allocation:**

   Does the grade-3 test-based retention policy lead to any change in teachers’ instructional time allocation? Does the policy effect depend on teacher expectations? Do the policy-by-teacher effects vary among different subjects, especially between tested (i.e. reading and math) and non-tested (i.e. science) subjects?

3. **Policy-by-teacher effects on student academic performance:**

   **Student overall academic performance.** Do the grade-3 test-based retention policy and teacher expectations jointly affect student overall academic performance in the promotional gate year? How do the policy-by-teacher effects differ between different subjects?

   **Student mastery of cognitive skills.** What are the joint effects of test-based retention policy and teacher expectations on student learning of each cognitive skill in tested subjects? How do the policy-by-teacher effects differ among skills?
**Student differential academic performance by prior ability.** How do the policy and teacher expectations affect the academic performance of students at each ability level? Do they produce differential effects across different ability subpopulations?

**Student long-term academic performance.** Are the effects on student overall academic performance and/or on student differential performance by prior ability persistent till two years later?

4. **Policy-by-teacher effects on student self-perceived competence and interests:**

   **Student overall self-perception.** Do the grade-3 test-based retention policy and teacher expectations jointly affect student self-perceived competence and interests toward tested subjects in grade 3? How do the policy-by-teacher effects differ between different subjects?

   **Student differential self-perception by prior ability.** How do the policy-by-teacher effects differ among students at different prior ability levels? Do they produce differential effects across different ability subpopulations?

   **Student long-term self-perception.** Do the effects sustain till two year later?

---

**Conceptual Clarification**

Before proceeding further, it is necessary to clarify several conceptual issues related to high-stakes testing, test-based retention policy, and teacher expectations.

**High-stakes Testing**

High-stakes testing, also known as test-based accountability, refers to a set of policies and procedures that “provide rewards and/or sanctions as a consequence of scores on large-scale achievement tests” (Hamilton et al., 2002, p. 3). The tests used for such policies are high-stakes tests, as opposed to low-stakes tests that are used only to provide information on student learning. In the current American educational system, high-stakes testing has two major
components: one is the high stakes associated with the existence of rewards/sanctions for good/poor student academic performance; and the other is the testing that relies on standardized tests to monitor the performance. In view of the voluminous literature on standardized testing (see review by Linn, 2001; Wang, Beckett, & Brown, 2006), my dissertation focuses on the effects of high stakes rather than on those of standardized testing. I resonate with Phelps (2005) that the dispute around standardized testing should be considered as a clash of interests between different educational communities and with Camilli (2003) that “the crux of the [high-stakes testing] matter is not the broad issue of testing, but high-stakes testing” (p. 36).

Judging from the prevalence of high-stakes testing in the American public school system, I situate my review and discussion of high-stakes effects in the U.S. Context and analyze a U.S, national dataset. I anticipate that my research may also have certain implications for other countries and for their use of high-stakes testing. But its findings should be evaluated with caution if being generalized to a different geographical context.

**Test-based Retention Policy**

High-stakes testing can target at students, teachers, schools, and school systems, etc. (AERA, 2000; Heubert & Hauser, 1999). In response to the No Child Left Behind (NCLB) Act of 2001 and/or its predecessors that stressed on school accountability, previous research on high-stakes testing was mostly dedicated to evaluating test-based decisions for schools or states. Considering that the effects of accountability policies ultimately depend on the learning efforts made by students (Chabran, 2003), this dissertation shifts attention to the incentives that can directly bring consequences to individual students. It highlights the test-based retention policy that requires students to achieve a certain minimum level on a standardized test before they can be promoted to the next grade (Greene & Winters, 2007; Thomas, 2005).
The purpose of using test-based retention policy is often two-fold in nature: (a) to motivate students and teachers with the threat of retention, and (b) to identify academically at-risk students and help them acquire necessary knowledge and skills for the success in higher grades by using grade repetition (i.e. the sanction of retention) (Heubert & Hauser, 1999; Thomas, 2005). The former justifies that the test-based retention policy shares with other high-stakes testing initiatives the same mechanism in regards to the policy effects on student academic performance. The latter explains why opponents of the test-based retention policy (e.g., Heubert & Hauser, 1999; Penfield, 2010) resorted to the evidence of negative effects of retention sanction in general and why many existing research on test-based retention policy (e.g. Greene & Winters, 2007; Greene, Winters, & Forster, 2004; McCombs, Kirby, & Mariano, 2010; Roderick & Nagaoka, 2005) attended to retained students and their learning as a result of grade repetition.

The current study is interested in the first-fold purpose and focuses on the threat of retention, but not the sanction. I believe that the effects of the test-based retention policy, if there are any, are not limited to the time period after retention decision is made, but rather will occur immediately following the introduction of the policy; the effects are not only concentrated on academically at-risk students, but also may expand to other students who are subject to the test-based retention policy in the promotional gate grade.

**Teacher Expectations**

The current study evaluates, given student composition and school characteristics, how differences in teacher expectations may interfere with the policy implementation and may lead to different student outcomes under high-stakes regime. I consider teacher expectations as one narrow aspect of instructional capacity. In Chapter 2, my review of literature provides evidence that teacher expectations reflect some important attributes of individual teachers’ capability to
teach, including their knowledge, skills, and certain beliefs. By underscoring these attributes, I do not mean to devalue other attributes of instructional capacity, especially of institutional capacity. But I believe that individual teachers’ skills and will are as influential as institutional environment. In order to improve student learning under institutional pressure, every teacher first has to have a positive expectations of his/her students’ academic ability.

Emphasizing on teacher expectations, this study, for the first time, brings together a macro-level policy and a micro variation among teachers in attempting to understand the relationship between high-stakes testing and student academic performance. I expect that the findings will have important implications for professional capacity building under the current test-based accountability system.

**Organization of the Dissertation**

The remainder of the dissertation is organized as follows. Chapter 2 develops a conceptual framework for evaluating the effects of the grade-3 test-based retention policy based on the past evidence on high-stakes testing and test-based retention policy. To supplement the extant theories, I theoretically analyze the role of instructional capacity and teacher expectations in explaining the policy effects. Chapter 3 critically reviews the strengths and weaknesses of previous research methods and identifies methodological challenges for studies of high-stakes testing. Chapter 4 describes the data source, measures, causal inference strategies, and analytic procedure that I employ in the current study. Chapter 5-8 presents the results corresponding to the four sets of research questions; at the end of each of the four chapters, I also provide a summary of the findings. Chapter 9 links and interprets the results from the previous chapters. I discuss their conceptual and methodological implications and limitations and propose directions for future studies.
Chapter 2 develops a conceptual framework for evaluating the effects of grade-3 test-based retention policy. The framework places teachers at the center of the policy implementation and highlights the interaction between the external accountability policy and internal individual instructional capacity as indicated by teacher expectations of students’ learning ability. As illustrated in Figure 2.1, the effects of the test-based retention policy are expected not only to channel through teachers’ instructional practices, but also to be modified by teacher expectations. The underlying idea is that teachers’ instructional capacity determines whether and how teachers will respond to the policy and partly explains variations in instructional practices and hence differences in student learning and motivation under the high-stakes regime. The conceptual framework as well attends to learning and motivation characteristics of grade-3 students. It views an indicator of student motivation, student self-perceived academic competence and interests as an outcome of instructional practices and does not assume the student motivation can be entirely translated into learning for third graders.

Figure 2.1. Conceptual framework for evaluating the effects of grade-3 test-based retention policy.
To justify the framework, I first take stock of existing evidence and theoretical explanations for the relationship between high-stakes testing and student academic performance. After identifying limitations in the previous literature, I present variations in instructional practices and capacity as a supplementary account for the effects of high-stakes testing. Subsequently I highlight teacher expectations as an indicator of individual teachers’ instructional capacity and situate the theoretical framework in grade 3. By doing so, I establish my hypotheses regarding the joint effects of the grade-3 test-based retention policy and teacher expectations.

**Effects of High-stakes Testing and Test-based Retention Policy: A Retrospect**

*Past Evidence on High-stakes Testing and Student Academic Performance*

Over the last decade, there have been continuous controversies surrounding the benefits and problems of high-stakes testing (e.g. Camilli, 2003; Cizek, 2001, 2003, 2005; Rich, 2003; also see review by Wang et al., 2006). Most research attention has been directed to some immediate negative reactions of teachers and/or schools, such as aggravated dropout and test-exclusion problems (e.g. Cullen & Reback, 2006; Figlio & Getzler, 2002; Heilig & Darling-Hammond, 2008; Jacob, 2005), narrowed curriculum and inappropriate test preparation (e.g. Au, 2007; Shepard & Dougherty, 1991; Stecher, 2001), and even teacher cheating during high-stakes tests (e.g. Jacob & Levitt, 2003, 2004). A smaller body of work has been concentrated on the ultimate policy outcome, that is, student academic performance (see review by Harris & Herrington, 2006; Lee, 2008a; Loveless, 2005; National Research Council, 2011).

The available work, however, has produced a rather confusing picture regarding the effects of high-stakes testing on student overall academic performance. While quite a few studies (e.g. Betts & Danenberg, 2002; Bryk, 2003; Camilli & Bulkley, 2001; Hanushek & Raymond, 2005; Ladd, 1999; Raymond & Hanushek, 2003; Roderick et al., 2002; Rosenshine, 2003;
Winter, Ritter, Greene, & Marsh, 2009) indicated that high-stakes testing brought out positive changes in student learning, some other studies (e.g., Jacob, 2001; Lee & Wong, 2004) did not detect any significant effects of the policy. In addition, many researchers noticed that the overall policy effects might vary by different contexts including grades and subjects (e.g. Amrein & Berliner, 2002, 2003; Carnoy & Loeb, 2002; CEP, 2007; Dee & Jacob, 2011; Greene, 2001; Nichols, Glass, & Berliner, 2006; Winfield, 1990). The inconsistency implies that a polarized view of the policy effects may underestimate the complexity of high-stakes testing issues.

Previous findings suggest that student learning may not be uniform across different cognitive skills (e.g., Frederiksen, 1994; Mangino & Babcock, 1986) or between tested and nontested subjects (e.g., Jacob, 2005; Reback, 2008); but controversies still exist regarding which aspects of student learning high-stakes testing could benefit or hurt. Although past research provides a hint that the policy effects may change over time (e.g., Bryk, 2003), there is no sufficient evidence for us to understand the long-term trend of the policy effects.

Moreover, researchers found that students may respond to high-stakes incentives differently depending on their ability levels (e.g., Jacob, 2001; Roderick et al., 2002); and the differential responses may be more pronounced in low-performing schools (e.g. Betts & Danenberg, 2003; Jacob, 2005; Roderick et al., 2002; Springer, 2008). However, even after taking the differences in students and/or schools into account, the researchers cannot agree on a specific differential pattern of the policy effects. Several studies (Reback, 2008; Neil and Gayler, 2001) showed that lowest-ability students were the only group that benefited from high-stakes testing; but a majority of studies (Betts & Danenberg, 2003; CEP, 2009; Jacob, 2005; Neal & Schanzenbach, 2010; Springer, 2008) found students of average- or lower-than-average-ability level benefited the most from the policy. Some (e.g., Neal & Schanzenbach, 2010) revealed that
improvement of average-ability students was produced at the expense of lowest-ability students. Others (CEP, 2009; Springer 2008) concluded that the positive gain of the average- or lower-than-average-ability students did not cost the learning of their peers in any other ability groups.

**Past Evidence on Test-based Retention Policy and Student Academic Performance**

Studies specific to test-based retention policy, albeit very limited, similarly revealed a complex relationship between the policy and student academic performance (Bryk, 2003; Jacob, 2003, 2005; Neal & Schanzenbach, 2010; Roderick et al. 2002). The available research evidence was exclusively from the evaluations of Chicago’s Ending Social Promotion reform, the reform that mandated summer school and ultimately retention for Chicago students who could not meet minimum test-score cutoffs of the Iowa Test of Basic Skills (ITBS) in grade 3, 6 and 8.

When examining the general trend of ITBS test scores, researchers (Bryk, 2003; Jacob, 2005) found a substantial score gain since the inception of the policy. However, according to Bryk (2003), the dramatic score gain occurred right after the announcement of the high-stakes initiatives, but started to decline after two years. By comparing the 1998 cohort with the 1994 and 1996 cohorts in terms of the proportion of students who correctly answered the ITBS items, Jacob (2005) showed that the high-stakes policy improved both basic and higher-order skills but the size of the improvement varied by subjects. In math, the improvement in basic skills was twice as much as in higher-order skills; in reading the improvement was distributed equally across item types. Jacob further contrasted the ITBS score gain in tested (i.e. reading and math) versus nontested subjects (i.e. science and social studies) – the former was about two to four times larger than the latter; in addition, the increases of math and reading scores were more significant in low-performing schools (i.e. schools had less than 20% students above the 50th percentile) than in high-performing ones while there was no such a difference in nontested
Jacob claimed that the differential effects by prior school performance were larger than those by student prior ability and suspected that responses to the accountability policy took place at the school level rather than at the individual student level. However, even so, given school prior performance, there was a considerable variation in student responses, especially in reading. Controlling for school differences, in each promotional gate grade, students whose prior achievement stood at the 10th-50th percentile were found to perform significantly better in reading than their peers who had scored either below the 10th percentile, or above the 50th percentile. In math, the differential pattern was not that obvious, but with two exceptions: one is that grade-8 students in the 26-50th percentile seemed to fare considerably better than the others in the rest distribution; the other exception is that grade-6 students in the 10-25th percentile showed more improvement than those above the 50th percentile.

Neal and Schanzenbach’s analysis (2010) of the 1996 grade-6 cohort revealed a somewhat different yet clearer pattern of the differential effects by student prior ability; despite an increase in scores at every decile, students in the third and fourth deciles of prior achievement distribution improved the most in both reading and math ITBS scores while those at the two tails of the distribution gained the least. Such a pattern was also observed on the Grade-6 IGAP (Illinois Goals Assessment Program) Math.

Roderick et al.’s evaluation (2002), however, complicated the picture of the policy effects by taking into account differences in both students and schools. They found that the differential pattern was more distinct in low performing school, typically schools that served high-proportion of at-risk students and that were subject to probation or to the threat of probation. By grouping students based on their risk of being retained, their study suggested that the differential effect
pattern would differ by grades and subjects. In reading, although moderate- and some-risk students benefited from the test-based retention policy regardless of grade levels, the policy effects on the bottom and top two groups (i.e. no-risk and high-risk) were not consistent across grades: for high-risk students, the policy appeared to be beneficial in grade 3 and 6, but ineffective in grade 8; however, for no-risk students, the policy effect was negative in grade 3, insignificant in grade 6, but positive in grade 8. In math, all grade-3 students seemed to benefit from the policy with high-risk student showing the largest score gain; nevertheless, in grade 6 and 8, except for no-risk students that showed an improvement, all other-risk students suffered from a learning loss under the high-stakes policy. The findings regarding the differential policy effects, especially the effects on grade-3 students, according to Roderick et al., are inconclusive and thus present a need to further understand the relationship between test-based retention policy and student academic performance.

**Conventional Accounts**

There is no doubt that high-stakes testing, including test-based retention policy is a well-intentioned endeavor to better education quality. However, my review of the past evidence above has shown that the policy effects on student academic performance is not as linear and as rosy as policymakers have anticipated. Previous literature has documented two potential explanations for the discrepancy between the intended and observed learning outcomes: one focuses on student motivation (e.g. Roderick et al, 2002) and is framed by educational motivation theories such as the self-determination theory (Deci & Ryan, 1985) and the goal theory (Ames, 1992); the other explanation looks into instructional responses to incentives (e.g. Jacob, 2005; Neal & Schanzenbach, 2010) and is premised on economic incentive theories of principal-agent relationship (Holmstrom, 1979; Holmstrom & Milgrom,1991; Stiglitz, 1987).
**Student motivation and motivation theories.** Student motivation and educational motivation theories, in certain contexts, may help interpret differential and long-term academic performance under high-stakes testing. Motivational goal theory (Ames, 1992) suggests that high-stakes policy may differentially affect students’ motivation. As most current high-stakes tests focus on basic skills and impose uniform standards, the goal of passing the tests, though not creating much challenge for high-ability students, may seem too difficult to attain for lowest-ability students. As demonstrated in Roderick and Engel’s study (2001) of 6th- and 8th-graders under the Chicago test-based retention policy, the policy motivated most of the students to work hard, but could hardly influence the high-ability students; more importantly, even with a desire to avoid sanction, the students at the lowest ability level were unable to translate the desire into substantial work. Self-determination theory (Deci, Koestner, & Ryan, 1999; Deci & Ryan, 1985; Ryan & Deci, 2000) implies that high-stakes testing, which heavily relies on the power of external incentives, promotes short-life extrinsic motivation at the expense of long-term intrinsic motivation (also see discussion by Jones, Jones, & Hargrove, 2003; Kohn, 1993; Madaus & Clarke, 2001). In the long run, due to reduced intrinsic motivation and lack of continued external encouragement, students may withdraw effort from their future study after high-stakes tests.

There are two major limitations in using students’ motivation to explain their academic performance under high-stakes testing. First, student motivation is a complex construct that depends on many other factors in addition to high-stakes incentives. Focusing on students’ motivational responses may overlook potential differences in teachers and schools and may introduce much uncertainty in predicting students’ learning pattern in high-stakes context. Second, it is unclear whether and how high-stakes testing influences student motivation in early or middle elementary grades. When presenting their theoretical framework for the effects of
Chicago high-stakes testing, Roderick et al. (2002) argued that younger students may be less sensitive to incentive threats and less capable of shaping their own learning and/or overcoming barriers through efforts. In addition, some educational psychologists (e.g. Dweck, 2001; Stipek, 2002) informed us that the motivation of young children may not exactly reflect their learning ability. I contend that motivational theories may not be able to serve as an explanation for grade-3 students’ short-term academic performance, but it may provide some clues to their long-term learning trend as students develop their intrinsic motivation along their academic pathways. Nevertheless, considering the two limitations, cautions need to be taken when we apply the motivational theories to the study of high-stakes testing.

**Instructional response and economic incentive theories.** One can also seek explanations by examining instructional responses to high-stakes testing. Accountability policies hinge on the implementation by teachers (O’Day, 2004; Cohen, 1996a; Elmore, 2008). Teachers and their teaching practices play an important role in students’ academic life and directly determine learning opportunities in school (Nye, Konstantopoulos, & Hedges, 2004; Palardy & Rumberger, 2008; Rowan, Chiang, & Miller, 1997). It has been well documented that for better or for worse, high-stake testing could shape classroom instruction (see review by Au, 2007, 2009; Herman, 2004; Jones et al., 2003; NBETPP, 2003). Because a teacher’s success is often defined by student performance (Cohen, 1996a; Finnigan & Gross, 2007; Johnson, 1986; Lortie, 1975), even when there is no explicit sanctions or rewards for teachers, the policy (e.g. test-based retention policy) can still influence teaching behaviors with the high-stakes consequences for students (Diamond, 2007; Hanushek & Raymond, 2005; Jacob, Stone, & Roderick, 2004).

Educational researchers (Koretz & Hamilton, 2006; Lindquist, 1951) have long realized that high-stakes incentives can shift teaching attention to the immediate goal of improving test
performance. From the perspectives of economic incentive theories (Baker, 1992; Holmstrom, 1979; Moe, 1984, 2003; Stiglitz, 1987), such shifts are a predictable result since teachers are agents of student performance and are evaluated based on test scores with high-stakes consequences for their students and/or for themselves. Holmstrom and Milgrom’s multitask principal-agent model (1991) asserts that when one engages in a complex job involving various tasks and when incentives for performance are given based on limited dimensions of output measures, agents tend to focus their effort on the task that has greater potential to improve their own performance rating and disregard the other tasks that offer relatively small incentives. This model suggests that under high-stakes testing, teachers have a tendency to reallocate resources across/within curriculum and/or among students in order to boost up average test performance. Quite a few studies (see review of Au, 2007; 2009) revealed that teachers traded off between tested and nontested subjects under high-stakes regime, and assigned more instructional time to tested subjects (e.g., McNeil & Valenuela, 2001; Shepard & Dougherty, 1991; Stecher & Barron, 2001; Taylor, Shepard, Kinner, & Rosenthal, 2001); they tended to drill students more on tested skills that were relatively easier to improve and to downplay specific skills that were not emphasized by the test (e.g., Abrams, Pedulla, & Madaus, 2003; Barksdale-Ladd & Thomas, 2000; Koretz, Barron, Mitchell, & Stecher, 1996; McNeil, 2000; Shepard & Dougherty, 1991); teachers might also prioritize marginal students whose expected performance were close to the passing threshold and who were most likely to improve the aggregate test scores (Booher-Jennings, 2005).

The economic theories of principal-agent relationship are useful for understanding the phenomena of shifted teaching focus and resource reallocation in preparing for high-stakes tests. They as well provide an insight into the potential distribution of the policy effects on student
academic performance. However, the theories seem unable to reconcile the inconsistency between the previous findings, for example, why students of the similar ability and in the same policy context exhibited different learning outcomes in different studies (e.g. Jacob, 2005; Neal & Schanzenbach, 2010; Roderick et al. 2002). Also, it remains unanswered why the instructional changes under various high-stakes initiatives were more often found in teaching focus and content than in pedagogy – even if high-stakes testing appeared to improve teaching methods, the changes were limited to a superficial alignment with test format and did not add to the complexity and depth of instructional content and pedagogy (e.g. Diamond, 2007; Firestone, Mayrowetz, & Fairman, 1998; Grant, 2001). Furthermore, by assuming a universal differential pattern of instructional responses by tasks and students, the economic theories seem to diagnose all teachers as intentionally undermining the policy implementation process or simply not mobilized to exert effort to improve student learning. This perspective, in my opinion, has oversimplified teaching behaviors and more importantly, failed to address potential variations in instructional practices and capacity under a high-stakes regime.

**Variations in Instructional Practices and Capacity: A Supplementary Account**

According to Fuhrman (2001), variations are a typical response to policy change. Here I consider the variations in instructional practices and capacity under high-stakes testing as a supplementary account for the extant theories, especially for the economic incentive theories that address changes in teaching behaviors. I see instructional capacity not only as an important source of such variations, but also as a moderator of the policy effects. I argue that although high-stakes testing shifts teaching attention to test scores, the incentive itself does not improve instructional capacity and hence is not able to advance pedagogy and produce dramatic improvement in overall student academic performance; but teachers’ test preparation practices
will be shaped by their existing experience, belief, knowledge, and as a result, will lead to
diverse learning pattern.

**Variations in Instructional Practices**

As Cohen and Hill (2001) commented on the California instructional reform, teachers are
not educational clerks who either follow or disobey orders; rather, they should be viewed as
“active players who judge and use policy, and make policy as they respond to others’ initiatives”
(p. 85). Broader literature on school reform and teacher change (e.g. Cohen & Hill, 2000; EEPA,
1990; Jennings, 1996; Spillane, 2001; Spillane & Jennings, 1997) told us that instructional
decisions, though to a certain extent constrained by social and organizational conditions, were
not uniform across schools and teachers. Coburn’s investigation of the teaching experiences of
three California elementary teachers (2004) has also shown that teachers who were facing similar
instructional challenges reacted to external pressures in various ways: they might transform their
existing knowledge structure to accommodate new information or experiences, provide symbolic
responses that were decoupled with the instructional core, create two or parallel instructional
approaches that corresponded with different pressure or priorities, assimilate new knowledge and
experience to fit existing schemas or practices, or even reject approaches that were not congruent
with their beliefs or preexisting instructional approach; the choice of action varied from teacher
to teacher and did not solely depend on the institutional environment.

When it comes to instructional practices under high-stakes testing, despite a prevalent
research trend in seeking a common response pattern (see review by Au, 2007; Herman, 2004;
Jones et al., 2003), previous study results have displayed variations in teachers’ test preparation
behaviors. For example, Diamond (2007) noticed that the Chicago policy influenced the
instructional content for a large proportion of teachers but did not completely dominate all
teachers’ decision making; there was a considerable variation in teaching strategies and teacher-students classroom interaction under the policy. In a study of the Kentucky Instructional Results Information System (KIRIS), Koretz et al. (1996) found that with overwhelmingly reduced emphasis on nontested areas, less than half of the teachers focused a great deal on using KIRIS-like tasks in regular instruction; 40 percent of the teachers reported increasing emphasis on the content alignment between instruction and the KIRIS tests; but about 58 percent reported paying more attention to general instructional improvements. According to Monfils and colleagues (2004), instructional responses were not uniform toward the New Jersey testing policy. There were two distinct test preparation approaches under the state high-stakes testing policy: one is an embedded inquiry-oriented approach that incorporates test preparation into the regular curriculum and reflects some ideas about national reforms in mathematics and science; the other, in contrast, is a direct, decontextualized test preparation that consists of special ‘cramming’ before test administration and/or intensification of conventional didactic instruction. Monfils et al. discovered that while teachers in poor school districts were more sensitive to the state policy than their colleagues in wealthier districts, within the poor districts, teachers still exhibited the divergent teaching approaches. Grant’s portrait of two social studies teachers under New York Regents testing program (2001) further unveiled that even with the similar academic background, similar classroom composition, and from the same school, teachers would adopt different teaching approaches and respond differently to the high-stakes testing. One teacher, though did not give explicit classroom attention to the high-stakes test, implicitly aligned his instruction content with the test and used a teacher-centered approach to faithfully follow each chapter of the textbook; the other teacher, however, was more student-centered; she not only designed various activities to actively engage students during most of the instruction, but also
accommodated the demand of the test by leaving out a proportion of class periods for practicing on test items and teaching test-taking strategies only.

McLaughlin (1990) stressed that in policy implementation, “local variability is the rule; uniformity is the exception” (p. 13). To understand the nature of variations and the potential consequences on student learning, it is necessary to unpack the sources of the variations and to examine their potential role in explaining the effects of high-stakes testing.

**Instructional Capacity as a Source of Variations**

When summarizing a series of studies on the effects of standards-based reform in the yearbook *From the Capitol to the Classroom: Standards-based Reform in the States*, Fuhrman (2001) pointed out that given an accountability design, the variations in policy responses mainly stem from differences in capacity at all levels of the accountability system. In the midst of diverse definitions of capacity (e.g. Corcoran & Goertz, 1995; Ford 1992; Newmann, King, & Rigdon, 1997; Spillane & Thomas, 1997; also see review of Adams, Jr. & Kirst, 1999), Elmore (2008) concluded that “capacity is defined by the degree of successful interaction of students and teachers” (p. 118); and “capacity both inheres in teachers and students and it comes to them from external sources” (p. 222). By centering the definition of capacity around students and teachers, Elmore zoomed in on the within-school instructional capacity and emphasized it as a key determinant for school improvement and as a major explanation for variability in policy responses between and within schools.

Previous research on high-stakes testing, though often lacked explicit attention to instructional capacity, has suggested several characteristics that can differentiate teachers’ responses to high-stakes testing. Most of them are concerning collective school properties, such as average SES status, student composition, aggregated prior performance, and/or collective
expectations. When focusing on school SES status or student composition (e.g. Diamond, 2007; Firestone, Schorr, Monfils, 2004), an underlying belief is that schools from a poor neighbourhood and/or with a high-concentration of disadvantaged students may suffer more from insufficient educational resources and may face more challenges in maintaining classroom order, breaking culture barriers, and/or retaining high-quality teachers. It is found that these schools tended to adopt didactic instruction rather than inquiry-oriented instruction. With a similar rationale, some studies as well took into account differences in school aggregated performance (e.g. Brown & Clift, 2010; Diamond & Spillane, 2004; Jacob, 2005; Roderick et al., 2002; Springer, 2008). While attending to the association between school performance and school SES status and racial composition, the researchers also agreed that teachers in low-performing schools are under greater pressure to improve student test performance and may be demoralized by public labels of failure; thus the teachers may resort to more extreme ‘teaching to test’ practices; but it is debatable whether high-performing schools are associated with more positive instructional responses (Diamond & Spillane, 2004) or will remain unchanged in an attempt to maintain the current performance status (Brown & Clift, 2010).

Another important characteristic that emerged from the past studies (e.g. Hess, Jr., 1999; Valli, Croninger, Chambliss, Graeber, & Buese 2008) is school shared expectation, for example, a collective sense of what constitutes quality teaching and adequate student learning as well as a shared belief of teaching capability and student learning ability. A higher shared expectation toward teaching and learning signals a higher level of teacher commitment and a better understanding of policy and practices as well as a healthier professional community that facilitates communication and learning between teachers (Elmore, 2008). It also predicts a greater level of school achievement (Hoy, Sweetland, & Smith, 2002; Lee & Smith, 1996). With
a high shared expectation, schools and teachers will be able to respond constructively to the external accountability policy; however, with low and/or incoherent expectations, accountability for student learning will devolve to an individual sense of responsibility; instructional improvement then becomes a personal choice and is subject to individual teachers’ own skills and will (Ablemann, Elmore, Even, Kenyon, & Marshall, 1999; Debray, Parson, & Woodworth, 2001).

All the school-level characterises identified above are important and necessary attributes of instructional capacity (Cohen & Moffitt, 2009) and are helpful in explaining variations in school responses to high-stakes testing. However, they are not sufficient accounts for differences between teachers. We know that teachers with similar school resources and similar students may nonetheless act differently in classroom instruction (e.g. Gant, 2001; Spillane, 2001) and may produce different learning outcomes (Konstantopoulos, 2006). Access to conventional school resources, such as books, facilities, class size, and time, will make a difference in teaching and learning only when individual teachers use and notice the resources (Cohen et al., 2003). Organizational structure and environment are influential in teachers’ work, but how they influence the work depends in part on what individual teachers make of them (Fairman & Firestone, 2001; Spillane, 2001). Gant (2001) showed us that teachers’ personal beliefs of subject matter and student ability shaped their instructional responses to New York Regents testing program (2001). Rex and Nelsons’ ethnography of two high school English teachers (2004) demonstrated that in the classroom practices, the teachers integrated their beliefs to accommodate school-wide test preparation for the Michigan Educational Assessment program; the beliefs include their senses of responsibility for students, personal views about teaching purposes and schooling, perceptions of subject matter, and self-evaluation of own agency.
Diamond’s study of the Chicago high-stakes testing program (2007) also revealed that teachers drew on their prior experiences and beliefs as well as interactions with their teaching colleagues and administrators for making instructional changes under the policy.

Cohen and Moffitt (2009) pointed out that capacity arises not only from organizations and environments, but also from individual attributes, including teachers’ values, interests, dispositions, skills and knowledge; these attributes interact with each other in influencing teachers’ practices. From the teaching and learning perspective as suggested by Elmore (2008), I consider instructional capacity in the context of high-stakes testing as an integrated construct that is specific to individual teachers and that can well predict and differentiate their effort for achieving the policy goal, i.e. improved student learning. Focusing on instructional capacity of individual teachers echoes the findings from the implementation of instructional reform in California (e.g. EEPA, 1990), Michigan (e.g. Jennings, 1996) and South Carolina (e.g. Spillane, 2001); that is, policy successes relied on “street level bureaucrat” (Lipsky, 1980) or individuals who were responsible for delivering policy at the ultimate end of the system (McLaughlin, 1987); the specific way in which policy was digested and acted on in the classroom was framed by teachers’ inherent knowledge, belief and practice (Cohen & Ball, 1990).

**Instructional Capacity as a Moderator of Incentives Effects**

Previous research told us that not every teacher believes in student potentials and knows how to effectively bring out desired learning changes (e.g. Kannapel, Aagaard, Coe, & Reeves, 2001; Spillane, 2001). The effects of various incentives that target at improving student learning are contingent on the capacity of individuals who mediate the messages that the incentives carry (Cohen, 1996a, 1996b; Elmore, 2008). I view instructional capacity as a moderator, but NOT a mediator of incentive effects. In other words, the effects of high-stakes incentives on teaching
and learning may be modified by instructional capacity, but are unlikely to occur through enhancing teachers’ capability to teach. My perspective is consistent with the literature on policy implementation and sensemaking. Although sensemaking researchers (e.g. Coburn, 2001, 2004, 2006; Spillane, 1999, 2002; Spillane & Callanhan, 2000; Louis, Febey, & Schroeder, 2005) emphasize less on individual traits than on social interaction processes in which teachers interpret and act on policy messages, instructional capacity (“existing cognitive structures” and “situation” as Spillane et al. termed, 2002, p. 388), in their framework, is still considered as an important factor in shaping the process. It is believed that extensive opportunities for professional communication and learning are necessary in order to improve teachers’ knowledge and skills and/or to change their ingrained values and beliefs. I contend that such opportunities unfortunately are not a part of the schema of high-stakes testing. As Cohen (1996a, 1996b) and Elmore (2003, 2008) reminded us, high-stakes incentives function to mobilize existing capacity, but not produce new capacity. Although sometimes test-based incentives were accompanied with other policy initiatives for professional development (see review of Berry, Turchi, Johnson, Hare, Owens, & Clements, 2003), there is little evidence supporting that performance awards/sanctions themselves are the causes of these initiatives and/or can help improve teaching capability, especially teaching skills and beliefs.

It is well established that instructional capacity is an important determinant of whether teachers will be able and willing to challenge deep-seated norms, to acquire new ideas, and to renovate their teaching methods (Cohen & Hill, 2001; EEPA, 1990; Jennings, 1996; Spillane, 2001). The limited power of high-stakes testing in influencing instructional capacity probably explains why the policy was found ineffective in advancing teachers’ pedagogy in many studies (e.g. Diamond, 2007; Firestone, Mayrowetz, & Fairman, 1998; Grant, 2001). It also suggests that
the policy may not be able to produce dramatic changes in student overall learning. We have ample evidence to confirm that high-stakes testing can push teachers to invest more instructional time in some tested areas (e.g. Jacob et al., 2004; Koretz et al., 1996; Taylor et al., 2001). But it does not mean teachers know how to efficiently use the time and/or to respond constructively to performance awards/sanctions (Cohen, 1996a). Without any improvement in teachers’ pedagogy, the increased instructional time may not necessarily be translated into effective student learning.

Nonetheless, high-stakes testing may still exert effects on teaching focus and content through mobilizing resources and values. This perspective has been well supported by previous research (see review by Au, 2007; Herman, 2004; Jones et al., 2003) and by economic theories of principal-agent relationship (Holmstrom, 1979; Holmstrom and Milgrom, 1991). To supplement the theories, I argue that organizational and external influences may play a part in shaping the mobilization process, but individual teachers’ will and skills are as well influential. While high-stakes incentives draw their attention to student performance, teachers may also resort to their own belief, knowledge and skill to decide what to teach and how to teach. If teachers want to emphasize skills that can improve average test performance, their perceptions of student learning ability and of their own teaching capability then will come into play in the choice of teaching content. Similarly, if high-stakes testing push teachers to prioritize marginal students who are at risk of failing the tests but who could avoid the failure with reasonable investment of teaching and learning, teachers may need to use discretion in identifying and providing support to this group of students and in choosing whether to sacrifice the learning of other students. The instructional decisions that teachers made according to their own instructional capacity will lead to various differential learning patterns and will inevitably put certain students and/or certain aspects of learning at a disadvantage. Unlike what was implied by the economic incentive
theories, I see the potential learning loss not as a result of teachers’ resistance to and/or intentional undermining of high-stakes testing, but rather as arising from teachers’ own capacity constraint.

Although previous studies have generated some interesting findings regarding differential learning under high-stakes testing, instructional capacity of individual teachers has rarely been included in their theoretical framework. If highlighting the instructional capacity as a moderator of the policy effects, we can anticipate incentive effects on teaching practices become relatively more predictable; as a result, we may be able to observe a clearer pattern between high-stakes testing and student academic performance.

**Grade-3 Test-based Retention Policy and Teacher Expectations: A Situated View**

This study highlights teacher expectations as an indicator of instructional capacity and evaluates how it modifies the effects of grade-3 test-based retention policy. I present and justify my research hypotheses following a brief review of the literature on teacher expectations and discussion of specific concerns related to the evaluation of teaching, learning and student motivation in the third grade.

**Teacher Expectations**

Teacher expectations are defined as “inferences that teachers make about the future behaviour or academic achievement of their students, based on what they know about these students now” (Good & Brophy, 1997, p. 79). There are three major types of teacher expectations (Cooper & Tom, 1984): (a) beliefs of students’ general academic competence, (b) prediction of how much progress students will make over a specified period of time, and (c) the degree to which teacher over- or under-estimate students’ present performance level. In this study, I focus on the first type of teacher expectations because teachers’ negative beliefs of
student general academic ability are often considered as one major obstacle to school reform.

Educational policy researchers (e.g. Fuhrman, 2001; Hess, Jr. 1999; Porter, Archbald, & Tyree Jr., 1990; Spillane, 2001) believe that when teachers do not expect that their students can learn, they would not have a strong sense of responsibility and would not try hard to improve instruction and/or to make challenging curricular content accessible for all students.

**Construct of teacher expectations.** Teachers form their expectations based on various sources of information in relation to students’ social and academic characteristics (see review by Good, 1987; Jussim, Robustelli, & Cain, 2009; Schunk, Pintrich, & Meece, 2008). Teacher expectations, though are most strongly predicted by students’ actual ability, are not always accurate. Researchers (e.g. Diamond, Randolph, & Spillane, 2004; McKown & Weinstein, 2008; Ready & Wright, 2011) found that classroom and school contexts are a major source of the inaccuracy: teachers in low-SES status and low-achieving contexts tend to underestimate their students’ ability. In addition to classroom and school characteristics, certain attributes of individuals may also account for teachers’ perception of students’ learning ability. One such attribute may be teachers’ knowledge and skills. According to Ready and Wright’s Analysis (2011) of the Early Childhood Longitudinal Study Kindergarten Cohort (ECLS-K), after student composition is controlled for, new teachers on average tend to overestimate students’ literacy skills over the course of a school year; and compared to those with only a BA degree, teachers with graduate degrees are more likely to overestimate the skills of Black students.

It is also generally accepted that teacher expectations are made up of beliefs and actions based on those beliefs (Good & Brophy, 1997) though literature is not well established regarding the composition of the beliefs. Several studies (e.g. Tiedemann, 2002, van den Bergh et al., 2010) suggested that given student characteristics, teachers’ perceptions of student learning
ability are influenced by their inherent beliefs of ethnicity and/or gender differences. In addition, Hughes et al. (2005) found that controlling for students’ actual abilities, parent education level, and child ethnicity and gender, teacher expectations are predicted by teachers’ perception of student-teacher support (i.e. perceived closeness between teacher and students) and by teachers’ perception of parent-teacher alliance (i.e. their comfort with parents). Teacher expectations were also observed to have a stronger relationship with teachers’ perceived parent-teacher alliance than with teachers’ self-report of parents’ actual involvement. This observation seems to imply that teacher expectations may not be fully explained by external factors and may as well be associated with some personal attributes of the teachers, for example, their self-efficacy (i.e. the perception of their own capability in bringing out desired student outcomes). Teacher expectations may be related to teachers’ self-efficacy because if teachers do not hold positive expectations of their students’ learning ability, they will hardly believe that their teaching effort will lead to better educational attainment by the students (Mohrman & Lawler III, 1996).

Timperly and Philips (2003) justified the relationship between teacher expectations and self-efficacy based on their similarity in influencing teachers’ goals for students, the effort teachers invest in teaching, teaching behaviors in classroom, and student achievement outcomes; there was evidence that both teacher expectations and self-efficacy can be changed through professional development; and the change in one perception seem to be accompanied by the change in the other. Hoy, Hoy, and Kurz (2008) also discovered that individual teachers’ estimates of students’ future academic performance are positively associated with their academic optimism, a general construct that encompasses teachers’ self-efficacy, their trust in parents and students, and the sense of academic emphasis (i.e. the degree to which teachers find ways to engage students in appropriate academic tasks, Beard, Hoy, & Hoy, 2010). Although more
evidence is needed to confirm the relationship between teacher expectations and the academic
optimism or the attributes of the academic optimism, my review of literature above suggests that
teachers’ perception of students’ academic competence is shaped not only by classroom and
school contexts, but also by some important attributes of individual teachers such as teachers’
knowledge and skills, their beliefs of gender/ethnicity differences, and their self-efficacy. I
therefore consider teacher expectations as a proxy indicator of teachers’ instructional capacity
though the latter is a quite broad concept that involves much more attributes than those
embedded in teacher expectations.

**Teacher expectations and student academic performance.** In their seminal study
*Pygmalion in the Classroom*, Rosenthal and Jacobson (1968) concluded that teacher expectations
can act as self-fulfilling prophecies; artificially high expectations can lead to enhanced student
performance. For the past over 40 years, the study has sparked an extensive body of research and
debate regarding the effects of teacher expectations on student outcome in the field of
educational psychology. Although researchers agree that teacher expectations can and sometimes
do affect students and teacher-student interaction, there is hardly a consensus on the actual power
and prevalence of the expectation effects; but it seems that teacher expectations are more likely
to exert influence in subjects areas that allow variations in instructional styles (e.g. reading) (see
review by Copper & Tom, 1984; Good, 1987; Jussim et al., 2009; Rosenthal, 1991; Schunk et
al., 2002). Despite focusing on individual students as the targets of teacher expectations in most
psychological research, a series of studies by Rubie-Davies (2006, 2007; Rubie, 2004) suggested
that teacher expectations of student learning may be class-centered, that is, teachers may have
high or low expectations for all students in their classroom. She also identified considerable
instructional differences associated with different teacher expectations: compared to teachers
with low expectations, teachers with high expectations were more aware of student learning needs and more likely to improve students’ self-perceptions of academic competence. But Rubie-Davies also found that high expectations are not always associated with positive teaching practices that emphasize on motivating and engaging students; nor do they necessarily lead to improved student academic performance. However, because Rubie-Davies defined teacher expectations as discrepancies between teachers’ beliefs of student competency and student actual performance, the generalizability of her findings to the current study remains a question.

Spillane’s (2001) investigated how teacher expectations of student academic ability would influence teaching practices in the context of North Carolina’s instructional reform. He found that teachers who did not believe in student learning ability tended to view learning as a linear and hierarchical process where mastery of basic skills should precede the learning of more intellectually challenging material; these teachers deprived students of opportunities to study advanced concepts and skills and spent most of the classroom time in emphasizing basic skills despite new state instructional standards. Although low expectations were a prevalent problem among teachers with a large proportion of African American students and/or low SES students, Spillane discovered two teachers (i.e. one math teacher and one reading teacher) with similarly disadvantaged students, but with a distinct understanding of teaching and learning. The teachers did not underestimate students’ learning ability and instead considered intellectually rigorous and engaging instruction as essential for the students’ success. The math teacher did not drill basic skills as a prerequisite for understanding difficult mathematics concepts; and the reading teacher provided students with both essential reading skills and literacy experiences to help make sense of the texts they read. Spillane’s study offered us some clues regarding how negative expectations would obstruct the implementation of policy initiatives and delineated a very
promising picture of improved instruction under new visions of teaching and learning. But it is not known whether the observed effective teaching approaches can be attributed only to teachers’ positive expectations of student learning ability, or as well to their professional development opportunities and/or their different perception of subject matters.

Cohen et al. (2003) theorized that teachers tended to calibrate instruction to their views of students’ capability. It is likely that teachers with negative expectations of student average learning ability, same as shown in Spillane’s study (2001), will focus more on basic skills while teachers with positive expectations may provide access to more advanced knowledge. Existing research has little to tell us about the teaching and learning consequences of this potential instructional calibration for students in general and more importantly, for students of different academic ability. A few quantitative large-scale studies (e.g. Lee & Smith, 1996; Palardy & Rumberger, 2008) treated teacher expectations of student learning ability as one dimension of a larger construct of teacher beliefs and appeared to suggest that higher expectations may lead to higher student achievement. However, we do not know whether the same class-level expectations will have different implications for students at different ability levels. Neither do we know how the expectation effects will vary depending on the school context (e.g. schools with or without high-stakes testing). Also, besides teachers who hold positive or negative expectations of student learning ability, there may be teachers who do not know and/or do not care about whether students can learn or not (i.e. indifferent-expectation teachers). So far no research has ever examined this group of teachers and their influence on students’ learning.

**Teaching, Learning, and Student Motivation in Grade 3**

To evaluate the impact of grade-3 test-based retention policy, it is necessary to consider relevant teaching characteristics and developmental concerns specific to grade-3 students.
**Subject matters.** Previous literature mostly agrees that student performance in tested and nontested subjects is unlikely to be the same. In addition to this common view, my review of the existing evidence also unveiled considerable differences between tested subjects, including grade 3 reading and math (e.g. Roderick et al., 2002). One explanation is that reading in the elementary grades focuses largely on activities that are not specifically geared to particular test items while math skills are more specific and more subject to test preparation; another explanation is that reading performance is under the influence of many factors outside schools (see discussion by Jacob, 2005). Spillane’s study (2005; also see Burch & Spillane, 2003) of elementary school leadership practices suggested an extra explanation: school leaders and teachers in primary grades may have different visions about reading and math. Primary-grade teachers do not have well-defined subject matter specialties and school support to establish any subject identities. But Spillane discovered that elementary school leaders often viewed reading/literacy as a subject that pervades the entire curriculum as well as an important measure of student and school progress. The leaders emphasized teachers’ professional development and participation in school decision-making about literacy. In contrast, math was also considered as a priority, however, mainly due to the emphasis of external policies (e.g. testing program). The school leaders tended to view math instruction as sequential and directed most of the attention to math textbooks and teachers’ implementation of the textbooks without much internal professional support. While teachers had more integrated advisory networks for literacy instruction, math instruction mostly consisted of isolated practices and depended on curricula and training provided by external programmes. Therefore, I suspect that under the grade-3 test-based retention policy, teaching to test is more pronounced in math than in reading and teachers’ individual instructional capacity may exert more influence in students’ math performance than in their reading performance. The
instructional capacity may also have considerable effects on students' science learning because teaching of nontested subjects is unlikely to receive as much attention and as much professional communication as that of reading. In addition, it should be also noted that most primary-grade teachers are responsible for multiple subjects; if teachers have developed certain expectations about students' learning ability, such expectations may be spread out to all subjects that the teachers are teaching, including subjects that are not to be tested in high-stakes exams.

**Student self-perceived competence and interests.** Another concern related to the evaluation of grade-3 test-based retention policy is student motivation. Instead of student motivation in general, this study focuses on student self-perceived competence and interests. Student self-perceived competence plays an important role in the development of student motivation. It predicts whether students will seek and enjoy challenges in learning and be resilient in the face of setback (Dweck, 2001). Self-perceived competence is inherently related to personal interests. Many researchers (e.g. Harter, 1992; Mac Iver, Stipek, & Daniels, 1991; Skaalvik & Rankin, 1995) have provided evidence that higher interests are correlated with higher self-perceived competence; students will not enjoy learning in a specific subject if they do not feel competent in that subject. The role of self-perceived competence and interests has seldom been examined in the previous research on high-stakes testing.

According to reviews by Stipek (2002) and Dweck (2001), grade-3 students are in the process of developing awareness that ability is an internal and relatively less observable quality. The students are able to perceive their ability based on teachers’ feedback and parents’ appraisal and to some extent, through social comparison. They may exhibit helpless motivational patterns and lose confidence when encountering a series of salient failure or criticism (Dweck & Master, 2009). But they tend to credit their academic performance to efforts instead of ability due to an
undifferentiated view between these two constructs (Nicholls, 1978, 1984; Nicholls & Miller, 1984). Therefore, grade-3 students’ self-perceptions may not be a strong predictor of their short-term academic performance. Stipek (2002) pointed out that the development of students’ self-perceptions of their academic ability can be partly explained by the between-grade differences in organizational, instructional and evaluation practices. Students in early or middle elementary grades are often protected by their teachers from academic competition and receive little information to challenge their optimistic views of own academic competencies. Stipek reasoned that with increased tasks and evaluation practices that stress on a uniform task structure, these students will have more differentiated view of ability and efforts; and there will be more within-classroom variations in academic self-perception because the same tasks in the same classroom make the performance more comparable and more salient than tasks that vary and/or that are individualized. However, no research has ever empirically confirmed this theory. Neither is there any literature addressing the consequences of this premature formation of ability conception.

**Grade-3 Test-based Retention Policy and Teacher Expectations: Research Hypotheses**

The review of teacher expectation and the discussion of grade-specific characteristics finally lead to my research hypotheses regarding the relationship between and the effects of grade-3 test based retention policy and teacher expectations. As shown in my summary of the past evidence, the effects of test-based retention policy have been under-examined. Although several researchers have conducted sophisticated analysis of the Chicago policy effects, their findings, including those related to grade-3 students, still remain inconclusive and are limited in the local context. Moreover, none of the researchers has attended to the variations in teachers. In this study, I emphasize individual teachers’ instructional capacity, in particular, teacher expectations when examining the effects of the test-based retention policy. Corresponding to the
four sets of the research questions that I raised in the first chapter, here I present the hypotheses of the relationship between the policy and teacher expectations and of their joint effects based on the extant economic and motivational theories and with extra attention to the moderating role of teacher expectations.

1. **Relationship between the policy and teacher expectations**

   Incentive effects are limited to mobilizing rather than enhancing capacity (Cohen, 1996a, 1996b; Elmore, 2003, 2008). Since high-stakes incentives themselves do not provide extra resources for teachers to improve their professional capability, I hypothesize that the retention incentive attached to grade-3 student test performance would not change teachers’ existing beliefs of student learning ability.

2. **Policy-by-teacher effects on instructional time allocation**

   Economic incentive theories (Holmstrom, 1979; Stiglitz, 1987) suggest that high-stakes testing can shift teacher attention to the immediate goal of boosting up test scores. Although reading and math are both the focuses of the current accountability system, test preparation may be more pronounced in math than in reading because teachers often lack within-school professional support for math teaching yet are under great pressure to improve students’ math scores (Spillane, 2005). I therefore predict that the test-based retention policy will push all teachers to allocate more time in math.

3. **Policy-by-teacher effects on student academic performance**

   There are four aspects of student academic performance related to the policy-by-teacher effects: student overall academic performance, student mastery of cognitive skills, student differential academic performance by prior ability, and student long-term academic performance. The policy effects on students taught by teachers with indifferent expectation are hard to predict
given the scarce literature on indifferent teacher expectation. I will investigate school, teacher, and classroom characteristics associated with indifferent teacher expectation (see the section Measures of Chapter 4) before conducting further causal analysis.

**Student overall academic performance.** As previously discussed, incentives are unlikely to improve instructional capacity and teachers’ pedagogy (see the section Variations in Instructional Practices and Capacity: A Supplementary Account of the current chapter). Access to conventional school resources, such as instructional time, is a necessary but not sufficient condition for instructional improvement (Cohen et al., 2003). Therefore I hypothesize that during the promotional gate year (i.e. grade 3), the test-based retention policy may not produce a significant gain in student overall learning in tested subjects; but may hurt learning in non-tested subjects as a result of reduced educational resources, such as time and/or attention in these subjects.

**Student mastery of cognitive skills.** Holmstrom and Milgrom’s multitask principal-agent model (1991) told us that high-stakes testing can drive teachers to focus their efforts on teaching of the skills that have the greater potential to boost up student average test scores and neglect the other skills that contribute less to the test results. I reason that teachers will judge which skills are essential for improving student test performance in light of their own perceptions of student learning ability. Teachers who have positive expectations of students’ academic competencies (i.e. positive-expectation teachers) will respond to the test-based retention policy by introducing more advanced knowledge and skills and probably by deemphasizing basic skills that they believe most of their students have mastered. On the contrary, teachers who have negative expectations (i.e. negative-expectation teachers) will narrow the curriculum and ignore advanced skills because they perceive their students as incapable of learning advanced skills. Rather, they
tend to believe that teaching less advanced skills is essential for improving the average test performance. Considering teaching practices as a mediator through which teacher expectations may influence student learning, I therefore establish my hypothesis regarding students’ skill learning as follows: in grade 3, the test-based retention policy may generate improved learning in the advanced skills with positive-expectation teachers; it may decrease learning in the same skills but increase learning in other lower-level skills with negative-expectation teachers; the effects pattern will be more prominent in math where skill learning is often viewed as hierarchical and sequential (Spillane, 2005).

**Student differential academic performance by prior ability.** It is also suggested by the multitask principal-agent theory that high-stakes incentives may direct more attention to marginal students who contribute the most to the average test performance and thus create differential learning pattern among students of different learning ability. I add that those teachers may have to define marginal students and/or to intentionally or unintentionally differentiate students depending on their expectations of student academic competencies; also, as primary-grade teachers are often responsible for multiple subjects, their decisions on how to treat students may be consistent across tested and nontested subjects. In the case of the grade-3 test-based retention policy, positive-expectation teachers may not explicitly prioritize any students. However, being passionate or optimistic of student learning, their emphasis on advanced skills and de-emphasis on basic skills may benefit average-ability students but may put the lowest-ability students at a certain disadvantage. Negative-expectation teachers, unlike positive-expectation teachers, will have to differentiate students as they perceive limited educational resources and difficulties in managing a high proportion of ‘low-ability’ students. In order to raise average test scores, they may define marginal students as those with lower than average
ability and may direct more resources to this group of students. However without improvement in the teachers’ pedagogy, the increased resources to this group of students may not be necessarily translated into improved learning, but will sacrifice the learning opportunities of the students at the higher end of the ability distribution. In short, my hypothesis is that, with positive-expectation teachers, the test-based retention policy will have a positive influence on the learning of average-ability students, and probably will have negative influence on the learning of the lowest-ability students; with negative-expectation teachers, the policy will negatively affect the learning of the average-ability students, but may have zero influence on the lower-than-average-ability students.

**Student long-term academic performance.** Motivational self-determination theory (Deci, Koestner, & Ryan, 1999; Deci & Ryan, 1985; Ryan & Deci, 2000) suggests that high-stakes testing would hurt students’ long term intrinsic motivation and hence will be detrimental to their future learning. I argue that this depends on what type of teachers the students encounter in the promotional gate grade. I suspect that positive-expectation teachers, given their confidence in student ability, may stress less on high-stakes tests and resort less to perverse test preparation practices in the classroom; and as a result may protect students’ intrinsic motivation. In contrast, negative-expectation teachers will practice more teaching-to-test and emphasize more the consequences of failing the high-stakes tests in class. The negative effects of such teaching behaviors on student intrinsic motivation and hence on student learning may not be observed instantly in grade 3, but may start to appear sometime after the promotional gate grade. I hypothesize that if students were taught by positive-expectation teachers in grade 3, in a long run, they will bounce back from short-term negative policy effects, if there were any. I also hypothesize that if students were taught by negative-expectation teachers, the test-based retention
policy will have negative residual effects on students’ long-term learning, that is, the short-term learning gain may not sustain overtime while the learning loss may last till several years later.

4. Policy-by-teacher effects on student self-perceived competence and interests

Three aspects of student self-perceived competence and interests toward tested subjects will be examined in the current study: student overall self-perception, student differential self-perception by prior ability, and student long-term self-perception. Again, due to the scarcity of literature about the indifferent expectation, I present the hypothesis about student self-perception related to negative and positive teacher expectations only.

**Student overall self-perception.** Adopting the same rationale as that for the policy-by-teacher effects on student overall academic performance, that is, incentives have limited power in changing instructional capacity and teachers’ pedagogy, I hypothesize that the grade-3 test-based retention policy will not improve students’ overall self-perceived competence and interests in tested subjects.

**Student differential self-perception by prior ability.** As previously discussed, each type of teachers will likely change their teaching practices by reallocating curricular emphasis and by intentionally or unintentionally differentiating students in responding to the test-based retention policy. I suspect that such changes may have differential effects on students’ self-perception according to their prior ability levels. However with limited theoretical understanding of student self-perceived competence and interests, I am not able to present any hypothesis regarding the specific differential pattern under each type of teacher expectations.

**Student long-term self-perception.** According to the developmental characteristics of student academic conception (Nicholls, 1978, 1984; Nicholls & Miller, 1984), students will become more motivationally mature and will start to differentiate ability and efforts as they grow
up. As previously discussed, the policy may hurt students’ long-term intrinsic motivation if implemented by a negative-expectation teacher, but may protect students’ intrinsic motivation if implemented by a positive-expectation teacher. I therefore predict that if operated with the negative expectation, the test-based retention policy may have negative residual effects on student self-perceived competence and interests after the promotional gate grade; however, if operated with the positive expectation, the policy may not have any long-term negative effects on student academic self-perception.
CHAPTER 3
METHODOLOGICAL CHALLENGES IN STUDIES OF HIGH-STAKES TESTING

In this chapter, I identify three major methodological challenges for estimating the casual effects of high-stakes testing, including (a) measures of student academic performance, (b) problem of selection bias, and (c) generalizability of research findings. With a goal to seek directions for more rigorous research methods, my discussion is mainly framed through analyzing the methodological strengths and weaknesses of previous large-scale quantitative studies on high-stakes testing and student academic performance.

Measures of Student Academic Performance

One major challenge for studies of high-stakes testing is to find an acceptable measure of student academic performance. In search for such a measure, decisions have to be made regarding the choice of achievement test, score metric, and unit of measures.

Choice of Achievement Test

In general, the previous quantitative studies relied on either local achievement tests (Betts & Danenberg, 2002; Bishop, Mane, Bishop, & Moriarty, 2001; Bryk, 2003; Camilli & Bulkley, 2001; Greene, 2001; Jacob, 2001, 2005; Ladd, 1999; Roderick et al., 2002) or nation-wide assessments to measure student learning (Amrein & Berliner, 2002, 2003; Braun, 2004; Carnoy & Loeb, 2002; Dee & Jacob, 2011; Frederiksen, 1994; Hanushek & Raymond, 2005; Lee & Wong, 2004; Neil & Gayler, 2001; Nichols, Glass, & Berliner, 2006; Raymond & Hanushek, 2003; Rosenshine, 2003; Winfield, 1990).

Local achievement tests. When employing local achievement tests, a major concern is that the test results were often used as a basis of high-stakes decision making. Being both the means and the ends of accountability policies, the high-stakes test scores were prone to the
problem of score inflation, that is, “performances on the tested sample increases substantially more than proficiency in the domain about which inferences are drawn” (Koretz, 2005, p. 104). In high-stakes contexts, test scores gains may be caused by a variety of construct irrelevant factors and are likely to reflect growing test familiarity and inappropriate test preparation rather than genuine difference in learning (Koretz, 2002, 2005; Shepard, 1990). In the case of the Chicago Ending Social Promotion program, researchers (Bryk 2003; Jacob, 2003, 2005) found that the substantial gain in test performance following the announcement of the policy could not be explained by students’ and schools’ prior-policy characteristics and was related to students’ increased test-taking motivation and test-specific skills; the gain was not transferable to the moderate-stakes tests of the Illinois Goals Assessment Program (IGAP); neither did it sustain over time even with the existence of similar policy initiatives. The problem of score inflation was also obvious in other studies of high-stakes testing. For example, in regards to the Dallas school accountability (also known as Texas test-based accountability policy in general), Klein, Hamilton, McCaffrey and Stecher (2000) discovered that the large improvement indicated by TAAS (Texas Assessment of Academic Skills) was not shown in other low-stakes tests such as NAEP (National Assessment of Educational Progress). The huge discrepancy between TAAS and NAEP, which was as well confirmed by other researchers (e.g., Haney, 2000; Peterson & Hess, 2005), was not captured by the skill and format differences between the two tests (Jacob, 2007). In evaluating the school-based accountability in California, Betts and Danenberg (2002) similarly questioned the dramatic increase in SAT9 (Stanford9) math scores and expressed a concern about its real meaning. When it comes to Florida’s school voucher program, Green’s use of FCAT (Florida Comprehensive Assessment Test) scores (2001) was equally problematic because the at-risk schools in her study made gains “even larger than would have been expected
simply given how low their previous scores were” (p. 11).

Another problem accompanying the use of high-stakes tests is distorted test-taking pool. Researchers (e.g., Allington & McGill-Franzen, 1992; Cullen & Reback, 2006; Figlio, 2006; Figlio & Getzler, 2002; Haney, 2000; Heilig & Darling-Hammond, 2008; Wheelock, 2003) have discovered that the high-stakes testing that provides rewards or sanctions on the basis of average student scores may create incentives for schools to manipulate the test-taking population. Schools would retain students that might be expected to have low achievement, place them into special education program, or encourage them to transfer to other schools or drop out before the grade in which the most important tests are given. By weeding out low-achieving students, schools may look better on grade-equivalent test scores, making a false impression of score gains. For example, there was indisputable evidence of score-boosting games along with the use of TAAS to estimate the impact of the accountability policy in Texas (e.g., Cullen & Reback, 2006; Haney, 2000; Heilig & Darling-Hammond, 2008). Concerning the utilization of ITBS to examine the Chicago policy, Jacob (2005) revealed an increase in the special education rates among bottom quartile students in high-achieving schools and an increase in grade retention in both high- and low-achieving schools before the promotional gate grades. The same problem existed with FCAT in Florida: Figlio and Getzler (2002) provided evidence that the rates of special education placement were higher in grades that entered into the accountability system; Haney (2008) found that disproportionately high numbers of minority student were flunked out and were required to repeat grade three. Although it is still debatable whether grade retention and dropouts were keys to improvement in high-stakes tests scores (Carnoy & Loeb, 2002; Haney, 2000; Toenjes & Dworkin, 2002), high-stakes testing clearly motivated schools and teachers to push out low-achieving students from the test-taking pools and thereby threatened the construct
validity of the test results.

**Nationwide assessments.** Many researchers have attempted to walk around the problems of score inflation and distorted test-taking pool by resorting to NAEP and NELS (National Educational Longitudinal Survey), the two nationwide assessments that are generally considered as low-stakes independent tests. In low-stakes contexts, teachers are unlikely to coach students as vigorously as in high-stakes tests; schools have fewer motives to dissuade low-achieving students from the tests taking. Although students are suspected to have reduced motivation in low-stakes tests (Kiplinger & Linn, 1993; O'Neil Jr., Sugrue, & Baker, 1995), the motivation decrement can be assumed relatively consistent overtime and thus will not bias the inference of score gains (Koretz, 2002). However, there are still two limitations in using NAEP and/or NELS to measure student performance. One is that the inference of student learning based on these tests is only limited in the knowledge and skills covered in the tests. If test items do not fully captured the latent achievement traits required by accountability standards, the test scores will not be a good indicator of student academic performance (Jacob, 2003). To understand the scores meaning, there is a need to examine student performance on different types of test items or to differentiate the performance by different skills – neither approach seems feasible with the composite measures provided by NAEP and NELS. The second limitation is specific to NAEP and is relevant to its important role in the public accountability debate. Because NAEP is relatively reliable, valid, and readily available tool to compare results across states (Lee, 2008b; Linn, 2000), more and more studies started using NAEP to estimate the effectiveness of state accountability, which consequently makes the tests no longer low-stakes as it used to be. Therefore, it is still a challenge for future research to find an appropriate test for studying the effects of high-stakes testing, that is, a test with true low stakes and a test that enables
investigation of student performance by items or skills.

**Choice of Score Metric**

Another concern related to the measures of student academic performance is the choice of score metric. The previous research presents two options of score metric in measuring student academic performance, scale scores (e.g. Amrein & Berliner, 2002, 2003; Bishop et al., 2001; Braun, 2004; Bryk, 2003; Camilli & Bulkley; Greene, 2001; Dee & Jacob, 2011; Hanushek & Raymond, 2005; Jacob, 2001, 2005; Lee & Wong, 2004; Neil & Gayler, 2001; Nichols, Glass, & Berliner, 2006; Raymond & Hanushek, 2003; Roderick et al., 2002; Rosenshine, 2003; Winfield, 1990) and/or proficiency rate (i.e. proportion of students scoring at/above certain proficiency level) (Betts & Danenberg, 2002; Carnoy & Loeb, 2002; CEP, 2007; Ladd, 1999). Comparison of studies by score metric did not reveal notable differences in causal effects estimates. However, Neil and Gayler’s (2001) study suggested that the estimates may be sensitive to the choice of score types; they did not observe any policy effects with state-level scale scores, but found positive effects on the percentages of students that moved from lower levels to ‘basic or above’ level in grade 4. In my opinion, the two score metrics are measuring different aspects of the policy effects. The scale scores, depending on whether they were derived through scaling procedure of the Item Response Theory (IRT), are appropriate either for characterizing learning trend overtime or for comparing magnitude of academic performance between different treatment conditions. Differently, the proficiency rate is a categorical measure of student academic performance and works the best for examining differential learning by proficiency levels or by student prior ability. In the studies of high-stakes testing, different metrics should be used together so that the results can complement each other and provide more comprehensive explanations of the policy effects.
**Unit of Measures**

Besides the choice of achievement tests and score metric, unit of measures should as well be chosen with deliberation. In the past literature, very limited empirical studies examined student academic performance at individual level (e.g. Jacob, 2001, 2005; Roderick et al, 2002). Instead, most of the attention has been directed to state- or school-level measures, which is probably due to the constraint of the available dataset. For example, NAEP test results are only reported at the state level and do not contain information about individual students. The aggregated test outcomes, however, may mask the underlying heterogeneous effects on individual students and prevent us from pinpointing the source of any high or low performance (Hanushek & Raymond, 2003). In addition, Kane and Staiger (2002) noted that the variance of average measurement error on a test is inversely related to the number of students tested. When scores are aggregated across a smaller number of students, the variance of measurement error increases and can directly affect our effects estimate. These potential problems suggest the importance in using individual students as units of measures in the studies of high-stakes testing.

**Problem of Selection Bias**

Estimates of treatment effects are considered biased when the treated and control groups differ systematically in terms of their pretreatment characteristics (Rosenbaum, 2002). The problem of selection bias is another challenge in the studies of high-stakes testing due to the inapplicability of randomized experiments. Below I review existing research designs and statistical approaches and discuss their potential and limitations in removing the selection bias.

**Inapplicability of Randomized Experiments and Problems in Post Facto Designs**

Under Rubin’s causal model (Holland, 1986; Rubin, 1978), the causal effect of one treatment versus another is the difference between the respective potential outcomes of a student
given the treatment settings. In an ideal world, under randomization, for each treatment group, one can have an unbiased estimate of the corresponding average population potential outcome presented by the sample. However, it is impractical and problematic to conduct large-scale experiment and to randomly assign intact schools to different testing and accountability situations. Moreover, in the current education system, high-stakes testing is so pervasive that almost every school/state had certain experience of one or even more types of high-stakes testing policies, which makes it hard to construct a control group for examining the policy impact. Therefore it is not surprising to find that almost all the past studies have relied on non-experimental data and have mostly used a post facto design.

Past studies on high-stakes testing often employed between-subject comparison or pre- and post-policy or longitudinal comparison for estimating the policy effects. With between-subject comparison (Bishop et al., 2001; Braun, 2004; Camilli & Bulkley, 2001; Carnoy & Loeb, 2002; Frederiksen, 1994; Greene, 2001; Hanushek & Raymond, 2005; Ladd, 1999; Lee & Wong, 2004; Neil & Gayler, 2001; Raymond & Hanushek, 2003; Winfield, 1990), researchers utilized cross-sectional data to compare groups with or without high-stakes testing; and many of them did not control for students’ prior academic status; as a result, the treatment groups in these studies were hardly made comparable. With pre- and post-policy or longitudinal comparison (Amrein & Berliner, 2002; Betts & Danenberg, 2002; Bryk, 2003; CEP, 2007; Jacob, 2001, 2005; Lee, 2008a; Roderick et al., 2002), researchers mostly focused on the score trend of a same cohort. A major problem of this design is potential interference from other accountability initiatives. In the case of the Chicago studies, test-based retention mingled with other accountability policies, such as remedial education program, extended learning time, and school probation (Jacob & Lefgren, 2004; Smith, Roderick, & Degener, 2005); hence the observed difference may not be attributed
to the test-based retention policy only. When comparing pre-post policy trend, it is possible to add in one comparison group. As a part of his exploratory analysis, Jacob (2005) used a panel of district-level data to compare achievement trends in Chicago to those in other mid-western cities where the similar accountability policy initiatives were not mandated during the study period. However, the extra comparison group might not serve as a clean control condition since the current educational system has been pervaded with all kinds of school-based accountability initiatives. Even with an ideal control condition, selection bias may still be a problem due to unobserved confounding variables that were correlated with high-stakes testing policy and that as well influenced student outcomes. For example, if struggling schools were more likely to adopt the policy, without fully capturing the school characteristics, the estimates of policy effects would be biased. From the perspective of causal inference, the post facto designs are in one way or another troublesome. To ensure the causal validity of findings from the analysis of non-experimental data, it is important to supplement the designs with advanced statistical methods.

**Existing Statistical Solutions to Problem of Selection Bias**

A common statistical solution to the problem of selection bias is to directly control for differences in schools and student characteristics (e.g. Bishop et al., 2001; Carnoy & Loeb, 2002; Hanushek & Raymond, 2005, Ladd, 1999; Lee, 2008a; Lee & Wong, 2004; Nichols et al., 2006; Winfield, 1990). However, statistical adjustment for a limited number of pretreatment covariates could hardly be relied upon to remove the bias. Moreover, model-based linear extrapolation hinges on the assumption that each covariate-outcome association is the same for the units in different treatment conditions. This assumption may not be warranted when the distributions of the two treatment conditions do not entirely overlap. Some of the schools/states in one policy condition may never have a chance to adopt another type of policy given their pretreatment
characteristics. The assumption is unlikely to hold either for the pre-post policy or longitudinal single-cohort studies, because the covariate-outcome association may not stay the same once other accountability initiatives start to exert their efforts during the study period; as a result, the extrapolation from one time point to another may not be justified.

There also existed several other sophisticated statistical approaches that intended to improve the quality of the effects estimation in the studies of high-stakes testing. For example, Jacob (2001) employed errors in variables analysis in his study of high-school graduation exams. While aiming to remove the selection bias related to prior achievement status, his approach took into account potential underestimation of the policy effects due to the ambiguous policy reference point. The approach was useful for removing a certain amount of bias and for avoiding over-control of possible post-treatment outcomes. But it still relied on direct statistical adjustments, and thus similarly suffered from the problems of unwarranted model-based extrapolation.

In view of the difficulty in constructing a natural control group for longitudinal studies, researchers of the Chicago Ending Social Promotion program (Bryk, 2003; Roderick et al., 2002) resorted to value-added models to predict ‘gate-year’ scores given the pretreatment status and relevant characteristics. By comparing the observed and predicted score gain (i.e. value added), the researchers were seeking a causal explanation of the policy effects. However, as Rubin, Stuart, & Zanutto (2004) pointed out, value-added models “should not be seen as estimating causal effects of teachers or schools, but rather as proving descriptive measures” (p. 113). Without carefully considering the assumptions underlying the intended causal inference and without appropriate control for sufficient number of covariates, the results from the value-added models are as well prone to the problem of selection bias (also see Reardon & Raudenbush,
When analyzing the effects of the NCLB Act on state-level academic achievement, to remove bias related to other nationwide changes, Dee and Jacob (2011) adopted a comparative interrupted time series (CITS) design and compared states with and without school-accountability systems prior to the NCLB in terms of their achievement gains associated with the NCLB provision. However, the use of CITS design requires a nonequivalent control group and is often constrained by the data available.

Regression discontinuity analysis is another potential approach to improve the accuracy of the effects estimates. This approach, albeit efficient in removing the bias, can only focus on the sample around the cutoff threshold, which explains why it was mostly seen in studies that examined the policy effects on at-risk students only (Greene & Winters, 2007; Jacob & Lefgren, 2004; Roderick & Nagaoka, 2005). To sum up, the previous research designs and statistical methods had limited power in dealing with the problem of selection bias and thus present a need for an alternative statistical approach for estimating the causal effects of high-stakes testing.

**Generalizability of Research Findings**

We learnt from the past research that the effects of high-stakes testing varied among different racial subgroups (e.g. Dee & Jacob, 2011; Hanushek & Raymond, 2005; Ladd, 1999; Nichols et al., 2006) and were not the same for students at various stages of developmental courses (e.g. Carnoy & Loeb, 2002; Dee & Jacob, 2011; Jacob, 2005; Roderick et al., 2002); neither were the effects consistent for students from different classroom contexts and/or school environment (see discussion in Chapter 2). In addition to the differences in study population, the generalizability of previous research findings may also be limited by the conceptual ambiguity of high-stakes testing.
A brief review of the past studies reveals that except for the studies that focused on high-school graduation exams (Bishop et al., 2001; Jacob, 2001; Neil & Gayler, 2001), most of the previous research ignored the heterogeneity of test-based initiatives. This is especially true with the cross-states studies that examined high-stakes testing as a general state-level phenomenon (see review of Lee 2008b). Although various incentives (e.g. financial awards to schools/students, school probation, teachers/students replacement, etc.) were applied to different organization levels, only three cross-state studies (Carnoy & Loeb, 2002; Lee & Wong, 2004; Nichols et al., 2006) experimented with continuous measures of accountability policies and considered the variation in the strength of policy implementation and/or in the dynamic nature of accountability system. The conceptual ambiguity problem was also present in the studies of local high-stakes testing system. In most of these studies, multiple types of stakes often coexisted in the same local program under investigation. For example, in addition to awards/sanctions to schools based on test performance, Florida enforced the policy with school vouchers (Camilli & Bulkley, 1999; Greene, 2001); California offered monetary incentives to high-performing students (Betts & Danenberg, 2002); Dallas provided bonuses to the staff of winning schools (Ladd, 1999); New York and North Carolina supplemented the school accountability with end-of-course graduation exams (Bishop et al., 2001); and Chicago tied grade promotion to students’ test scores (Bryk, 2003; Jacob, 2005; Roderick et. al., 2002). To make it more complicated, in the same program, multiple types of incentives sometimes were applied to districts at different times and/or with different strength. For instance, different test-based reform efforts were enacted from 1996 to 2001 in Chicago (Roderick et al., 2002) and from 1997 to 2000 in California (Betts & Danenberg, 2001). All the complexity and variability in conceptualizing high-stakes testing partially explain the current controversies regarding the policy effects and
present a challenge for generalizing the findings to a larger population. In order to maximize
generalizability, a potential solution is to be explicit about the construct and context to be
generalized and/or to examine the policy effects by isolating specific incentive types. Since
previous research has only focused on high-school graduation exam, more studies are needed to
address the effects of other types of incentives.
CHAPTER 4
RESEARCH METHODS

Data

Data came from the Early Childhood Longitudinal Study Kindergarten Class (ECLS-K) of 1998-99 released by the U.S. National Center for Education Statistics (NCES). The dataset contains repeated observation of a nationally representative sample of students, their teachers, and schools in a total of seven waves – two waves in kindergarten (fall 1998, spring 1999), two waves in grade 1 (fall 1999 and spring 2000), and one wave for each of grades 3, 5, and 8 (spring 2002, 2004, and 2007 respectively) (NCES, 2005). I employed the data from kindergarten to grade 5 but excluded the third wave (i.e. fall 1999) when only a random sub-sample of students participated in the ECLS assessment. I identified 9488 students who were enrolled in grade-3 classes \( n = 3324 \) in spring 2002 and whose schools \( n = 1498 \) reported having standardized testing in that year.

Measures

This study involved two treatment measures (i.e. grade-3 test-based retention policy and teacher expectations) and three types of outcomes (i.e. student academic performance, instructional time, and student self-perceived competence and interests). It as well took into account a list of pretreatment variables and utilized student prior academic ability as a moderator.

Grade-3 Test-based Retention Policy

The grade-3 test-based retention policy was measured based on responses to two items in the spring 2002 school administrator questionnaire: one item asked whether grade 3 students were tested with standardized tests; the other item inquired whether students could be retained
due to failing a school-wide standardized test. Among the schools selected, 278 (18.56%) schools established grade-3 test-based retention policy while 1220 (81.44%) adopted school-wide standardized testing but imposed no retention policy tied to test results. I treated ‘standardized testing only’ as a control condition and excluded schools without standardized testing ($n = 111$) from the analysis. I made this decision because of my focal interest in the effects of high-stakes incentives and because schools without standardized testing were found quite unique and largely incomparable to the other two policy groups\(^1\). In this study, the grade-3 test-based retention policy was viewed as a school-level treatment\(^2\). Because at present only a few school districts and states formally implemented the policy (Penfield, 2010), I consider the decision of adopting the test-based retention policy might be mostly initiated within individual schools.

**Teacher Expectations**

In the ECLS spring 2002 questionnaires, teachers were asked to rate their attitude toward the statement “Many of the children I teach are not capable of learning the material I am supposed to teach them” on a five-point Likert scale (1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree). I created the measure of teacher expectation\(^3\) by categorizing response options 1 and 2 as positive expectation ($n = 2510$, 75.51%), 3 as indifferent expectation ($n = 412$, 12.39%), and 4 and 5 as negative expectation ($n = 402$, 12.09%).

To understand how teacher expectations are associated with teacher, classroom, and school characteristics, I conducted two sets of exploratory analysis: one compared the teachers across the whole ECLS spring 2002 sample at a single level while the other explored how the teachers would differ if they taught in a same school using a two-level model (with teachers at
level 1 and schools at level 2). The latter assumed fixed school membership by centering the dummy indicators of teacher expectations at their respective school mean as predictors of teacher and classroom characteristics; hence it was only based on a sub-sample of schools \((n=452)\) in which more than one type of teacher expectation was available. The analysis suggests that a teacher’s expectation of students’ learning ability reflected his/her school settings, classroom composition, as well as the teacher’s will and skills to teach.

In general, positive expectation was prevalent in affluent schools and was associated with more socially and academically privileged students and more supportive and coherent school community. When the school settings are given, positive-expectation teachers encountered fewer Hispanic students than negative-expectation teachers, and served more Asian students and fewer misbehaved students than indifferent-expectation teachers. In a same school, positive-expectation teachers took more pride in and showed more enjoyment of the teaching career than did indifferent-expectation teachers; compared to negative-expectation teachers, positive-expectation teachers were more satisfied with the school environment and allocated more time to test preparation without compromising the teaching quality in math and reading.

Negative teacher expectation was often associated with less favorable school settings (e.g. higher minority concentration, more disciplinary problems, poorer leadership, higher teacher turnover rate, and/or less sufficient educational resources). Controlling for differences in school characteristics, negative-expectation teachers seemed more likely to be a white teacher and/or a primary teacher of language and literacy than positive-expectation teachers; these two expectation groups did not differ from each other in their self-efficacy and commitment to teaching, but negative-expectation teachers complained more about school environment and administrative duty and worried more about insufficient supply of instructional resources. In a
same school, negative-expectation teachers focused more on basic skills in reading instruction than their positive-expectation colleagues, though students of these two types of teachers entered into the grade 3 with a similar level of prior reading ability.

Indifferent-expectation teachers were distinguished by their schools’ tighter control over curriculum and teaching and the greater demand for improving student achievement as well as by their own lack of professional status and teaching morale. Indifferent-expectation teachers had similar chances as negative-expectation teachers to teach in a school with high proportion of disadvantaged students but appeared to have more educational resources. Compared to positive- or negative-expectation teachers, indifferent-expectation teachers seemed to view themselves as more important in influencing school policy, including determining discipline policy, deciding how some school funds will be spent, and assigning children to classes; but they were under more pressure to boost up students’ achievement and tended to depreciate professional development opportunities and to practice deskilling instructional strategies within classrooms. Compared to those of the other two types of teachers, the school administrators of indifferent-expectation teachers, were more capable to maintain school discipline and order; they placed greater emphasis on improving teachers’ professional capacity, but lacked enough attention to professional communication within school and to instructional strategies aligned with high standards; the administrators were more likely to believe that there was shared expectation among the teachers and staffs and/or that the school community was supportive of school goals and activities. If assigned to a same school with the other two types of teachers, indifferent-expectation teachers were less likely to be a regular full-time teacher or primary teacher of language and literacy despite the fact that they reported similar educational background and similar teaching certification and even had more teaching experience with the current school
and/or the current grade; the indifferent-expectation teachers also communicated less frequently with other school teachers, lacked a sense of belonging to the school community, and showed less respect to school administrators. Different from their positive-expectation colleagues, the indifferent-expectation teachers were less likely to enjoy their schools and teaching.

**Student Academic Performance**

Student academic performance was measured by ECLS direct assessments of reading, math, and/or science. The ECLS test specifications of each subject were derived mainly from the corresponding NAEP framework and accommodated the knowledge and skills that were typically emphasized in each of the sampled grades (NCES, 2002). ECLS reported assessment results at the individual student level. Due to the unique study design (i.e. only a limited number of students were sampled from each school; individual students, instead of classes or schools, were followed throughout the study), the assessment results were unlikely to bring high stakes to students, teachers, or schools and thus were less subject to the contamination of inappropriate test preparation practices and school gaming strategies such as excluding low-achieving students from test-taking pools (Cullen & Reback, 2006; Figlio & Getzler, 2002; Haney, 2000). My analysis employed three outcome metrics that were provided by the ECLS to evaluate different aspects of policy effects on student academic performance, namely, standardized T scores, Item Response Theory (IRT) \( \theta \) scores/Oral Language Development Scale (OLDS) scores, and students’ highest proficiency level.

**Standardized T scores.** The standardized T scores in reading, math, and science reflect a student’s relative standing within each wave and were standardized with a mean of 50 and a standard deviation of 10 for the whole ECLS sample. I used the T scores to compare students’ performance across reading \((M = 51.79, SD = 9.01)\), math \((M = 51.82, SD = 9.16)\), and science
(M = 51.49, SD = 9.51) in grade 3.

**IRT \( \theta \)/OLDS scores.** To characterize students’ overall academic growth, I relied on the IRT \( \theta \) scores and/or OLDS scores\(^4\). In the first four waves of the ECLS study, language minority students were screened with an Oral Language Development Scale first; On a scale ranging from 0 to 60 for the OLDS tests, only those who reached 37 or beyond were allowed to take the direct assessments and had IRT \( \theta \) scores available. Therefore, my analysis incorporated both types of scores as students’ repeated academic outcomes. Table 4.1 lists the respective distributions, sample sizes and reliability estimates for the IRT \( \theta \) scores and OLDS scores used in the current analysis. According to the ECLS psychometric report (NCES, 2005), the IRT \( \theta \) scores were estimated using a three-parameter IRT model. It was not clear whether the OLDS scores were also vertically scaled. Judging that the OLDS scores measured students’ language ability instead of academic achievement, I consider vertical scaling not necessary in this case.

**Proficiency levels.** To evaluate a student’s mastery of specific skills in reading and math, I examined the highest proficiency level each student achieved. ECLS constructed this measure based on a set of dichotomous pass/fail scores and defined the proficiency levels for each subject on a hierarchical scale: if a student demonstrated mastery of certain skill and knowledge at one level, he/she should have passed the items from the lower levels as well (NCES, 2005). Judging from the student distribution and the curricular emphasis in grade 3, I combined levels 1-4 (i.e. letter recognition, beginning sounds, ending sounds, and sight words) of reading and levels 1-3 (i.e. number and shape, relative size, and ordinality, sequence) of math as the lowest reading/math proficiency level for the third graders.

By the end of grade 2, most of the students should have mastered basic reading skills including familiarity with print and recognition of letters and phonemes (i.e. levels 1-4, the
lowest reading proficiency level); the focal emphasis of a typical grade-3 reading curriculum is shifted to comprehension skills including inferential understanding of text and developing interpretation (e.g. levels 6-7) (NCES, 2005). Since vocabulary knowledge is important to the development of reading comprehension skills (NICHD, 2000), the skill of reading words in context (i.e. level 5), which most of elementary students have started to learn in their early grades, continues to be a focus in some grade-3 classrooms. On the other hand, the evaluation skill (i.e. level 8) involves a relatively higher level understanding of the written text and requires students to connect the information in the text with their own personal background knowledge; whether to emphasize this type of skill in the grade-3 reading instruction might be subject to the decisions of individual teachers and/or of their schools.

As for the math skills, ECLS defined levels 1-3 (the lowest math proficiency level) as the dominant emphasis only of the K-1 curriculum (NCES, 2005). According to the standards of the National Council of Teachers of Mathematics (NCTM, 2009), most students should be able to solve simple addition and subtraction problems (i.e. level 4) by the end of grade 2; the focal emphasis of grade-3 math curriculum is on gaining fluency in using all basic multiplication and division facts (i.e. level 5); efficient use and in-depth understanding of the operations also require conceptual knowledge of place value (level 6). Compared to the other levels, using knowledge of measurement and rate to solve math problems (i.e. level 7) requires more advanced problem-solving skills and might not be a curricular emphasis in some grade-3 classrooms.
Table 4.1

Summary of K-5 $\theta$ and OLDS Scores

<table>
<thead>
<tr>
<th>Round of Assessment</th>
<th>Reading $\theta$ Scores</th>
<th>OLDS Scores</th>
<th>Math $\theta$ Scores</th>
<th>Science $\theta$ Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n$</td>
<td>Mean (SD)</td>
<td>Reliability</td>
<td>$n$</td>
</tr>
<tr>
<td>Fall 1998 (Kindergarten)</td>
<td>7946</td>
<td>-1.23 (0.49)</td>
<td>0.92</td>
<td>2170</td>
</tr>
<tr>
<td>Spring 1999 (Kindergarten)</td>
<td>8881</td>
<td>-0.65 (0.46)</td>
<td>0.95</td>
<td>1553</td>
</tr>
<tr>
<td>Spring 2000 (Grade 1)</td>
<td>9221</td>
<td>0.19 (0.39)</td>
<td>0.96</td>
<td>540</td>
</tr>
<tr>
<td>Spring 2002 (Grade 3)</td>
<td>9488</td>
<td>0.83 (0.28)</td>
<td>0.94</td>
<td>---</td>
</tr>
<tr>
<td>Spring 2004 (Grade 5)</td>
<td>9488</td>
<td>1.07 (0.28)</td>
<td>0.93</td>
<td>---</td>
</tr>
</tbody>
</table>

Note. According to the ECLS grade 3 psychometric report (NCES, 2005), Cronbach Alpha reliability were provided for $\theta$ scores while OLDS scores were evaluated with split half reliability; There are more students with math theta scores than with reading $\theta$ scores in the first three waves. This is probably because some language minority students took math and general knowledge assessments in Spanish; Science was not tested before grade 3.
## Table 4.2

### Student Distribution over the Proficiency Levels

<table>
<thead>
<tr>
<th>Proficiency Levels</th>
<th>Skill Content</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reading</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1) Letter recognition</td>
<td>identifying upper- and lower-case letters by name</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>2) Beginning sounds</td>
<td>associating letters with sounds at the beginning of words</td>
<td>6 (0.1)</td>
</tr>
<tr>
<td>3) Ending sounds</td>
<td>associating letters with sounds at the end of words</td>
<td>58 (0.6)</td>
</tr>
<tr>
<td>4) Sight words</td>
<td>recognizing common words by sight</td>
<td>264 (2.8)</td>
</tr>
<tr>
<td>5) Comprehension of words in context</td>
<td>reading words in context</td>
<td>1700 (17.9)</td>
</tr>
<tr>
<td>6) Literal inference</td>
<td>making inferences using cues that were directly stated with key words in text</td>
<td>2416 (25.5)</td>
</tr>
<tr>
<td>7) Extrapolation</td>
<td>identifying clues used to make inferences</td>
<td>2768 (29.2)</td>
</tr>
<tr>
<td>8) Evaluation</td>
<td>demonstrating understanding of author’s craft and making connections between a</td>
<td>2276 (24.0)</td>
</tr>
<tr>
<td></td>
<td>problem in the narrative and similar life problems</td>
<td></td>
</tr>
<tr>
<td><strong>Math</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1) Number and shape</td>
<td>identifying some one-digit numerals, recognizing geometric shapes, and one-to-one-counting of up to 10 subjects</td>
<td>2 (0.0)</td>
</tr>
<tr>
<td>2) Relative size</td>
<td>reading all single-digit numerals, counting beyond ten, recognizing a sequence of patterns, and using nonstandard units of length to compare subjects</td>
<td>6 (0.1)</td>
</tr>
<tr>
<td>3) Ordinality, sequence</td>
<td>reading two-digit numerals, recognizing the next number in a sequence, identifying the ordinal position of an object, and solving a simple word problem</td>
<td>299 (3.2)</td>
</tr>
<tr>
<td>4) Addition/subtraction</td>
<td>solving simple addition and subtraction problems</td>
<td>1555 (16.4)</td>
</tr>
<tr>
<td>5) Multiplication/division</td>
<td>solving simple multiplication and division problems and recognizing more complex number patterns</td>
<td>2953 (31.1)</td>
</tr>
<tr>
<td>6) Place value</td>
<td>demonstrating understanding of place value in integers to hundreds place</td>
<td>3047 (32.1)</td>
</tr>
<tr>
<td>7) Rate and measurement</td>
<td>using knowledge of measurement and rate to solve word problems</td>
<td>1626 (17.1)</td>
</tr>
</tbody>
</table>

*Note.* The description of the skill content is from the ECLS grade 3 psychometric report (NCES, 2005).
**Instructional Time**

Instructional resource allocation was characterized by how grade-3 teachers allocated instructional time to reading ($M = 385.49$, $SD = 134.80$), math ($M = 268.63$, $SD = 100.82$), and science ($M = 103.25$, $SD = 73.14$) over a week. The measures of instructional time were constructed based on teachers’ responses to two sets of items in the 2002 spring teacher questionnaire: one item required teachers to rate on a 5-point scale how frequently they taught each subject per week, and the other was about the duration of instruction in each subject during a typical instructional day using a four-point scale. I assigned a middle value to each category in the frequency measure (never = 0, less than once a week = 0.5, one to two times a week = 1.5, three-four times a week = 3.5, daily = 5) and in the duration measure (1 to 30 mins a day = 15, 30-60 mins a day = 45, 61 to 90 mins a day = 75, more than 90 mins a day = 105), and obtained continuous measures of instructional time by multiplying the recoded frequency and duration for instruction in each subject.

**Student Self-perceived Competence and Interests**

The measure of student self-perceived competence and interests \(^5\) is a composite score of eight items assessing students’ self-perception of their grades of each tested subject (i.e. reading and math), the difficulty of reading/math work, as well as their interests in and enjoyment of reading/math. The reading self-reported scores had a mean of 3.26 with a standard deviation of 0.65 and a reliability of 0.87 in spring 2002, and a mean of 2.85 with a standard deviation of 0.79 and a reliability of 0.90 in spring 2004. The math scores had a mean of 3.14 with a standard deviation of 0.79 and a reliability of 0.90 in spring 2002, and a mean of 2.91 with a standard deviation of 0.78 and a reliability of 0.92 in spring 2004. These measures reflected more of academic self-perception other than students’ true academic ability as they were weakly related
to teachers’ rating of students’ academic skills in reference to the grade average, and to students’
direct assessment scores, both with a correlation of around .20.

**Student Prior Academic Ability**

The measure of student prior academic ability was used as a moderator to examine the
hypothesis that the policy produced differential academic and motivational patterns among
students at different ability levels. One desirable characteristic of a moderator variable is to be
uncorrelated with the treatment (Baron & Kenny, 1986). This study defined student prior ability
as the end-of-grade 2 reading ability that was unaffected by the grade 3 test-based retention
policy. Because the ECLS study did not assess students between grade 1 and 3, using a weighted
value-added model-based approach (see Appendix A for the nonlinear growth model
specifications), I first examined whether students’ K-1 reading growth was subject to the
influence of the grade 3 test-based retention policy. As the policy effects turned out to be
negligible (see Appendix A for the results), with a same model I then estimated each student’s
pre-policy reading performance by extrapolating the K-1 growth trajectory to the end of grade 2.
Based on the distribution of the predicted end-of-grade 2 reading scores, I classified the students
into five equal-sized ability groups. Dividing sampled students into four to five ability groups
has been one common approach in past studies on students’ differential performance under high-
stakes testing (e.g., Reback, 2008; Roderick et al., 2002).

**Other Pretreatment Variables**

Pretreatment variables have to occur before the treatment and/or be immune to the
influence of the treatments. The ECLS-K study did not provide information on the year when the
grade-3 test-based retention policy was first introduced into each treated school. It is likely that
some of the sampled schools might have started implementing the policy earlier than fall 1998,
the beginning of the ECLS study. The policy might have affected students and teachers in a
treated school before they entered into the promotional gate grade, and might have altered the
whole school climate, including instructional practices, school policy, teacher assignment within
the school, and teacher and principal turnover. Therefore this study only used two types of
pretreatment variables that are unlikely the results of the policy (see Appendix B for the list of
the pretreatment variables). One is demographic measures of students, classes and schools\(^7\). I
view this type of measures relatively stable overtime and thus relied on the information from
spring 2002 to obtain a more accurate and more recent snapshot of the sample characteristics.
The other is students’ prior academic and social emotional status as well as their previous
learning experience, which were measured at the time of kindergarten entry (i.e. fall 1998). The
two types of pretreatment variables were considered either for constructing propensity score
models or as covariates to control in the final outcome models. It has to be pointed out that
measures of teacher characteristics were not selected as pretreatment variables in this study
because they might be closely related to teacher expectations and could be viewed as part of the
class-level treatment.

I used a multi-stage imputation procedure to impute the missing values of the selected
pretreatment variables and to take into account their interdependence in a multilevel setting (see
Appendix C for the imputation strategy).

Causal Inference Strategies

Ideal Research Designs

Corresponding to the four sets of research questions, I conducted four phases of data
analyses. Phase I evaluated the relationship between the grade-3 test-based retention policy and
teachers’ perception of their students’ learning ability. Phase II examined the joints effects of the
policy and teacher expectations on the amount of instructional time that the teachers allocated to
tested and nontested subjects. Phase III focused on the joint effects of the policy and teacher
expectations on grade-3 students’ academic performance in different subjects and on their
mastery of different cognitive skills in tested subjects. In general, reading and math represented
tested subjects while science was viewed as a nontested subject. Additionally, I explored how the
short-term effects varied for students at different ability levels as well as whether the effects
would be sustained till two years later. Phase VI looked at students’ self-perceived academic
competence and interests on average and by academic prior ability; I also compared the self-
perception pattern in the promotional gate year vs. two years later. The four phases of analysis
approached the issue of the test-based retention policy from different angles, seeking an enriched
and elaborated understanding of the causal relationship between the policy, teacher expectations,
and student academic outcomes.

In order to generate unbiased causal inference from the sample statistics, my analysis of
the observational data aimed to approximate three experimental designs. Phase I represented a
randomized design in which schools were randomly assigned to the ‘test-based retention’ group
or the ‘standardized testing only’ group. In the rest three phases, the analysis of the overall joint
effects of the school policy and teacher expectations for students/classes on average was
analogous to a sequential randomized design\(^8\) where schools were first assigned to one of the
two policy conditions, and then classes within each policy group were randomly assigned to
different levels of teacher expectations; in the analysis of the moderated treatment effects by
student prior ability, I constructed a treatment-by-aptitude design in which treatments were
sequentially assigned and were then crossed with five levels of student prior ability.
**Potential Outcomes and Casual Estimands**

To mimic the intended experimental designs using observational data requires an understanding of the potential outcomes and causal estimands in the current study and of the corresponding assumptions and strategies for causal inference. For ease of clarification, here I focus most of the discussion on evaluations of the joint effects of test-based retention policy and teacher expectations on student-level outcomes. The same discussion can be easily tailored to causal analysis of the effects of the school policy only or to on class-level outcomes.

In this study each school was characterized by a dummy indicator of the school policy, with $Z = 1$ if the school imposed test-based retention policy in grade 3 and $Z = 0$ if it adopted school-wide standardized testing but did not retain grade-3 students based on test results. Meanwhile, each class has been assigned to one of the three types of teacher expectations denoted by $T$: $T = 1$ for negative expectation, 2 for indifferent expectation, and 3 for positive expectation. Provided that a class has a possibility of receiving each of the three teacher expectations regardless of school policy and that its school has a possibility of adopting either of the policies, every student in that class will have six corresponding potential learning/social emotional outcomes at the end of grade 3 (see the upper panel of Table 4.3). Hence if we take the expectation of each outcome across all the individual students in the population, there will be six population average potential outcomes: $E[Y(Z=1,T=1)]$ is the average potential outcomes of all students if all the classes were exposed to the negative teacher expectation and all the school adopted the grade-3 test-based retention policy; $E[Y(Z=0,T=1)]$ is the average outcomes of all students if all the classes received the negative teacher expectation and were from schools with standardized testing only. Likewise I define the population average potential outcomes $E[Y(Z=1,T=2)]$, $E[Y(Z=0,T=2)]$, $E[Y(Z=1,T=3)]$, and $E[Y(Z=0,T=3)]$. 
Table 4.3
Potential Outcomes and Causal Estimands

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Label</th>
<th>Potential Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z=0, T=1</td>
<td>negative expectation under standardized testing only</td>
<td>Y(Z=0, T=1)</td>
</tr>
<tr>
<td>Z=0, T=2</td>
<td>indifferent expectation under standardized testing only</td>
<td>Y(Z=0, T=2)</td>
</tr>
<tr>
<td>Z=0, T=3</td>
<td>positive expectation under standardized testing only</td>
<td>Y(Z=0, T=3)</td>
</tr>
<tr>
<td>Z=1, T=1</td>
<td>negative expectation under test-based retention policy</td>
<td>Y(Z=1, T=1)</td>
</tr>
<tr>
<td>Z=1, T=2</td>
<td>indifferent expectation under test-based retention policy</td>
<td>Y(Z=1, T=2)</td>
</tr>
<tr>
<td>Z=1, T=3</td>
<td>positive expectation under test-based retention policy</td>
<td>Y(Z=1, T=3)</td>
</tr>
</tbody>
</table>

Causal Estimands

- effects of test-based retention policy: \( E[Y(Z=1, T=1)-Y(Z=0, T=1)] \); \( E[Y(Z=1, T=2)-Y(Z=0, T=2)] \); \( E[Y(Z=1, T=3)-Y(Z=0, T=3)] \);
- effects of teacher expectations:
  - \( E[Y(Z=1, T=3)-Y(Z=1, T=1)] \);
  - \( E[Y(Z=1, T=2)-Y(Z=1, T=1)] \);
  - \( E[Y(Z=0, T=3)-Y(Z=0, T=1)] \);
  - \( E[Y(Z=0, T=2)-Y(Z=0, T=1)] \);
- interaction effects:
  - \( E[Y(Z=1, T=3)-Y(Z=0, T=3)] - E[Y(Z=1, T=1)-Y(Z=0, T=1)] \);
  - \( E[Y(Z=1, T=2)-Y(Z=0, T=3)] - E[Y(Z=1, T=2)-Y(Z=0, T=2)] \);
  - \( E[Y(Z=1, T=2)-Y(Z=0, T=2)] - E[Y(Z=1, T=1)-Y(Z=0, T=1)] \);

Note: Here I only present the interaction effects of interest in this study, i.e. how the policy effects depend on teacher expectations.

Rubin (1974, 1978, & 1980) defined population average causal effect (i.e. causal estimand) of one treatment versus another as the difference between the two potential average outcomes corresponding to the two treatments. Under Rubin’s causal model, the population average effects of the grade-3 test-based retention policy are viewed as the difference in expected learning/motivational outcomes if all students were exposed to the test-based retention policy versus if all of them received the standardized testing only under each teacher expectation. For example, the causal estimands for the policy effects are \( E[Y(Z=1, T=1)-Y(Z=0, T=1)] \) under
negative expectation, which is equivalent to \( E[Y(Z=1,T=1)] - E[Y(Z=0,T=1)] \). Similarly, population average effects of teacher expectations are estimated through contrasting every pair of corresponding average potential outcomes under each school policy; for instance, the causal estimands for the effects of positive expectation versus negative expectation under the test-based retention policy, \( E[Y(Z=1,T=3) - Y(Z=1,T=1)] \), are the differences in the expected outcomes if all classes were taught by teachers with the positive expectation versus if all classes received the negative expectation. In addition, to evaluate how the population average effects of test-based retention policy depend on teacher expectations, one can compare the causal estimands of the policy effects across teacher expectations, for example, \( E[Y(Z=1,T=3) - Y(Z=0,T=3)] - E[Y(Z=1,T=1) - Y(Z=0,T=1)] \) is a comparison between the positive and negative expectations. I have listed all the causal estimands of interest on the lower panel of Table 4.3.

Nevertheless, the causal analyses are hindered by a “fundamental problem of causal inference” (Holland, 1986); that is, for each student, depending on which combination of treatments have been selected for his/her school and class, we can observe only one of the six potential outcomes. To tackle the unobservability of a large proportion of individual potential outcomes (i.e. counterfactual outcomes), one statistical approach is to replace population average potential outcomes with observed aggregated outcomes, for example, \( E[Y(Z=1,T=1)] - E[Y(Z=0,T=1)] = E[Y(Z=1,T=1) | z=1, t=1] - E[Y(Z=0,T=1) | z=0, t=1] \). This approach is feasible when stable-unit-treatment-value assumption (SUTVA, Rubin, 1976) holds. SUTVA assumes that that there is a single value of each potential outcome corresponding to each treatment for each experimental unit and the potential outcomes of a unit are independent of the treatments assigned to other units. The assumption, however, is subject to scrutiny in the study of the test-based retention policy. In current educational systems, within the same school/class, it is a hard
to avoid rivalry or sharing of resources between students. And classes/schools that have adopted the same policy may differ in how they implement the policy and how they allocate educational resources. Therefore, there may be multiple potential outcomes associated with each combination of treatments for each individual; and the potential outcomes may depend on the classroom/school that an individual attended as well as on his/her peers that attended the same class/school. SUTVA is unlikely to hold given the potential variations in policy implementation and the interdependency between students/teachers.

**Assumptions for Casual Inference**

To relax SUTVA and to obtain unbiased estimates of the causal effects of the grade-3 test-based retention policy and teacher expectations, I adopted Hong and Raudenbush’s extended causal framework (2006) and refined the casual estimands in a multilevel setting under three assumptions: (a) no interference between schools, (b) intact schools and classes, and (c) ignorable treatment assignment. Although these assumptions are not directly testable with the current observational data, I consider them plausible given the richness of ECLS data and the current nature of educational system.

**No interference between schools.** It is assumed that the treatment assignment of students attending other schools does not affect the potential outcomes of students in a given school. Unlike SUTVA, I allow potential outcomes of each student to be affected by the treatment assignment and identities of students attending the same school and class, but constrain that there is no interference in students’ outcomes between intact schools. Therefore, for student $i$ attending class $j$ of school $k$ receiving treatments $z$ and $t$, the potential outcome is $Y_{ijk} (z_k, t_{jk}, a_k^*, b_{jk}^*)$, where $a^*$ and $b^*$ are the vectors of school and class assignments observed in the sample.

**Intact schools and classes.** I also assume that intact schools were assigned to one of the
two policy conditions and intact classes within each policy group were assigned to different levels of teacher expectations; in other words, students have already been enrolled in the existing schools and classes without being randomly assigned. Hence the average causal effects of interest will be conditioned on students’ current school and class membership. Take the causal estimand for the effects of test-based retention policy under negative expectation $E[Y(Z=1,T=1) - Y(Z=0,T=1)]$ as an example: I modify it as $E[Y(z=1, t=1, a^*, b^*) - Y(z=0, t=1, a^*, b^*) | A = a^*, B = b^*]$.

**Ignorable treatment assumption.** I as well postulate that given the observed covariates, school assignment to standardized testing only versus test-based retention policy was ignorable; and so was class assignment to teacher expectation levels within a school policy condition (Rubin, 1978; Imbens, 2000). In the current study where the two treatments were not assigned at the same level, within all relevant levels of covariates (e.g. class- and school- level covariates, denoted by vectors $W$ and $V$ respectively), ignorable treatment assignment assumption is required for both school assignment and class assignment within schools such that $E[Y(z) | D(z) = 1, V = v] = E[Y(z) | V = v]$ and $E[Y(t, z) | D(t) = 1, D(z) = 1, W = w, V = v] = E[Y(t, z) | D(z) = 1, W = w, V = v]$. Here I use a dummy indicator $D(z)$ to indicate each of the two policy conditions and $D(t)$ to denote each of the three expectations.

**Multilevel Marginal Mean Weighting Through Stratification (MMW-S) Method**

To remove selection bias in estimating the causal effects of the test-based retention policy and teacher expectations, this study applied a newly developed propensity score-based causal inference method, the marginal mean weighting through stratification method (MMW-S, Hong, 2010a, 2011; Huang, Frangakis, Dominici, Diette, & Wu, 2005; Zanutto, Lu, & Hornik, 2005). I extended the method to an evaluation of the multilevel treatments and approximated a
sequential random assignment mechanism.

**Propensity score methods for removing selection bias.** Rosenbaum and Rubin (1983, 1984) introduced the propensity score method where all pretreatment covariates are summarized in a unidimensional balancing score (i.e. propensity score) to characterize the conditional probability of treatment exposure. In the case of a binary treatment, given the propensity score \( \theta_z = Pr (Z=1|X) \), the conditional distribution of all covariates \( X \) should be the same for the treated \((Z = 1)\) and control \((Z = 0)\). When propensity scores balance well on the covariates, with a large number of covariates, the propensity method is superior to analysis of covariance (ANCOVA) or multiple regression because the direct adjustment for covariates may bias treatment effect estimate if the functional form of outcome model is misspecified (Drake, 1993); and the adjustment also constrains the number of pretreatment variables that can be controlled for due to the concern of degrees of freedom for the treatment effect estimation. Among different variations of propensity score methods, stratification or matching based on propensity scores has been frequently practiced, but is usually restricted to evaluations of binary treatment. This is not only because of the cumbersome procedure in simultaneous adjustment for multiple sets of propensity scores but also due to some unresolved statistical issues such as inflated Type I error and potential sparseness of data in many cells (Hong, 2011; Joffe & Rosenbaum, 1999). To adjust for selection bias associated with multi-valued or multiple treatment(s), methods that use propensity scores in weighting overcomes the above limitations by assigning each unit a weight so that the weighted group composition approximates the composition of the whole population; instead of averaging conditional treatment effects across strata or matched pairs, the weighting methods first estimate the population average potential outcome of each treatment (condition) and then compute treatment effects by contrasting the estimated potential outcomes.
**MMW-S and its extension to evaluations of multilevel treatments.** There are two alternative propensity weighting strategies: one is inverse-probability-of-treatment weighting (IPTW), which was proposed by Rosenbaum (1987), developed by Robins and colleagues (Robins, 2000, Robins, Rotnitzky, & Zhao, 1995) and advanced by Imbens (2000); the other is the MMW-S method, which was originated from the field of epidemiology (Huang et al., 2005) and was recently formalized by Hong (2010a) in educational research. With IPTW, the propensity score for treatment group \( z \) (i.e. \( \theta_z \)) is directly used in the denominator with the average probability of treatment assignment as a denominator in computing the weight, i.e. \( Pr(Z = z) / \theta_z \), whereas with MMW-S, the propensity scores are used as a basis for stratification. The MMW-S method, which borrows the idea of post-stratification adjustment (Horvitz & Thompson, 1952) and adjustment for nonresponse through weighting (Little, 1982, 1986), conceptualizes the weight as \( Pr(\theta_z) / Pr(\theta_z | Z = z) \), a comparison of the distribution of the propensity scores in the population to its distribution in the treatment group \( z \). In practice, this weight can be obtained by computing the ratio of the total number of units in each stratum (\( n_s \)) to the number of treated units in that stratum (\( n_{z,s} \)), i.e. \( n_s \times Pr\{Z = z\} \). Hong (2010a) has proven that the two weights are inherently related, but the MMW-S outperforms the IPTW method due to the former’s nonparametric use of propensity scores. The MMW-S strategy not only can avoid the problem of distorted values in the IPTW by assigning a zero weight to units who have no counterfactual information under an alternative treatment condition, but also are robust to misspecifications of the functional form of a propensity model.

MMW-S has recently been applied in several educational studies and has shown promises as a viable solution for estimating treatment effects in single-level and multi-level non-
experimental data (Hong, 2010b; Hong & Hong, 2009; Hong, Corter, Hong, & Pelletier, 2011; Hong, Pelletier, Hong, & Corter, under review). Hong (2010a, 2011) also discussed its potentials for addressing various types of causal questions, including evaluations of multi-dosage or multiple concurrent treatments as well as estimating moderated or mediated treatment effects. But an unexplored scenario is its use for evaluating two treatments assigned at two different levels of the system. In this study, I extended MMW-S to a causal analysis where class-level treatment assignment is at a lower level than and is hence considered sub-sequential to school-level treatment assignment. The essence is to compute two-level propensity scores and conduct two-level stratification by first stratifying schools into a few clusters and then stratifying classes within each cluster. On the basis of the stratification results, two-level marginal mean weights will be computed and applied at corresponding treatment levels. It is expected that the weighted composition of each class-level treatment group within a school cluster will approximate the composition of that cluster and the weighted composition of each school-level treatment group will represent the school-level population composition. Therefore, the weighted marginal mean outcome of treatments \( z \) and \( t \), \( \bar{Y}_{MMW}(Z = z, T = t) \) will be an unbiased estimate of the population average outcome \( E[Y(Z = z, T = t)] \), which makes it possible to estimate the causal estimands through pairwise comparisons among different treatment combinations. Below I describe the procedure in applying the MMW-S to the evaluation of the multilevel treatments.

**Analytic Procedure**

The procedure involves 10 steps in analyzing the effects of the grade-3 test-based retention policy and/or teacher expectations:

1. Estimate school level propensity score. To ensure that the same analytic sample is used throughout the study, I employed a two-stage stepwise selection procedure, first to identify at
school level all outcome predictors from 81 pretreatment variables including school aggregates of students’ academic and motivational pretest scores\textsuperscript{10}, and then to further select confounders that were associated with both outcomes and the school-level policy treatment. Using a school-level binary logistic regression, I estimated the conditional probability that a grade-3 school would adopt the test-based retention policy given the confounders (\(n=11\), see Table D1 of Appendix D).

2. Conduct school-level stratification. From the original 1498 schools, I excluded, based on the logit of the school-level propensity scores, 169 schools\textsuperscript{11} that had a zero probability of adopting one of the two school policies or were hard to find matches in the alternative policy group. I then stratified the remaining schools into four clusters based on the logit scores and made sure that there was no significant within-cluster difference between the two policy conditions in terms of the mean and variance of the logit of propensity scores.

3. Evaluate the relationship between test-based retention policy and teacher expectations within and across the clusters. Based on the school-level stratification, I analyzed a two-level weighted ordinal logistic model with teacher expectations as an outcome at level 1 and the school policy as a predictor at level 2. The results suggest that teacher expectations were independent of the school policy both within and across the school clusters (see Chapter 5 for model specification and results).

4. Construct a separate class level propensity model within each school cluster. In each of the four school clusters, again through a two-stage stepwise procedure, at class level I identified from 81 school-level and 31 class-level covariates a set of confounders that predicted within-cluster teacher expectations (\(n = 5-8\) per cluster). Then using a two-level multinomial logistic model, I estimated within each cluster the propensity that a class would receive each of the three
teacher expectations given the selected school- and class-level confounders as well as the observed school-level assignment. Appendix D explains the rationale for estimating the propensity models at the both school- and class-levels and lists the confounders selected into the different models.

5. Identify a common support for class-level stratification. Employing an Empirical Bayes approach, without breaking the initial balance between $Z=1$ and $Z=0$, I identified in every school cluster a common support where every class will have a nonzero probability of assigning to each teacher expectation and will be able to find match(es) in terms of the pretreatment characteristics across the three expectations regardless of the actual school assignment. Appendix E explains the specific strategy for identifying the common support.

In total, 2015 classes of 781 schools were further removed from the analytic sample identified at step 2. As illustrated in Appendix E (Figure E2), the dramatic sample reduction was mainly due to largely non-overlapping distribution across the three expectations within each cluster. The rather spread-out distribution of the class-level logit scores for each expectation suggests that the measure of teacher expectations may reflect both teachers’ individual traits and their students’ collective ability in a class and also indicates a huge variation in pretreatment characteristics for classes under similar teacher expectation. By removing the classes with unmatchable characteristics, the causal inference procedure enabled me to minimize confounding effects associated with students’ collective ability and to compare among classes with similar student composition and school demographics yet subject to different judgment of students’ academic capability by teachers.

6. Stratify classes within each school cluster. To echo the intended sequential random assignment mechanism, this step was used to simulate randomization at class level. Since school
assignment to the policy conditions was found unrelated to teacher expectations and considered ignorable within clusters, I further stratified the classes of each school cluster into 2-4 strata only based on the logit of the corresponding class-level propensity scores. The stratification allowed me to achieve balance between each expectation group and the rest of the sample on the logit propensity scores within each stratum.

7. Identify the final analytic sample. To ensure consistency in interpreting the results, I decided to utilize the same analytic sample for evaluating the joint effects of the grade-3 test-based retention policy and teacher expectations. Although the common support identified at step 5 was appropriate for estimating overall treatment effects, in order to accommodate the evaluation of moderated treatment effects by student prior ability, the sample had to be further reduced so that at each ability level, students in a focal treated group will have at least a match in the rest of the sample in every stratum of a school cluster. After removing a total of 104 unmatchable students, my final analytic sample contained 2476 students from 1065 classes of 539 schools. As shown in Table 4.4, in this analytic sample, a majority of schools (70.50%) adopted standardized testing only; within each policy condition, most of the classes (65.35%) were assigned to positive-expectation teachers. I summarized in Table 4.5 the frequency of the students at each ability levels within the six treatment conditions in the final sample.

Table 4.6 shows the differences between the K-5 full student sample (n =17565) and the final analytic sample. In comparison with the schools excluded from the analytic sample, my final sample for estimating concurrent treatment effects was dominated by public schools and consisted of more schools in the South region, but fewer schools in the Northeast or Midwest region. It had fewer schools with sixth and above grade, but more urban schools and more schools with relatively larger size. The final analytic sample also had a higher proportion of
disadvantaged students, i.e. black students, Hispanic students and students from non-English-speaking families or families below poverty threshold.

Table 4.4

**Class Distribution across Treatment Conditions in the Final Analytic Sample**

<table>
<thead>
<tr>
<th>School Policy</th>
<th>Number of Classes</th>
<th>(%)</th>
<th>Number of Schools</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Negative Expectation</td>
<td>Indifferent Expectation</td>
<td>Positive Expectation</td>
<td>Total</td>
</tr>
<tr>
<td>Standardized Testing Only</td>
<td>142</td>
<td>117</td>
<td>470</td>
<td>729</td>
</tr>
<tr>
<td></td>
<td>(13.33)</td>
<td>(10.99)</td>
<td>(44.13)</td>
<td>(68.45)</td>
</tr>
<tr>
<td>Test-based Retention</td>
<td>61</td>
<td>49</td>
<td>226</td>
<td>336</td>
</tr>
<tr>
<td></td>
<td>(5.73)</td>
<td>(4.60)</td>
<td>(21.22)</td>
<td>(31.55)</td>
</tr>
<tr>
<td>Total</td>
<td>203</td>
<td>166</td>
<td>696</td>
<td>1065</td>
</tr>
<tr>
<td></td>
<td>(19.06)</td>
<td>(15.59)</td>
<td>(65.35)</td>
<td>(100)</td>
</tr>
</tbody>
</table>

Table 4.5

**Student Distribution across Ability Levels and Treatment Conditions in the Final Analytic Sample**

<table>
<thead>
<tr>
<th>Teacher Expectations</th>
<th>Ability 1</th>
<th>Ability 2</th>
<th>Ability 3</th>
<th>Ability 4</th>
<th>Ability 5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>negative</td>
<td>indifferent</td>
<td>positive</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized Testing Only</td>
<td>103</td>
<td>91</td>
<td>287</td>
<td>481</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ability 2</td>
<td>86</td>
<td>63</td>
<td>229</td>
<td>378</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ability 3</td>
<td>75</td>
<td>57</td>
<td>216</td>
<td>348</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ability 4</td>
<td>59</td>
<td>62</td>
<td>205</td>
<td>326</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ability 5</td>
<td>44</td>
<td>32</td>
<td>178</td>
<td>254</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test-based Retention</td>
<td>Ability 1</td>
<td>28</td>
<td>13</td>
<td>87</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>Ability 2</td>
<td>32</td>
<td>26</td>
<td>114</td>
<td>172</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ability 3</td>
<td>28</td>
<td>27</td>
<td>104</td>
<td>159</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ability 4</td>
<td>24</td>
<td>28</td>
<td>93</td>
<td>145</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ability 5</td>
<td>12</td>
<td>11</td>
<td>62</td>
<td>85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>491</td>
<td>410</td>
<td>1575</td>
<td>2476</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: students’ ability levels ranged on a continuum from 1 to 5 with 1 stands for the lowest ability and 5 for the highest.
### Table 4.6
Comparison between ECLS-K K-5 Full Sample and the Final Analytic Sample

<table>
<thead>
<tr>
<th></th>
<th>Full Sample</th>
<th>Final Analytic Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion of public schools</td>
<td>0.83</td>
<td>0.99***</td>
</tr>
<tr>
<td>Proportion of schools in Northeast region</td>
<td>0.19</td>
<td>0.15*</td>
</tr>
<tr>
<td>Proportion of schools in Midwest region</td>
<td>0.23</td>
<td>0.16***</td>
</tr>
<tr>
<td>Proportion of schools in South region</td>
<td>0.33</td>
<td>0.45***</td>
</tr>
<tr>
<td>Proportion of schools in West region</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>Proportion of urban schools</td>
<td>0.44</td>
<td>0.47*</td>
</tr>
<tr>
<td>Proportion of suburban schools</td>
<td>0.43</td>
<td>0.40</td>
</tr>
<tr>
<td>Proportion of rural schools</td>
<td>0.13</td>
<td>0.14</td>
</tr>
<tr>
<td>Proportion of schools with sixth and above grades</td>
<td>0.21</td>
<td>0.10***</td>
</tr>
<tr>
<td>Proportion of small-size schools (less than 150 students)</td>
<td>0.05</td>
<td>0.01***</td>
</tr>
<tr>
<td>Average age at grade 3 entry</td>
<td>8.46</td>
<td>8.46</td>
</tr>
<tr>
<td>Proportion of girls</td>
<td>0.49</td>
<td>0.49</td>
</tr>
<tr>
<td>Proportion of white</td>
<td>0.69</td>
<td>0.58***</td>
</tr>
<tr>
<td>Proportion of black</td>
<td>0.12</td>
<td>0.18***</td>
</tr>
<tr>
<td>Proportion of Hispanic</td>
<td>0.19</td>
<td>0.24***</td>
</tr>
<tr>
<td>Proportion of Asian</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td>Proportion below poverty</td>
<td>0.19</td>
<td>0.26***</td>
</tr>
<tr>
<td>Proportion speaking non-English at home</td>
<td>0.13</td>
<td>0.20***</td>
</tr>
</tbody>
</table>

Note: the analytic sample was compared with the excluded cases in the full sample  * p<.05. ** p<.01. *** p<.001

8. Compute multilevel weights with the final analytic sample. Appendix F explains how the multilevel weights were computed and presents the corresponding distribution. In brief, based on the distribution of the school clusters in the final analytic sample, I computed school-level marginal mean weights as a ratio of total number of units in each cluster to the number of the treated units in that cluster. When it comes to class-level weights, I treated every school cluster as a pseudo population and similarly obtained class-level weights within each pseudo population. I computed marginal mean weight for each student as a function of the student’s prior ability level, treatment group membership, and stratum membership.
9. Conduct weighted multilevel balance checking. To examine whether the computed multilevel marginal weights balanced pretreatment composition between the weighted treatment groups with the whole analytic sample, I conducted two sets of weighted multilevel balance checking, first on the mean logit scores and then on all the observed covariates. The multilevel weighting adjustment effectively eliminated the mean differences in the logit propensity scores between the initial treatment groups. Adopting a significance level of $p = .05$, I found the two policy groups resembled each other in about 91% of the 81 observed school- covariates; and balance was achieved between the three expectation levels in approximately 94% of the 112 observed school-and class-covariates under standardized testing only and in about 91% of the covariates under test-based retention policy. Appendix G specifies the statistical models used for each set of balance checking and the corresponding results.

10. Estimate the weighted joint effects of the test-based retention policy and teacher expectations. I combined the marginal mean weights with multilevel modeling in estimating the joint treatment effects. Unless specified, in the statistical models for examining the overall effects of the school policy and teacher expectations, I simultaneously applied the school- and class-level weights generated at step 8; when examining the corresponding moderated effects by students’ prior ability, I employed the school- and student-level weights. In order to further reduce bias and improve precision in estimation, I combined the MMW_S method with regression adjustment as suggested by Rubin & Thomas (2000) and controlled in the final outcome models for several important covariates where the MMWs failed to achieve balance. Since my propensity score models only considered school- and class-level covariates, at the student level of the final outcome models, I also controlled for differences in student demographics, including gender, ethnicity, SES status, home language, kindergarten entry age,
ECLS pretest scores and the fall 1998 rating of social and emotional behavior. Appendix H lists all the controlled variables in each outcome model. These variables were selected through systematic model comparisons and were identified as strong predictors of the focal outcome(s) (i.e. prognostic variables).

The specific statistical models are presented together with the results in Chapters 5-8. In those models, besides $Y_{ijk}$ for the observed outcome of student $i$ taught by teacher $j$ in school $k$, $Z$ for school policy, and $T1-T3$ for negative, indifferent, and positive teacher expectations respectively, I used $A1-A5$ to label five ability groups, with 1 for the lowest group and 5 for the highest; in addition, $X$, $W$, or $V$ denote prognostic variables to be controlled for at student, class, and/or school level; correspondingly $nx$, $nw$, and $nv$ refer to their respective number.

The analysis of the policy-by-teacher effects involves 2 x 3 treatment conditions and needs to differentiate the five ability populations (e.g. analysis of the differential effects by student prior ability). To avoid inflated type I error with multiple comparisons, when examining the effects of teacher expectations, I used Bonferroni correction to adjust the significance level of the pairwise comparisons. When examining the effects of the grade-3 test-based retention policy in Chapter 6-8, I began with an omnibus likelihood ratio test through models comparison. That is, I first compared the models with or without the policy predictor $Z$ to roughly detect whether there were any policy effects for students in general and/or for each ability subpopulation. Significant omnibus test results would signal the existence of at least one of the following three possibilities: (a) significant conditional policy effect after the differences between teacher expectations were controlled for, (b) significant policy effect(s) under at least one type of teacher expectations, and/or (c) significant difference(s) in the policy effects across the three teacher expectations (i.e. interaction effects). Following the omnibus likelihood tests, I explored each of
the three possibilities. I examined the conditional effects of the test-based retention policy by constraining the parameters of the policy effects as equal across the three teacher expectations for the whole student population or for each ability subpopulation. I also examined the policy effects by each type of teacher expectations as shown by the multilevel modeling results. In addition, I employed multivariate hypothesis testing to test whether the policy effects across the three expectations were equal to zero. Only when the multivariate hypothesis tests were significant, would I proceed to post-hoc analysis to compare each pair of teacher expectations in terms of the policy effects.

Notes to Chapter 4

1 Compared to schools with standardized testing, schools without the testing seem mainly to be special education schools. They were more likely to provide gifted program service and to locate in south and west regions. They had a smaller proportion of girls and/or LEP students and a smaller amount of black, Hispanic and white students. Their school enrollment was smaller and thus hired less regular teachers. Their students were more likely to come from low-income families, but less likely to be qualified for free or reduced lunch program. The schools tended to suffer more from insufficient educational resources and had fewer computers and regular school facilities. However, safety issues were less problematic in those schools.

2 ECLS dataset did not allow empirical investigation of the actual level where the grade-3 test-based retention policy was initiated. Neither did it provide any information regarding whether the policy was associated with other high-stakes accountability initiatives, such as teacher evaluation and school probation.

3 As teacher expectation was measured at the end of spring 2002, there is a possibility that it would be an outcome of instructional experience with the sampled students. However, as student composition in a school was unlikely to change drastically over time, I assume teachers' perception of student average learning ability was relatively stable during the school year.

4 Reardon (2008) pointed out that the relationship between the IRT scale scores and student ability estimate (θ) is not linear in the ECLS study because the ECLS tests contain more ‘difficult’ items than ‘easy’ items; as a result, this type of scores is difficult to interpret as an interval-scaled metric and thus is not appropriate for making anything other than ordinal comparison. Following Reardon’s advice, I decided to use θ scores provided by ECLS to characterize students’ overall academic growth overtime.

5 Student self-perceived competence and interests actually measures two closely related constructs: student academic conception and their interests. However, at the time of my data analysis, ECLS had not released item-level data from students’ self-description questionnaire and did not allow me to differentiate the two constructs.
6 Student reading skill is an important indicator of their academic ability in the early grades; school districts and/or states with high-stakes testing policy (e.g., Chicago, California, Texas) tended to weigh more on students’ reading/literacy performance for identifying high-performing vs. low-performing schools. Ability classification within school therefore is likely to reflect this emphasis.

7 In current educational practices, students are assigned to schools mostly based on their residence location. Therefore school enrollment and their student demographic composition are unlikely to change overtime, neither as a consequence of grade-specific high-stakes testing policy.

8 An alternative strategy is to consider the two sets of treatment assignment occurred simultaneously, assuming that all intact classes were assigned to the three expectations regardless of the school-level assignment. But I consider a sequential random design fits better with the multilevel structure of the treatments as well as the current research focus on the dependency of the policy effects on teacher expectation.

9 Imbens (2000) stated this as a weak ignorability assumption. Unlike in Rubin’s strongly ignorability assumption (1978), this assumption does not require each treatment to be independent of the entire set of potential outcomes. Instead, it only assumes pairwise independence of the treatments with each of the potential outcomes.

10 Since students’ prior academic achievement and motivation are often regarded as among the strongest confounders that predict any educational outcome (Cook & Steiner, 2010), in each level of variable selection, I included the corresponding aggregates of students’ spring 1998 academic and motivational scores. In view of limited number of students sampled within each school and potential missing not at random, I as well considered the proportion of missingness.

11 Among the 169 schools excluded from the school-level stratification, 166 were from standardized testing only condition and 3 were schools with test-based retention policy.

12 Among the 2015 classes excluded from the class-level stratification, 184 were from the negative expectation group, 230 were from the indifferent expectation group, and 1601 from positive expectation group.

13 With the final analytic sample, after conditioning on a full set of classroom and school pretreatment variable, the within-school variance in teacher expectations was .63, accounting for 99.74% of the total variance.

14 According to Cochran (1968), by dividing the distribution of pretreatment covariates into five subclasses, we can remove about 90% bias associated with the pretreatment covariates. However, to preserve more units in each class-level strata and to ensure sufficient statistical power for estimating treatment effects, I tried to stratify schools/classes into as few as possible number of clusters/strata on the condition that the stratification can balance the corresponding logit propensity scores between treatment groups within each cluster/stratum. Final balance checking suggests that this approach was able to remove more than 90% of the bias either between or within school policies.

15 To evaluate the relationship between the test-based retention policy and teacher expectations, I used the sample identified at step 2. It includes 3101 classes from 1329 schools.

16 The removal of 104 students did not significantly change the sample composition. Neither did it decrease the effectiveness in reducing the selection basis. Multilevel balance checking results are similar
with or without the 104 students.

Although more than 5% of the pretreatment covariates showed statistically significant difference either between the policy groups or between the expectation levels within each policy condition, further analysis indicated that among the 7 covariates that failed to balance $Z = 1$ and $Z = 0$, none was significant predictors of the outcomes and treatments; under $Z = 0$, there were 6 significant confounders out of the 7 unbalanced covariates; under $Z = 1$, only 2 out of the 10 unbalanced covariates could confound the treatment effects estimation.
CHAPTER 5
GRADE-3 TEST-BASED RETENTION POLICY AND TEACHER EXPECTATIONS

Chapter 5 analyzes the relationship between the grade-3 test-based retention policy and teacher expectations. In addition to the causal analysis results, it also provides a brief description of the pretreatment characteristics that predict the treatments assignment.

Association with Pretreatment Characteristics

In the ECLS sample, the schools or classes were not randomly assigned to the corresponding treatment conditions. Here I describe the pretreatment differences between schools/classes that adopted different treatment conditions. The description was outlined based on the bivariate association between pretreatment covariates and the school- or class-level treatment.

Pretreatment Characteristics Predicting Grade-3 Test-based Retention Policy

In general I found that schools with the test-based retention policy were often from suburban area and located in West or Middle West region. The policy was more prevalent in private and catholic schools, which explains why schools with the policy were less likely to participate in USDA (U.S. Dept. of Agriculture) breakfast program and often not subject to Title I funds options. In addition, schools were more likely to adopt the test-based retention policy if having a higher concentration of disadvantaged students, such as minority students (especially black and Hispanic students), students qualified for reduced-price or free lunch program, students with limited English proficiency, and students with lower general knowledge scores at the kindergarten entry. These schools usually involved sixth and above grades, and hence had a larger student enrollment and more regular teachers. They tended to suffer from limited educational resources (e.g. smaller school site and no service for students with special needs),
problem of student mobility, and/or serious violence and crime within and around.

**Pretreatment Characteristics Predicting Teacher Expectations**

Compared to other classes, classes taught by positive-expectation teachers had a lower concentration of disadvantaged students (e.g. minority student, LEP students, above-grade-age students, and students from low SES background); their students had better academic performance and/or social skills at the kindergarten entry. This type of classes was less frequently found in public schools, or in schools with more severe problems of student mobility or violence within and around. Instead, they were prevalent in schools with smaller student enrollment or in schools with more sufficient educational resources. Classes with negative- or indifferent-expectation teachers were mostly the same in their student composition and school characteristics; but the classes subject to indifferent expectation were exposed to relatively better school orders; they were usually identified in Northeastern or Western schools and/or suburban schools while classes with negative expectation were mostly from urban schools.

**Relationship between Grade-3 Test-based Retention Policy and Teacher Expectations**

Did the policy have any influence on teachers’ expectations of student learning ability? Table 4.4 in Chapter 4 has shown the distribution of the classes across the six treatment conditions. To evaluate the relationship between the grade-3 test-based retention policy and teacher expectations, I used a two-level ordinal logistic model with teachers at level 1 and schools at level 2. I controlled for several important prognostic variables, $W$ at class level and $V$ at school level (see Appendix H for the list of variables). Let $T_{jk}$ be the three categories of teachers’ expectations, taking on a value of $m$ ($m = 1-3$, denoting the negative, indifferent, or positive expectation respectively). I specified its cumulative probabilities as:
\[
\phi^*_m = \text{Prob}(Y_{mjk} = 1) = \phi_{tjk} + \ldots + \phi_{mjk}, \text{ where } Y_{mjk} = 1 \text{ if } T_{jk} \leq m,
\]

\[
Y_{mjk} | \phi_{mjk} \sim [\phi_{mjk}, \phi_{mjk} (1 - \phi_{mjk})]. \quad (5.1)
\]

With the cumulative probabilities, I posed the following level-1 structural model:

\[
\eta_{mjk} = \log \left( \frac{\phi_{mk}^*}{1 - \phi_{mk}^*} \right) = \beta_{ok} + \sum_{p=1}^{m} \beta_{pk} W_{pk} + D_{2jk} \delta_{2k}, \text{ where } D_{2jk} \text{ is an indicator for } m = 2. \quad (5.2)
\]

To remove the selection bias associated with school assignment, I applied the school-level marginal mean weight at level 2:

\[
\beta_{ok} = \gamma_{00} + \gamma_{01} Z_k + \sum_{g=2}^{g+1} \gamma_{0g} V_k + u_{0k}, \quad u_{0k} \sim N(0, \tau_{00}), \quad (5.3)
\]

\[
\delta_{2k} = \delta_2.
\]

The weighted analysis did not find any significant relationship between the grade-3 test-based retention policy and teacher expectations (coefficient = -.06, SE = .15, t = -.42, p > .5)\(^2\).

Alternatively, I specified an unweighted level-2 model with direct adjustment for the dummy indicators of the four school-level clusters denoted by C1-4:

\[
\beta_{ok} = C_{1k} \ast (\gamma_{01} + \gamma_{02} \ast Z_k) + C_{2k} \ast (\gamma_{03} + \gamma_{04} \ast Z_k) + C_{3k} \ast (\gamma_{05} + \gamma_{06} \ast Z_k) + C_{4k} \ast (\gamma_{07} + \gamma_{08} \ast Z_k) + \sum_{g=5}^{g+1} \gamma_{0g} V_k + u_{0k}, \quad u_{0k} \sim N(0, \tau_{00}), \quad (5.4)
\]

\[
\delta_{2k} = \delta_2.
\]

The results from the alternative model suggested that the two policy groups did not differ significantly by teacher expectations in any school cluster: coefficient = -.32, SE = .38, t = -.85, p = .40 in cluster 1, coefficient = -.05, SE = .21, t = -.23, p > .5 in cluster 2, coefficient = .15, SE = .18, t = .81, p = .42 in cluster 3, and coefficient = -.20, SE = .28, t = -.70, p = .48 in cluster 4.

Since I failed to reject the hypothesis \( \gamma_{02} = \gamma_{04} = \gamma_{06} = \gamma_{08} = 0 \) (\( \chi^2 = 1.94, \text{ df} = 4, p > .5 \)), I constrained the four slopes as equal and confirmed no significant relationship between the two
treatments for the whole analytic sample (coefficient = -.03, \( SE = .12 \), \( t = -.27 \), \( p > .5 \)). The results were consistent with those produced by the weighted analysis.

**Summary**

Through examining the bivariate association between the pretreatment covariates and school- or class-level treatment, I revealed the systematic differences in pretreatment characteristics for adopting one treatment versus another. The differences signaled the existence of the selection bias and justified the use of causal inference strategies for the current study.

When evaluating the relationship between the grade-3 test-based retention policy and teacher expectations, the two alternative models generated consistent results and confirmed my hypothesis that the grade-3 test-based retention policy had no influence on teachers’ perception of their students’ learning ability.

**Notes to Chapter 5**

1. Note that the sample for this part of analysis was defined at step 2 of the analytic procedure that I described in the Methods Chapter. It includes 3101 classes from 1329 schools, larger than what I used for the analysis in Chapter 6-8 (i.e. 1065 classes from 539 schools). The school-level marginal mean weights used here were computed based on the strata sizes identified at that stage. Hence they were slightly different from what I presented in Table F1 of Appendix F. But the both sets of school-level weights were used to serve the same purpose, i.e. to remove selection bias associated with the school assignment for the corresponding analytic sample.

2. I also tested this model with the final analytic sample and the final school-level marginal mean weight. The conclusion remains the same (coefficient = -.11, \( SE = .15 \), \( t = -.73 \), \( p > .5 \)).
CHAPTER 6
EFFECTS OF TEST-BASED RETENTION POLICY AND TEACHER EXPECTATIONS ON INSTRUCTIONAL TIME ALLOCATION

Chapter 6 highlights the amount of instructional time teachers allocated to different subjects per week and examines whether and how it was subject to the joint influence of the test-based retention policy and teacher expectations.

Instructional Time Allocation in Each Subject

Did the grade-3 test-based retention policy lead to any change in teachers’ instructional time allocation? Did the policy effects depend on teacher expectations? To answer these questions, for each subject I used a two-level weighted univariate model where classes were specified at level 1 and schools were at level 2; each level was adjusted by applying the corresponding marginal mean weights and by controlling for some important prognostic variables, i.e. $W$ for class-level prognostics and $V$ for school-level ones. Considering that less than half of the schools had two or more classes sampled, to allow more degrees of freedom, I assumed a constant within-school variance by using a pseudo-intercept at level 1 and specifying it as a random error at level 2.

Level 1:

$$Y_{jk} = \beta_{0k} + \beta_{1k}(T1)_{k} + \beta_{2k}(T2)_{k} + \beta_{3k}(T3)_{k} + \sum_{p=4}^{mv+3} \beta_{pk}W_{k} + e_{jk}, \; e_{jk} \sim N(0, \sigma^{2}),$$

(6.1)

Level 2:

$$\beta_{0k} = u_{0k}, \; u_{0k} \sim N(0, \tau_{\beta})$$

$$\beta_{hk} = \gamma_{h0} + \gamma_{h1}Z_{k} + \sum_{g=2}^{mv+1} \gamma_{hg}V_{k}, \; \text{for } h = 1 - 3$$

$$\beta_{pk} = \gamma_{p0}$$

(6.2)
**Table 6.1**

*Weighted Analysis of the Policy-by-Teacher Effects on Instructional Time Allocation*

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Reading Time</th>
<th>Math Time</th>
<th>Science Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff.</td>
<td>SE</td>
<td>t</td>
</tr>
<tr>
<td><strong>Negative expectation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control Condition</td>
<td>379.01</td>
<td>13.22</td>
<td>28.66***</td>
</tr>
<tr>
<td>Policy Effect</td>
<td>-22.20</td>
<td>36.00</td>
<td>-0.62</td>
</tr>
<tr>
<td><strong>Indifferent expectation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Policy Effect</td>
<td>-7.29</td>
<td>27.74</td>
<td>-0.26</td>
</tr>
<tr>
<td><strong>Positive expectation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control Condition</td>
<td>399.80</td>
<td>6.75</td>
<td>59.19***</td>
</tr>
<tr>
<td>Policy Effect</td>
<td>6.83</td>
<td>12.69</td>
<td>0.54</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effect</th>
<th>Variance</th>
<th>df</th>
<th>$\chi^2$</th>
<th>Variance</th>
<th>df</th>
<th>$\chi^2$</th>
<th>Variance</th>
<th>df</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1, r</td>
<td>14035.31</td>
<td></td>
<td></td>
<td>8130.08</td>
<td></td>
<td></td>
<td>3627.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2, $u_0$</td>
<td>3441.68</td>
<td>539</td>
<td>799.94***</td>
<td>2152.07</td>
<td>539</td>
<td>819.61***</td>
<td>1736.37</td>
<td>539</td>
<td>1048.04</td>
</tr>
</tbody>
</table>

Note: * $p<.05$. ** $p<.01$. *** $p<.001$. 
Figure 6.1. Estimated effects of the test-based retention policy and teacher expectations on instructional time by subjects.
Effects of teacher expectations. The estimated main effects of teacher expectations are illustrated in Figure 6.1. To understand the effects of teacher expectations in the standardized testing only condition, I conducted pairwise comparisons of $\gamma_{h0}$, adopting a Bonferroni alpha level of .017. I found that when no high-stakes policy was imposed, teachers mostly resembled each other in their decisions on how to allocate instructional time except that positive-expectation teachers assigned 40.94 more minutes per week ($SE = 14.07, \chi^2 = 8.47, p < .01$) to teaching reading than indifferent-expectation teachers; the discrepancy approximated one third of a standard deviation of the population reading time.

To examine the effects of teacher expectations in the test-based retention condition, I replaced the predictor $Z$ with 1-$Z$ in the equation 6.2 and then used a similar strategy to compare $\gamma_{h0}$ in each of the three subjects. The pairwise comparisons suggested no difference between the three types of teachers in terms of their instructional time allocation in the high-stakes context.

Effects of the grade-3 test-based retention policy. To evaluate the effects of the grade-3 test-based retention policy, I first performed omnibus likelihood ratio tests to compare the model specified in Equation 6.1-6.2 with the model without the policy predictors $Z$. The results implied significant policy teacher effects in math ($\chi^2 = 8.94, df = 3, p < .05$), but not in reading ($\chi^2 = 1.53, df = 3, p > .5$) or in science ($\chi^2 = .61, df = 3, p > .5$).

As shown in Table 6.1 and Figure 6.1, the test-based retention policy in general did not change significantly the reading and science instructional time. But it produced a notable rise in the math instructional time across the three expectations: negative-expectation teachers raised their weekly math time by 25.18 minutes ($SE = 22.38, t = 1.13, p = .26$), indifferent-expectation teachers raised by 34.84 minutes ($SE = 19.44, t = 1.79, p = .07$) and positive-expectation teachers by 21.47 minutes ($SE = 11.62, t = 1.85, p = .07$) – each respectively accounted for about 34%,
47%, and 29% of one standard deviation of the population math time. After constraining $\gamma_{hi}$ as equal in the math outcome model (see Equation 6.1-6.2), I found a significant conditional policy effect (coefficient = 23.80, $SE = 10.06$, $t = 2.37$, $p < .05$) on math instructional time. It suggested that controlling for teacher expectations, the policy on average significantly increased the time for math instruction by 23.80 minutes per week; and the effect size was about one third of the population standard deviation. To be consistent, I also constrained $\gamma_{hi}$ in the other two subject outcomes models: the estimated conditional policy effects were not significant: coefficient = .13, $SE = 13.29$, $t = .01$, $p > .5$ for reading instructional time, coefficient = -4.11, $SE = 8.63$, $t = .48$, $p > .5$ for science instructional time.

**Moderating role of teacher expectations.** To examine the role of teacher expectations in the relationship between the test-based retention policy and teachers’ instructional time allocation, I further employed multivariate hypothesis testing of $\gamma_{i1} = \gamma_{i2} = \gamma_{i3} = 0$ for the analysis of each subject. None of the comparisons was significant, indicating that the policy effects on the instructional time allocation were independent of the teacher expectations.

**Between-subjects Comparison of the Policy-by-Teacher Effects**

Did the policy-by-teacher effects vary among different subjects, especially between tested (i.e. reading and math) and nontested (i.e. science) subjects? I addressed the question with a two-level weighted multivariate model, using dummy subject indicators $Dread$, $Dmath$, and $Dsci$ to combine the three class-level outcomes and specifying heterogeneous level-1 variance with two of the three indicators:
Level 1
\[
Y_{mjk} = \text{Dread}_{mjk} * (\beta_0 + \beta_1 (T1)_{jk} + \beta_2 (T2)_{jk} + \beta_3 (T3)_{jk} + \sum_{h=4}^{n_w+3} \beta_{hk} W_{hk} ) + D\text{math}_{mjk} * (\beta_{(n_w+4)k} + \beta_{(n_w+5)k} (T1)_{jk} + \beta_{(n_w+6)k} (T2)_{jk} + \beta_{(n_w+7)k} (T2)_{jk} + \sum_{h=n_w+8}^{2n_w+7} \beta_{hk} W_{hk} ) \]
\[+ D\text{sci}_{mjk} * (\beta_{(2n_w+8)k} + \beta_{(2n_w+9)k} (T1)_{jk} + \beta_{(2n_w+10)k} (T2)_{jk} + \beta_{(2n_w+11)k} (T2)_{jk} + \sum_{h=2n_w+12}^{3n_w+11} \beta_{hk} W_{hk} ) \]
\[+ e_{mjk}, \ e_{mjk} \sim N(0, \sigma^2) ; \]
\[
\log(\sigma_{mjk}^2) = \alpha_{0jk} + \alpha_{1jk} * \text{Dread}_{mjk} + \alpha_{2jk} * \text{Dmath}_{mjk} \]

(6.3)

(6.4)

At level 2, I used a same school-level sub-model for each outcome as in Equation 6.2.

In either policy condition, reading instruction always received the most attention while science received the least regardless of teacher expectations (see Figure 6.2). According to multivariate hypothesis testing, the policy effects were significantly different between subjects for the negative expectation (\(\chi^2 = 9.85, df = 3, p < .05\)) and positive expectation (\(\chi^2 = 9.79, df = 3, p < .05\)), and marginally significant for the indifferent expectation (\(\chi^2 = 7.69, df = 3, p = .05\)).

Follow-up post-hoc analyses revealed that the policy effects were significantly larger on the math instruction than the science instruction regardless of teacher expectations: \(\chi^2 = 3.81, \ p < .05\) under the negative expectation, \(\chi^2 = 3.95, \ p < .05\) under the indifferent expectation, and \(\chi^2 = 9.05, \ p < .01\) under the positive expectation; The effect was also larger on the math instruction than that on the reading instruction under the negative expectation (\(\chi^2 = 7.59, p < .01\)) or the indifferent expectation (\(\chi^2 = 5.64, p < .05\)).
Figure 6.2. Between-subjects comparison of the estimated policy-by-teacher effects on instructional time allocation.

Summary

Phase II analysis revealed that the test-based retention policy uniformly directed teachers’ attention to math instruction. It confirmed my hypothesis that all teachers will increase the amount time for teaching math under the high-stakes pressure. The extra time investment in math was not accompanied by dramatic changes in the instructional time for reading and science. However, the results from the between-subjects comparisons showed that the changes in math time were significantly larger than those in science time under each teacher expectation. They were also larger than those in reading time under negative or indifferent expectation.

In addition, the comparison of the three teacher expectations suggested that the expectation effects depended on the policy context. In standardized testing only condition, positive-expectation teachers spent more time on reading instruction than indifferent-expectation
teachers. However, there was no statistically significant difference between teachers in terms of their time use as a response to the test-based retention policy.

**Notes to Chapter 6**

1 For the whole ECLS grade 3 sample, the standard deviation for reading, math, and science instructional time is 133.16, 73.51, and 72.73 respectively.

2 The multivariate model includes three slopes at level 1, but less than half of schools had 2 or more classes sampled and were able to contribute to the estimation of variance-covariance matrix. Although point estimates from the multivariate model are similar to those from the univariate models, I refer to the univariate analysis results for treatment effect estimates (both in the table and figure) and only use multivariate models to test the significance between the effects across subjects. The same rationale and principle apply to the reporting of findings in Chapter 7 and 8.
CHAPTER 7
EFFECTS OF TEST-BASED RETENTION POLICY AND TEACHER EXPECTATIONS ON STUDENT ACADEMIC PERFORMANCE

Chapter 7 is devoted to analyzing the effects of the grade-3 test-based retention policy and teacher expectations on student academic performance. Corresponding to the research questions 3, I present the analysis results in four sections: (a) student overall academic performance, (b) student mastery of cognitive skills, (c) student differential academic performance by prior ability, and (d) student long-term academic performance.

Student Overall Academic Performance

This section focuses on student overall academic performance as measured by reading, math, and science standardized T scores.

Student Overall Academic Performance in Each Subject

I first asked: did the policy and teacher expectations jointly affect student overall academic performance in the promotional gate year? I started the investigation with a weighted three-level model for each subject. I specified level 1 as a student-level cross-sectional model instead of a growth model to avoid control of any potential policy outcomes between kindergarten and grade 3.

Level 1:
\[ Y_{ijk} = \pi_{o,j,k} + \sum_{l=1}^{n_s} \pi_{l,j,k} X_{ij,k} + e_{ij,k}, \quad e_{ij,k} \sim N(0, \sigma^2) \]  

(7.1)

At each level I controlled for several important prognostic variables \( X, W \) or \( V \) (see Appendix H for the list of the variables). Levels 2 and 3 included 539 classes from 1065 schools and were weighted by the class- and school-level marginal mean weights. Considering that less than half of the schools had two or more classes sampled, to allow more degrees of freedom, I included a
pseudo-intercept at level 2 and specified it as a random error at level 3. By doing so, I assumed a constant within-school variance between classes.

Level 2:
\[ \pi_{jk} = \beta_{00k} + \beta_{01k} (T1)_{jk} + \beta_{02k} (T2)_{jk} + \beta_{03k} (T3)_{jk} + \sum_{p=4}^{m+3} \beta_{0pk} W_{jk} + r_{0jk}, r_{0jk} \sim N(0, \tau_n), \]  
\[ \pi_{ijk} = \beta_{10k} \]  

Level 3:
\[ \beta_{00k} = u_{00k}, u_{00k} \sim N(0, \tau) \]
\[ \beta_{0hk} = \gamma_{0h0} + \gamma_{0h1} Z_k + \sum_{g=2}^{m+1} \gamma_{0hg} V_k, \text{ for } h=1-3 \]
\[ \beta_{0pk} = \gamma_{0p0} \]
\[ \beta_{10k} = \gamma_{100} \]  

Below I describe the results separately for the effects of teacher expectations, the effects of the grade-3 test-based retention policy, and the moderating role of teacher expectations.

**Effects of teacher expectations.** To examine the effects of teacher expectations in each subject, I conducted pairwise comparisons between the three expectations first in the standardized testing only condition and then in the test-based retention condition. For the first set of comparisons, I employed the model as specified in Equations 7.1-7.3. For the second set of comparisons, I replaced \( Z \) with \( 1-Z \) in the specified model. Adopting a Bonferroni adjusted alpha level of .017, the comparisons of \( \gamma_{0h0} \) did not reveal any significant differences in student overall academic performance between teacher expectations in either policy condition. The estimated effects of the test-based retention policy and teacher expectations are illustrated in Figure 7.1.

**Effects of the grade-3 test-based retention policy.** To examine the effects of the grade-3 test-based retention policy, I first compared the models with or without the policy predictors. The omnibus likelihood ratio tests were significant in reading \( (\chi^2 = 7.84, df = 3, p < .05) \) and science \( (\chi^2 = 9.88, df = 3, p < .05) \), but not in math \( (\chi^2 = 1.49, df = 3, p > .5) \).
Figure 7.1. Estimated effects of the test-based retention policy and teacher expectations on student overall academic performance by subjects.
### Table 7.1

**Weighted Analysis of the Policy-by-Teacher Effects on Student Overall Academic Performance**

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Reading</th>
<th></th>
<th>Math</th>
<th></th>
<th>Science</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff.</td>
<td>SE</td>
<td>t</td>
<td>Coeff.</td>
<td>SE</td>
</tr>
<tr>
<td>Negative expectation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control Condition</td>
<td>50.00</td>
<td>0.41</td>
<td>120.54***</td>
<td>49.82</td>
<td>0.48</td>
</tr>
<tr>
<td>Policy Effect</td>
<td>-1.06</td>
<td>0.70</td>
<td>-1.51</td>
<td>-0.11</td>
<td>0.66</td>
</tr>
<tr>
<td>Indifferent expectation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control Condition</td>
<td>48.97</td>
<td>0.43</td>
<td>114.61***</td>
<td>49.04</td>
<td>0.46</td>
</tr>
<tr>
<td>Policy Effect</td>
<td>0.99</td>
<td>0.70</td>
<td>1.41</td>
<td>-0.64</td>
<td>0.92</td>
</tr>
<tr>
<td>Positive expectation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control Condition</td>
<td>49.86</td>
<td>0.24</td>
<td>210.94***</td>
<td>49.80</td>
<td>0.26</td>
</tr>
<tr>
<td>Policy Effect</td>
<td>-0.98</td>
<td>0.63</td>
<td>-1.55</td>
<td>0.41</td>
<td>0.52</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effect</th>
<th>Variance</th>
<th>df</th>
<th>$\chi^2$</th>
<th>Variance</th>
<th>df</th>
<th>$\chi^2$</th>
<th>Variance</th>
<th>df</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1, $e$</td>
<td>40.23</td>
<td></td>
<td>41.37</td>
<td></td>
<td></td>
<td></td>
<td>37.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2, $r_0$</td>
<td>0.02</td>
<td>523</td>
<td>461.24</td>
<td>0.02</td>
<td>521</td>
<td>517.99</td>
<td>0.02</td>
<td>522</td>
<td>500.26</td>
</tr>
<tr>
<td>Level 3, $u_{00}$</td>
<td>5.55</td>
<td>539</td>
<td>857.11***</td>
<td>6.28</td>
<td>539</td>
<td>883.12***</td>
<td>5.42</td>
<td>539</td>
<td>911.72***</td>
</tr>
</tbody>
</table>

Note: * $p<.05$. ** $p<.01$. *** $p<.001$. 
Following the omnibus likelihood ratio tests, I explored whether there existed any
conditional policy effects given teacher expectations. I analyzed the model specified in Equation
7.1-7.3 by constraining $\gamma_{oh1}$ as equal for $h = 1-3$. The results showed that controlling for
teacher expectations, the test-based retention policy significantly decreased student science
performance in general (coefficient = -.93, $SE = .42$, $t = -2.22$, $p < .05$), but did not affect student
overall math performance (coefficient = .16, $SE = .46$, $t = .36$, $p > .5$) or reading performance
(coefficient = -.74, $SE = .51$, $t = -1.44$, $p = .15$).

I further analyzed the policy effects by teacher expectations. As shown in Table 7.1,
under the indifferent expectation, there was no significant policy effect in any subject. Under the
negative expectation, the policy appeared to exert no statistically significant influence on student
learning. However, its negative effect in science (coefficient = -1.17, $SE = .66$, $t = -1.78$, $p = .08$)
accounted for about 12% of a standard deviation of the population outcome; and the negative
effect in reading was about 11% of the population standard deviation. The effects sizes might
not be negligible because the negative-expectation group had a smaller sample size and a larger
standard error than the positive-expectation group and thus might not have enough power to
reach the significance level. Under the positive expectation, I detected significant policy effects
only on students’ overall science learning (coefficient = -1.20, $SE = .52$, $t = -2.28$, $p < .05$). The
effect size was equivalent to 12% of the population standard deviation.

Moderating role of teacher expectations. Subsequently I inspected the differences in the
policy effects across the three teacher expectations. Multivariate hypothesis testing of
$\gamma_{011} = \gamma_{021} = \gamma_{031} = 0$ suggested a significant role of teacher expectations in modifying the policy
effects in reading ($\chi^2 = 8.53$, $df = 3$, $p < .05$) and science ($\chi^2 = 8.99$, $df = 3$, $p < .05$), but not in
math ($\chi^2 = 1.60$, $df = 3$, $p > .5$). I therefore followed up with post-hoc tests only for student
reading and science performance.

In reading, the policy effects under the positive and negative expectations were similar to each other (contrast = .07, SE = .85, \( \chi^2 = .01, p > .5 \)), but were significantly worse than that under the indifferent expectation. The contrast in the policy effects between the negative and indifferent expectations was -2.05 points (SE = .92, \( \chi^2 = 4.91, p < .05 \)); and the contrast between the positive and indifferent expectations was -1.97 points (SE = .77, \( \chi^2 = 6.50, p < .05 \)).

In science, the indifferent expectation also showed an advantage over the positive expectation (contrast = 1.93, SE = .93, \( \chi^2 = 4.34, p < .05 \)) in terms of sheltering students from the negative policy effects. However, the policy effect under the indifferent expectation did not differ from that under the negative expectation (contrast = 1.90, SE = 1.00, \( \chi^2 = 3.63, p = .05 \)). There was no significant difference in the policy effects between the positive and negative expectations either (contrast = -.02, SE = .78, \( \chi^2 = .00, p > .5 \)).

**Between-subject Comparison of the Policy-by-Teacher Effects**

I then asked: How did the policy-by-teacher effects differ between different subjects? To address this question, I employed a weighted three-level multivariate model, again, applying the marginal mean weights at class and school levels. At level 1, I used three dummy indicators of reading (Dread), math (Dmath), and science (Dsci) to combine the three outcomes. To allow more degrees of freedom, this model assumes homogenous level-1 variance between subjects.

Level 1:

\[
Y_{mijk} = Dread_{mijk} * (\pi_{1j} + \sum_{h=2}^{n_1-1} \pi_{hjk} X_{ijk}) + Dmath_{mijk} * (\pi_{1(j+2)} + \sum_{h=n_1+2}^{2n_2+2} \pi_{hjk} X_{ijk}) + Dsci_{mijk} * (\pi_{1(2n_3+3)} + \sum_{h=2n_3+3}^{3n_3+3} \pi_{hjk} X_{ijk}) + e_{mijk}, \quad e_{mijk} \sim N(0, \sigma^2)
\]

At levels 2 and 3, for each outcome I used the same sub-models for each outcome as in
Equations 7.2 & 7.3.

Figure 7.2 compares student academic performance between subjects by teacher expectations. Between-subjects differences in the policy effects existed for the negative ($\chi^2 = 11.15$, $df = 3$, $p < .05$) and positive expectation ($\chi^2 = 13.55$, $df = 3$, $p < .01$), but not for the indifferent expectation ($\chi^2 = 5.41$, $df = 3$, $p = .14$).

![Figure 7.2. Between-subjects comparison of the estimated policy-by-teacher effects on student overall academic performance.](image)

If a teacher believed that most of his/her students were able to learn, the policy effect on the students’ math learning, albeit trivial, would be significantly different from the policy effect on the science ($\chi^2 = 7.68$, $df = 1$, $p < .01$) or reading learning ($\chi^2 = 5.00$, $df = 1$, $p < .05$). If a teacher held a negative view of his/her students’ capability to learn, the students’ science learning would be more subject to the negative policy effect compared to their math learning ($\chi^2 = 7.05$, $df = 1$, $p < .01$). Only when a teacher was indifferent to the students’ learning ability, would the policy have uniformly zero effect on the students’ learning across subjects.
Student Mastery of Cognitive Skills

In this part of analysis, I evaluated the policy-by-teacher effects on students’ mastery of different cognitive skills of tested subjects, i.e. reading and math.

Student Mastery of Specific Cognitive Skills

What were the joint effects of test-based retention policy and teacher expectations on students’ learning of each cognitive skills of reading/math? To answer the question, I first analyzed students’ mastery of each specific skill through comparing the proportion of students at or above a certain proficiency level with the proportion of those below the level.

For each specific skill, I used a three-level binary logistic regression model with students at level 1, classes at level 2 and schools at level 3, applying marginal mean weights at class and school levels. Although ECLS designed the cognitive skills of each subject as hierarchical, I did not employ ordinal logistic regression in the current analysis. By using only one equation across all skill levels within a subject, an ordinal model assumes the policy effect on the acquisition rate of a lower level skill is the same as that of a higher level skill. The use of binary logistic regression enables me to relax this assumption.

Here \( Y_{ijk} \) takes a value of 1 if student \( i \) of teacher \( j \) from school \( k \) has passed the targeted proficiency level or higher levels, or is equal to 0 otherwise.

Level 1

\[
\eta_{ijk} = \log \left( \frac{Pr(Y_{ijk} = 1)}{1 - Pr(Y_{ijk} = 1)} \right) = \pi_{0jk} + \sum_{l=1}^{n} \pi_{ljk} X_{ijk}
\]

Level 2

\[
\pi_{0jk} = \beta_{00k} + \beta_{01k}(T1)_{jk} + \beta_{02k}(T2)_{jk} + \beta_{03k}(T3)_{jk} + \sum_{p=4}^{n+3} \beta_{p0k} W_{jk} + r_{0jk},
\]

\[
\pi_{ljk} = \beta_{l0k}, \quad r_{0jk} \sim N(0, \tau_{\pi})
\]
Below I describe the results from the weighted binary logistic model for the effects of teacher expectations, the effects of the grade-3 test-based retention policy, and the moderating role of teacher expectations.

**Effects of teacher expectations.** Table 7.2 presents the estimated odds for mastering each cognitive skill by the six treatment conditions. Using the model specified in Equation 7.5-7.7, I first examined the effects of teacher expectations in the standardized testing only condition. Adopting an adjusted alpha level of .017, pairwise comparisons of $\gamma_{0h0}$ for $h = 1-3$ only suggested an advantage of the positive expectation over the indifferent expectation in acquiring the extrapolation skill in reading (contrast = .36, $SE = .15$, $\chi^2 = 6.05$, $p = .013$).

After substituting $1-Z$ for the predictor $Z$ in the specified model, I then compared the skill acquisition rate across teacher expectations in the test-based retention condition. As shown in Table 7.2, among all the expectations, the negative expectation was associated with the highest learning rate of two skills that are below the grade-3 curriculum standard, i.e. words comprehension in reading and addition/subtraction in math, though I found statistically significant difference only in words comprehension between negative and indifferent expectations (contrast = 1.53, $SE = .55$, $\chi^2 = 7.72$, $p < .01$). In contrast, positive expectation seemed to foster the learning of math advanced skills in the high-stakes context. Pairwise
comparisons suggested that students were more likely to learn rate and measurement under the positive expectation than under the negative expectation (contrast = 1.50, \(SE = .58, \chi^2 = 6.74, p < .01\)). In the test-based retention condition, the indifferent expectation seemed to differ from the other expectations in students’ learning of grade-level math skills: students would be less likely to learn multiplication/division under the indifferent expectation than under the negative (contrast = -1.59, \(SE = .40, \chi^2 = 15.96, p < .001\)) or positive expectation (contrast = -1.09, \(SE = .40, \chi^2 = 7.51, p < .01\)).

Table 7.2

*Estimated Odds of Skill Mastery Corresponding to the Policy-by-Teacher Treatments Conditions*

<table>
<thead>
<tr>
<th></th>
<th>Negative Expectation</th>
<th>Indifferent Expectation</th>
<th>Positive Expectation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Odds</td>
<td>n</td>
<td>Odds</td>
</tr>
<tr>
<td><strong>Reading Cognitive Skills</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lowest Reading Proficiency</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized testing only</td>
<td>--</td>
<td>67</td>
<td>--</td>
</tr>
<tr>
<td>Test-based retention</td>
<td>--</td>
<td>25</td>
<td>--</td>
</tr>
<tr>
<td>Comprehension of words in context</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized testing only</td>
<td>36.92</td>
<td>96</td>
<td>44.77</td>
</tr>
<tr>
<td>Test-based retention</td>
<td>100.62</td>
<td>33</td>
<td>43.60</td>
</tr>
<tr>
<td>Literal inference</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized testing only</td>
<td>3.65</td>
<td>102</td>
<td>3.36</td>
</tr>
<tr>
<td>Test-based retention</td>
<td>5.15</td>
<td>39</td>
<td>3.68</td>
</tr>
<tr>
<td>Extrapolation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized testing only</td>
<td>0.80</td>
<td>82</td>
<td>0.57</td>
</tr>
<tr>
<td>Test-based retention</td>
<td>0.93</td>
<td>21</td>
<td>0.67</td>
</tr>
<tr>
<td>Evaluation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized testing only</td>
<td>0.15</td>
<td>20</td>
<td>0.14</td>
</tr>
<tr>
<td>Test-based retention</td>
<td>0.16</td>
<td>6</td>
<td>0.11</td>
</tr>
<tr>
<td><strong>Math Cognitive Skills</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lowest Math Proficiency</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Negative Expectation</td>
<td>Indifferent Expectation</td>
<td>Positive Expectation</td>
</tr>
<tr>
<td>------------------</td>
<td>----------------------</td>
<td>-------------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td></td>
<td>Odds  n*</td>
<td>Odds  n</td>
<td>Odds  n</td>
</tr>
<tr>
<td>Standardized testing only</td>
<td>-- 49</td>
<td>-- 36</td>
<td>-- 139</td>
</tr>
<tr>
<td>Test-based retention</td>
<td>-- 16</td>
<td>-- 19</td>
<td>-- 81</td>
</tr>
<tr>
<td>Addition/subtraction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized testing only</td>
<td>52.04 96</td>
<td>64.89 73</td>
<td>50.34 302</td>
</tr>
<tr>
<td>Test-based retention</td>
<td>119.33 36</td>
<td>41.11 32</td>
<td>43.02 136</td>
</tr>
<tr>
<td>Multiplication/division</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized testing only</td>
<td>3.95 129</td>
<td>3.11 110</td>
<td>4.26 389</td>
</tr>
<tr>
<td>Test-based retention</td>
<td>5.90 43</td>
<td>1.20 22</td>
<td>3.56 141</td>
</tr>
<tr>
<td>Place value</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized testing only</td>
<td>0.58 74</td>
<td>0.49 75</td>
<td>0.59 244</td>
</tr>
<tr>
<td>Test-based retention</td>
<td>0.55 23</td>
<td>0.62 29</td>
<td>0.70 84</td>
</tr>
<tr>
<td>Rate and measurement</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized testing only</td>
<td>0.08 19</td>
<td>0.05 11</td>
<td>0.07 41</td>
</tr>
<tr>
<td>Test-based retention</td>
<td>0.02 6</td>
<td>0.04 3</td>
<td>0.10 18</td>
</tr>
</tbody>
</table>

Note: The lowest reading proficiency level includes associating letter with sounds at the beginning of end of words and recognizing common words by sight. The lowest math proficiency level includes identifying number and shape, recognizing relative size and identifying ordinality, sequence, etc.

*a The frequency n refers to the number of students whose highest proficiency level fell into the corresponding skill category within the treatment condition.

**Effects of the grade-3 test-based retention policy.** As an omnibus likelihood ratio test is not feasible for the multilevel logistic regression, to understand the effects of the grade-3 test-based retention policy, I first investigated the conditional policy effects by constraining $y_{0h1}$ as equal in the three-level binary logistic model (see Equation 7.5-7.7). After the differences between the expectations were controlled for, I did not observe any significant effects of the test-based retention policy on the math/reading cognitive skills.

Next, I evaluated the policy effects by teacher expectations. As shown in Table 7.3, the test-based retention policy produced a unique learning pattern under each expectation.
Under the negative expectation, the test-based retention policy appeared to impede the learning of advanced skills, but promote the learning of lower-level skills. This pattern seemed to be more pronounced in math than in reading and science. In math, if implemented by negative-expectation teachers, the policy would significantly lower the likelihood for students to learn rate and measurement (coefficient = -1.18, odds ratio = .52, \( t = -2.26, p < .05 \)); however, it would lead to insignificant, but notable improvement in acquiring subtraction/addition (coefficient = .83, odds ratio = 2.29, \( SE = .55, t = 1.52, p = .13 \)). In reading, the policy seemed to produce a considerable albeit insignificant improvement in the words comprehension skill (coefficient = .93, odds ratio = 2.54, \( SE = .53, t = 1.77, p = .08 \)).

Under the indifferent expectation, students’ learning of the reading and math skills were mostly unresponsive to the test-based retention policy except for the learning of multiplication/division in math; this grade 3-level math skill was less likely to be gained in the high-stakes context than in the standardized testing only condition (coefficient = -.95, odds ratio = .39, \( SE = .32, t = -2.95, p < .01 \)).

Under the positive expectation, the test-based retention policy negatively affected the learning of a low-level reading skill, i.e. comprehension of words in context (coefficient = -.89, odds ratio = .41, \( t = -2.58, SE = .35, p < .05 \)); but it significantly increased the probability of acquiring an advanced math skill, i.e. rate and measurement (coefficient = .43, odds ratio = 1.54, \( SE = .21, t = 2.06, p < .05 \)).
Table 7.3
Weighted Analysis of the Policy-by-Teacher Effects on Student Mastery of Each Cognitive Skill

<table>
<thead>
<tr>
<th>Cognitive Skill (from lowest to highest)</th>
<th>Negative Expectation</th>
<th>Indifferent Expectation</th>
<th>Positive Expectation</th>
<th>Multivariate Hypothesis Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff.</td>
<td>SE</td>
<td>Odds Ratio</td>
<td>Coeff.</td>
</tr>
<tr>
<td>Comprehension of words in context</td>
<td>0.93</td>
<td>0.53</td>
<td>2.54</td>
<td>-0.04</td>
</tr>
<tr>
<td>Literal inference</td>
<td>0.33</td>
<td>0.3</td>
<td>1.39</td>
<td>0.03</td>
</tr>
<tr>
<td>Extrapolation</td>
<td>0.10</td>
<td>0.23</td>
<td>1.10</td>
<td>0.18</td>
</tr>
<tr>
<td>Evaluation</td>
<td>0.07</td>
<td>0.42</td>
<td>1.08</td>
<td>-0.23</td>
</tr>
<tr>
<td>Reading Cognitive Skills</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Addition/ subtraction</td>
<td>0.83</td>
<td>0.55</td>
<td>2.29</td>
<td>-0.46</td>
</tr>
<tr>
<td>Multiplication/ division</td>
<td>0.40</td>
<td>0.32</td>
<td>1.49</td>
<td>-0.95**</td>
</tr>
<tr>
<td>Place value</td>
<td>-0.04</td>
<td>0.23</td>
<td>0.96</td>
<td>0.23</td>
</tr>
<tr>
<td>Rate and measurement</td>
<td>-1.18*</td>
<td>0.52</td>
<td>0.31</td>
<td>-0.21</td>
</tr>
</tbody>
</table>

Note: * p<.05. ** p<.01. *** p<.001.
**Moderating role of teacher expectations.** I first conducted multivariate hypothesis testing of $\gamma_{011} = \gamma_{021} = \gamma_{031} = 0$ to examine whether teacher expectations modified the policy effects. If a hypothesis test suggested significant differentiation between teacher expectations (see the results in the last column of Table 7.3), I subsequently employed post-hoc analyses to further explore the differences.

Specifically, in reading, the hypothesis tests signaled significant differences in terms of the policy effects on students’ learning of comprehension of words in context ($\chi^2 = 10.00, df = 3, p < .05$). The post-hoc analyses revealed that the test-based retention policy decreased the acquisition rate of words comprehension among students encountering positive-expectation teachers, significantly different from the increase among their peers encountering negative expectation teachers (contrast = 1.83, $SE = .63, \chi^2 = 8.52, p < .01$).

In math, significant differences in the policy effects across the expectations were found for students’ mastery of multiplication/division ($\chi^2 = 13.35, df = 3, p < .01$) as well as for the mastery of rate and measurement ($\chi^2 = 8.83, df = 3, p < .05$). The post-hoc analyses revealed that the negative policy effect on the learning of multiplication/division under the indifferent expectation was significantly different from the effect under the negative expectation (contrast = -1.35, $SE = .46, \chi^2 = 8.65, p < .01$); the positive effect on the learning of rate and measurement under the positive expectation was also significantly different from the effect under the negative expectation (contrast = 1.61, $SE = .59, \chi^2 = 7.40, p < .01$).

**Across-skills Comparison of the Policy-by-Teacher Effects**

How did the policy-by-teacher effects differ among skills? To answer this question, I ran a weighted two-level multinomial logistic model for each subject, applying student- and school-
level marginal mean weights. At level 1, $Y_{ijk}$ refers to the highest proficiency level that student $i$ of teacher $j$ from school $k$ has achieved (see Table 7.2 for students distribution across the skills). It takes a value of 1 to 5, with 1 denoting the lowest level and 5 the highest.

Level 1

$$\eta_{mjk} = \log \left( \frac{\phi_{mjk}}{\phi_{Mjk}} \right) = \log \left( \frac{Pr(Y_{ijk} = m)}{Pr(Y_{ijk} = 5)} \right) = \pi_{0,jk(m)} + \sum_{l=1}^{n_v} \pi_{ijkl(m)} X_{ijk} , \ m = 1, \ldots, 4 \quad (7.8)$$

Due to a small school-level variance, especially for the reading outcome, I combined school and class levels as level 2.

Level 2

$$\pi_{0,jk(m)} = T1_{jk} \star (\beta_{01(m)} + \beta_{02(m)} Z_k + \sum_{g=3}^{n_v+2} \beta_{0g(m)} V_k )$$

$$+ T2_{jk} \star (\beta_{0(nv+3)(m)} + \beta_{0(nv+4)(m)} Z_k + \sum_{g=nv+5}^{2nv+4} \beta_{0g(m)} V_k)$$

$$+ T3_{jk} \star (\beta_{0(2nv+5)(m)} + \beta_{0(2nv+6)(m)} Z_k + \sum_{g=2nv+7}^{3nv+6} \beta_{0g(m)} V_k)$$

$$+ \sum_{p=3nv+7}^{3nv+nv+6} \beta_{0p(m)} W_{jk} + r_{0,jk(m)},$$

$$r_{0,jk(m)} \sim N(0, \tau_\pi) \quad (7.9)$$

Figure 7.3 compares the estimated policy effects within each teacher expectation as log odds of achieving each skill in reference to the highest-level skill in reading (i.e. evaluation) or math (i.e. rate and measurement). It shows that the differential policy effects across the skills were more pronounced in math than in reading; the changes in the skills focus under the test-based retention policy depended on teacher expectations.

Under the negative expectation, the negative policy effect on the learning of the advanced rate and measurement skill was significantly different from the effects on grade 3-level skills, including multiplication/division (coefficient = -.97, odds ratio = 2.64, $SE= .46$, $t = -2.10, p < .05$) and place value (coefficient = -.93, odds ratio = 2.54, $SE= .42$, $t = -2.22, p < .05$).
Under the positive expectation, in reading, the test-based retention policy failed more students into the lowest reading proficiency category (see Table 7.3), which was significantly different from the policy effects on the odds of acquiring the extrapolation skill (contrast = -.86, \( SE = .44, \chi^2 = 3.85, p < .05 \)); in math, the positive policy effect on the mastery of rate and measurement was significantly different from that on the learning of multiplication/division (coefficient = .53, odds ratio = .59, \( SE = .24, t = 2.23, p < .05 \)).

Under the indifferent expectation, the negative policy effect on achieving multiplication/division was significantly different from the effect on the learning of addition/subtraction (contrast = -1.32, \( SE = .59, \chi^2 = 5.05, p < .05 \)).
Figure 7.3. Across-skills comparison of the estimated policy effects by teacher expectations on student mastery of cognitive skills.

Note: * $p<.05$. ** $p<.01$. *** $p<.001$. 
Student Differential Academic Performance by Prior Ability

At this stage of analysis, I break down the estimated overall policy effects by student prior academic ability.

Student Differential Academic Performance

First I asked: How did the policy and teacher expectations affect the academic performance of students at each ability level? The statistical model used here was similar to the univariate weighted model for analyzing the overall policy effects (i.e. Equations 7.1-7.3) except that at level 1 I added in five dummy indicators $A1$-$A5$ for student prior academic ability with $A1$ standing for the lowest ability and $A5$ for the highest; and the student- and school-level marginal mean weights were applied at levels 1 and 3 respectively.

Level 1:

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk} (A1) + \pi_{2jk} (A2) + \pi_{3jk} (A3) + \pi_{4jk} (A4) + \pi_{5jk} (A5)$$

$$+ \sum_{i=0}^{n+5} \pi_{ijk} X_{ijk} + e_{ijk}, \quad e_{ijk} \sim N(0, \sigma^2) \quad (7.10)$$

Level 2:

$$\pi_{0jk} = \beta_{00k} + r_{0jk}, \quad r_{0jk} \sim N(0, \tau),$$

$$\pi_{hk} = \beta_{h1k} (T1)_{jk} + \beta_{h2k} (T2)_{jk} + \beta_{h3k} (T3)_{jk} + \sum_{p=4}^{n+3} \beta_{hpk} W_{jk}, \quad h = 1-5$$

$$\pi_{ijk} = \beta_{10k} \quad (7.11)$$

Level 3:

$$\beta_{00k} = u_{00k}, \quad u_k \sim N(0, \tau)$$

$$\beta_{hqk} = \gamma_{h0} + \gamma_{hl1} Z_k + \sum_{g=2}^{m+1} \gamma_{hgg} V_k, \quad \text{for } q = 1-3$$

$$\beta_{hpk} = \gamma_{hp0} \quad (7.12)$$

$$\beta_{10k} = \gamma_{100}$$

Since only about 54% of the classes had two or more students selected into the final analytic sample, I assumed a constant within-class variance across the five ability levels and a constant
within-school variance across the three expectations. This was done by including a pseudo intercept at level 1 and then partitioning it into class and school random effects at levels 2 and 3.

Below I evaluate by sequence the effects of teacher expectations, the effects of the grade-3 test-based retention policy, and the moderating role of teacher expectations.

**Effects of teacher expectations.** Table 7.4 displays the estimated average academic performance in relative to the six treatment conditions for each ability subpopulation. In the standardized testing only condition, teacher expectations did not play a significant role in the math performance of any ability group, but affected the reading and science performance of the students in the middle of the ability distribution. For this group of students, pairwise comparisons indicated a benefit of the negative expectation in reading over the indifferent expectation (contrast = 2.56, $SE = .92$, $\chi^2 = 7.78$, $p < .01$) and over the positive expectation (contrast = 1.82, $SE = .70$, $\chi^2 = 6.76$, $p < .01$); the benefit of the negative expectation were also evident in science if compared to the positive expectation (contrast = 4.47, $SE = 1.23$, $\chi^2 = 13.22$, $p < .001$).

In the test-based retention condition, reading performance of each ability group was not affected by the teacher expectations. However in math and science, the positive expectation appeared to be beneficial to the ability 2 and 3 students in the high-stakes context. In math, it would work better for the ability 2 students than the indifferent expectation (contrast = 4.56, $SE = 1.17$, $\chi^2 = 15.13$, $p < .001$) and would as well work better for the ability 3 students than the negative expectation (contrast = 3.89, $SE = 1.40$, $\chi^2 = 7.69$, $p < .01$). In science, the positive expectation also exhibited an advantage over the indifferent expectation for the middle-ability students (i.e. ability 3, contrast = 3.78, $SE = 1.48$, $\chi^2 = 6.54$, $p = .01$); however, for the bottom-ability students (i.e. ability 1), the positive expectation did not work as well as the indifferent
expectation (contrast = -7.66, $SE = 2.16, \chi^2 = 12.58, p < .001$).

Table 7.4

Estimated Mean (Standard Error) of Student Academic Performance by Prior Ability and Treatment Conditions

<table>
<thead>
<tr>
<th>Ability</th>
<th>Reading</th>
<th>Math</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Negative Expectation</td>
<td>Indifferent expectation</td>
</tr>
<tr>
<td>Ability 1</td>
<td>Standardized testing only</td>
<td>44.58 (0.80)</td>
</tr>
<tr>
<td>Ability 1</td>
<td>Test-based retention</td>
<td>44.68 (0.94)</td>
</tr>
<tr>
<td>Ability 2</td>
<td>Standardized testing only</td>
<td>48.41 (0.70)</td>
</tr>
<tr>
<td>Ability 2</td>
<td>Test-based retention</td>
<td>49.57 (0.86)</td>
</tr>
<tr>
<td>Ability 3</td>
<td>Standardized testing only</td>
<td>52.29 (0.58)</td>
</tr>
<tr>
<td>Ability 3</td>
<td>Test-based retention</td>
<td>48.64 (0.98)</td>
</tr>
<tr>
<td>Ability 4</td>
<td>Standardized testing only</td>
<td>53.25 (0.70)</td>
</tr>
<tr>
<td>Ability 4</td>
<td>Test-based retention</td>
<td>54.53 (1.25)</td>
</tr>
<tr>
<td>Ability 5</td>
<td>Standardized testing only</td>
<td>55.84 (0.73)</td>
</tr>
<tr>
<td>Ability 5</td>
<td>Test-based retention</td>
<td>54.34 (1.18)</td>
</tr>
<tr>
<td>Total</td>
<td>Standardized testing only</td>
<td>50.00 (0.41)</td>
</tr>
<tr>
<td>Total</td>
<td>Test-based retention</td>
<td>48.95 (0.58)</td>
</tr>
<tr>
<td></td>
<td>Negative Expectation</td>
<td>Indifferent expectation</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>---------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Standardized testing only</td>
<td>51.14 (0.75)</td>
<td>48.73 (1.00)</td>
</tr>
<tr>
<td>Test-based retention</td>
<td>49.31 (0.81)</td>
<td>49.30 (1.40)</td>
</tr>
<tr>
<td><strong>Ability 4</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized testing only</td>
<td>52.35 (1.18)</td>
<td>51.60 (1.17)</td>
</tr>
<tr>
<td>Test-based retention</td>
<td>53.47 (1.26)</td>
<td>51.10 (1.06)</td>
</tr>
<tr>
<td><strong>Ability 5</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized testing only</td>
<td>51.91 (1.50)</td>
<td>53.52 (0.96)</td>
</tr>
<tr>
<td>Test-based retention</td>
<td>52.81 (2.31)</td>
<td>53.59 (1.11)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized testing only</td>
<td>49.82 (0.48)</td>
<td>49.04 (0.46)</td>
</tr>
<tr>
<td>Test-based retention</td>
<td>49.71 (0.46)</td>
<td>48.40 (0.83)</td>
</tr>
<tr>
<td><strong>Science</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Ability 1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized testing only</td>
<td>46.94 (0.95)</td>
<td>46.42 (0.68)</td>
</tr>
<tr>
<td>Test-based retention</td>
<td>46.15 (1.47)</td>
<td>51.78 (1.94)</td>
</tr>
<tr>
<td><strong>Ability 2</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized testing only</td>
<td>48.51 (0.61)</td>
<td>47.27 (0.77)</td>
</tr>
<tr>
<td>Test-based retention</td>
<td>47.46 (1.13)</td>
<td>45.70 (1.11)</td>
</tr>
<tr>
<td><strong>Ability 3</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized testing only</td>
<td>50.84 (0.66)</td>
<td>46.37 (1.09)</td>
</tr>
<tr>
<td>Test-based retention</td>
<td>48.68 (0.85)</td>
<td>47.22 (1.33)</td>
</tr>
<tr>
<td><strong>Ability 4</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized testing only</td>
<td>50.78 (0.64)</td>
<td>49.58 (0.82)</td>
</tr>
<tr>
<td>Test-based retention</td>
<td>51.80 (1.17)</td>
<td>50.34 (0.91)</td>
</tr>
<tr>
<td><strong>Ability 5</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized testing only</td>
<td>53.58 (1.35)</td>
<td>54.49 (0.95)</td>
</tr>
<tr>
<td>Test-based retention</td>
<td>50.78 (1.57)</td>
<td>53.47 (1.53)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized testing only</td>
<td>49.28 (0.38)</td>
<td>48.23 (0.44)</td>
</tr>
<tr>
<td>Test-based retention</td>
<td>48.11 (0.53)</td>
<td>48.96 (0.71)</td>
</tr>
</tbody>
</table>
Effects of the grade-3 test-based retention policy. As a prelude to my exploration of the policy effects, for each ability subpopulation I conducted omnibus likelihood ratio tests by comparing the models with or without the corresponding policy effects (i.e. $\gamma_{hq}$). There were significant policy effects on the ability 1 students in reading ($\chi^2 = 11.62$, $df = 3$, $p < .01$) and science ($\chi^2 = 22.17$, $df = 3$, $p < .001$), on the ability 2 students in math ($\chi^2 = 12.07$, $df = 3$, $p < .01$), and on the ability 3 students in reading ($\chi^2 = 10.35$, $df = 3$, $p < .05$) and math ($\chi^2 = 15.00$, $df = 3$, $p < .01$).

I examined the conditional policy effects given teacher expectations. After constraining teacher expectations as equal for each ability subpopulation (i.e. $\beta_{h1k} = \beta_{h2k} = \beta_{h3k}$), I identified significant conditional policy effects only in math and only for ability 2 (coefficient = 1.70, $SE = .71$, $t = 2.38$, $p < .05$) and ability 3 (coefficient = 1.77, $SE = .76$, $t = 2.32$, $p < .05$) students.

Subsequently I investigated how the test-based retention policy affected each ability subpopulation under different teacher expectations. To judge the significance of the estimated policy effects as shown in Table 7.5, I not only examined the $p$ value, but also considered the effects sizes using Cohen’s $d$ equal or larger than .20 as a criterion.

**Table 7.5**

*Estimated Effects of the Test-based Retention Policy on Student Academic Performance by Student Prior Ability and Teacher Expectations*

<table>
<thead>
<tr>
<th>Ability 1</th>
<th>Reading</th>
<th>Math</th>
<th>Science</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff.</td>
<td>SE</td>
<td>Coeff.</td>
</tr>
<tr>
<td>Negative expectation</td>
<td>0.11</td>
<td>1.17</td>
<td>-0.30</td>
</tr>
<tr>
<td>Indifferent expectation</td>
<td>1.69</td>
<td>1.99</td>
<td>-1.09</td>
</tr>
<tr>
<td>Positive expectation</td>
<td>-2.23</td>
<td>1.39</td>
<td>-1.55</td>
</tr>
<tr>
<td></td>
<td>Reading</td>
<td></td>
<td>Math</td>
</tr>
<tr>
<td>--------------------</td>
<td>---------</td>
<td>----------</td>
<td>------</td>
</tr>
<tr>
<td></td>
<td>Coeff.</td>
<td>SE</td>
<td>Coeff.</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\chi^2_{(3)} = 3.82$</td>
<td></td>
<td>$\chi^2_{(3)} = 2.68$</td>
</tr>
<tr>
<td>Ability 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative expectation</td>
<td>1.16</td>
<td>1.14</td>
<td>1.49</td>
</tr>
<tr>
<td>Indifferent expectation</td>
<td>-1.12</td>
<td>1.10</td>
<td>-2.01</td>
</tr>
<tr>
<td>Positive expectation</td>
<td>0.75</td>
<td>0.81</td>
<td>2.72^{**}</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\chi^2_{(3)} = 3.09$</td>
<td></td>
<td>$\chi^2_{(3)} = 14.81^{**}$</td>
</tr>
<tr>
<td>Ability 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative expectation</td>
<td>-3.65^{**}</td>
<td>1.12</td>
<td>-1.83</td>
</tr>
<tr>
<td>Indifferent expectation</td>
<td>1.76</td>
<td>1.25</td>
<td>0.57</td>
</tr>
<tr>
<td>Positive expectation</td>
<td>1.09</td>
<td>0.88</td>
<td>3.41^{**}</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\chi^2_{(3)} = 13.07^{**}$</td>
<td></td>
<td>$\chi^2_{(3)} = 11.32^{*}$</td>
</tr>
<tr>
<td>Ability 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative expectation</td>
<td>1.28</td>
<td>1.34</td>
<td>1.12</td>
</tr>
<tr>
<td>Indifferent expectation</td>
<td>1.61</td>
<td>0.99</td>
<td>-0.05</td>
</tr>
<tr>
<td>Positive expectation</td>
<td>-0.40</td>
<td>0.66</td>
<td>0.65</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\chi^2_{(3)} = 4.73$</td>
<td></td>
<td>$\chi^2_{(3)} = 1.31$</td>
</tr>
<tr>
<td>Ability 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative expectation</td>
<td>-1.51</td>
<td>1.23</td>
<td>0.90</td>
</tr>
<tr>
<td>Indifferent expectation</td>
<td>1.41</td>
<td>2.60</td>
<td>0.06</td>
</tr>
<tr>
<td>Positive expectation</td>
<td>0.50</td>
<td>0.96</td>
<td>0.78</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\chi^2_{(3)} = 2.27$</td>
<td></td>
<td>$\chi^2_{(3)} = 0.8$</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative expectation</td>
<td>-1.06</td>
<td>0.70</td>
<td>-0.11</td>
</tr>
<tr>
<td>Indifferent expectation</td>
<td>0.99</td>
<td>0.70</td>
<td>-0.64</td>
</tr>
<tr>
<td>Positive expectation</td>
<td>-0.98</td>
<td>0.63</td>
<td>0.41</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\chi^2_{(3)} = 8.53^{*}$</td>
<td></td>
<td>$\chi^2_{(3)} = 1.60$</td>
</tr>
</tbody>
</table>

Note: * $p<.05$. ** $p<.01$. *** $p<.001$.

Under the negative expectation, the policy was found detrimental to the students in the middle-ability group and the highest-ability group. As a result of the test-based retention policy,
the middle-ability students experienced a dramatic drop in the learning of reading (coefficient = -3.65, \(SE = 1.12, t = -3.28, p < .01\)) and in that of science (coefficient = -2.16, \(SE = 1.06, t = -2.05, p < .05\)). The magnitude of the learning losses approximated 37% and 22% of the population standard deviation in reading and science respectively. The highest-ability students (i.e. ability 5), if taught by negative-expectation teachers also suffered from a negative consequence of the high-stakes testing in their science learning, with an effect size equal to 28% of the population standard deviation.

Under the indifferent expectation, for students at the bottom ability level, the high-stakes testing radically raised their science performance by 5.36 points (\(SE = 2.06, t = 2.60, p < .01\)), which was more than a half of the population standard deviation. In addition, the policy led to a slight setback in the math learning of the ability 2 students (coefficient = 2.01, \(SE = 1.16, t = -1.73, p = .08\)); the effect size was equivalent to 20% of the math population standard deviation.

Under the positive expectation, the test-based retention policy appeared to be ineffective to the students on the higher end of the ability distribution (i.e. ability 4 & 5); it improved the learning of ability 2 and 3 students, but impeded the learning of the bottom-ability students in certain subject areas. The policy was found beneficial to the middle-ability students in math (coefficient = 3.41, \(SE = 1.09, t = 3.13, p < .01\)) and in science (coefficient = 1.92, \(SE = .82, t = 2.35, p < .05\)). The effect sizes were approximately one third of the math population standard deviation and one fifth of the science standard deviation respectively. The policy also significantly improved the math learning of the ability 2 students (coefficient = 2.72, \(SE = .86, t = 3.18, p < .01\)), with an effect size of more than one quarter of the population standard deviation. However, if taught by positive-expectation teachers, the lowest-ability students suffered from the high-stakes policy, especially in terms of their science learning (coefficient =
-2.48, $SE = .99, t = -2.51, p < .05$) as well as of the reading learning (coefficient = -2.23, $SE = 1.39, t = -1.61, p = .11$). Their learning loss in science and reading was about 25% and 22% of the corresponding population standard deviation.

**Moderating role of teacher expectations.** According to multivariate hypothesis testing of $\gamma_{b11} = \gamma_{b21} = \gamma_{b31} = 0$, teacher expectations moderated the policy effects only on average or lower-than-average ability subpopulations.

As shown in the rows of Hypothesis testing in Table 7.5, for students in the middle ability group, in every subject, the policy would work better under the positive expectation than under the negative expectation: the discrepancy in the policy effects between the two expectations was as much as 4.74 points ($SE = 1.53, \chi^2 = 9.61, p < .01$) in reading, 5.25 points ($SE = 1.69, \chi^2 = 9.59, p < .01$) in math, and 4.08 points ($SE = 1.23, \chi^2 = 10.98, p < .01$) in science; moreover, also for the middle-ability students, the policy effect would be less detrimental in reading if operated by the indifferent expectation than by the negative expectation (contrast = 5.41, $SE = 1.69, \chi^2 = 10.30, p < .01$). For students in the second ability group, in math, the positive expectation outscored the indifferent expectation in terms of the policy effects by 4.73 points ($SE = 1.38, \chi^2 = 11.80, p < .001$). For students at the bottom ability level, the indifferent expectation excelled the negative expectation (contrast = 6.15, $SE = 2.50, \chi^2 = 6.06, p < .05$) and the positive expectation (contrast = 7.83, $SE = 2.27, \chi^2 = 11.89, p < .001$) in the policy effects on the science performance.

**Across-ability levels Comparison of the Policy-by-Teacher Effects**

Did the policy-by-teacher treatments produce differential effects pattern across different ability subpopulations? To avoid inflated type I error, I first used multivariate hypothesis testing
to examine the differences in the policy effects across the five ability groups for each expectation level. I found that uneven distribution of the policy effects would occur in math ($\chi^2 = 21.62, df = 5, p < .0001$) and science ($\chi^2 = 19.84, df = 5, p < .01$) if teachers were positive about students’ average learning ability and would as well occur in reading ($\chi^2 = 15.53, df = 5, p < .01$) if teachers did not believe most of the students could learn.

As illustrated in Figure 7.4, if operated by the negative expectation, the test-based retention policy would hurt the reading performance of the middle-ability students more than that of the ability 1 (contrast = -3.76, $SE = 1.59, \chi^2 = 5.59, p < .05$), ability 2 (contrast = -4.81, $SE = 1.56, \chi^2 = 9.56, p < .01$), and ability 4 groups (contrast = -4.93, $SE = 1.60, \chi^2 = 9.50, p < .01$).

In contrast, the combination of the test-based retention policy and positive expectation seemed to prioritize the ability 2 and 3 students at the expense of other students, especially of the lowest-ability students. If taught by positive-expectation teachers, the middle-ability students would gain the most from the test-based retention policy. The learning gain of the middle-ability students would be significantly different from the learning loss of the bottom-ability students in math (contrast = 4.97, $SE = 1.57$, $\chi^2 = 10.00, p < .01$) and science (contrast = 4.39, $SE = 1.29$, $\chi^2 = 11.63, p < .01$). The gain would as well be different from the learning rate of the ability 4 students in math (contrast = 2.76, $SE = 1.26$, $\chi^2 = 4.84, p < .05$) and science (contrast = 2.93, $SE = 1.06$, $\chi^2 = 7.70, p < .01$) and from the learning rate of the ability 5 students in science (contrast = 3.31, $SE = 1.26$, $\chi^2 = 6.96, p < .01$). Under the positive expectation, there was also a significant contrast between the ability 2 and the bottom-ability students in terms of the policy effects on their math (contrast = 4.27, $SE = 1.33$, $\chi^2 = 10.32, p < .01$) or science learning (contrast = 2.83, $SE = 1.38$, $\chi^2 = 4.22, p < .05$).
Figure 7.4. Across-ability levels comparison of the estimated policy-by-teacher effects on student academic performance.
Student Long-term Academic Performance

The last part of the Phase III analysis utilized IRT $\theta$ scores (see Table 4.1 for the descriptive statistics) to explore whether the observed policy effects or no effects in grade 3 would be sustained over time. It has to be pointed out that the analysis essentially provides a cross-sectional perspective by looking at only two time points and having no control for any events in between. Therefore, the results should be considered as suggestive and be evaluated with cautions. To avoid over-interpretation of the causal relationship between the grade-3 test-based retention policy and student long-term academic outcomes, instead of presenting point estimates of long-term policy effects, my description of the results focuses on changes in the policy effects between spring 2002 and 2004 as indicated by multivariate hypothesis testing.

Long-term Overall Academic Performance

Were the policy effects on student overall academic performance persistent till two years later? I first examined the repeated measures of student overall academic performance by conducting separating analysis for reading, math, and science. The statistical model had the same structure as the three-level weighted model for the between-subjects comparison of student short-term academic performance (see Equations 7.4, 7.2, & 7.3). I found that the spring 2004 outcomes on average were unrelated to the test based retention policy regardless of teacher expectations. Although the policy reduced third-graders’ overall science performance under the positive expectation in spring 2002, this group of students bounced back with a learning gain of .08 in science ($SE = .03, \chi^2 = 9.55, p < .01$) over the two years. Other than this, I did not observe other significant long-term change in the policy effects under any expectation in any subjects.

Long-term Differential Academic Performance

Then were the policy effects on student differential performance by prior ability
persistent till two years later? I analyzed a three-level weighted multivariate model, applying the student- and school-level marginal mean weights at level 1 and 3 respectively. At level 1, I used dummy indicators to combine two sub-models (same as Equation 7.10) for evaluating differential policy effects by prior ability in spring 2002 and 2004, and assumed homogenous variance between years. Within each sub-model, the structure of levels 2 and 3 was the same as in Equations 7.11 and 7.12.

In reading and math, most of the effects and non-effects by student prior ability that I observed in spring 2002 appeared to last over time, but with two exceptions under the indifferent expectation. Among those who had experience with the test-based retention policy and indifferent expectation in grade 3, the bottom-ability students, though unaffected by the policy in the promotional gate year, showed significant improvement in reading performance in spring 2004 (contrast = .09, $SE$ = .04, $\chi^2 = 5.21, p < .05$); and the students in the ability group 2 also gained .10 in math after two years ($SE$ = .04, $\chi^2 = 5.71, p < .05$).

In science, for the middle-ability students who received the positive expectation in grade 3, the short-term positive policy effects seemed to get stronger after two years (contrast = -.15, $SE$ = .06, $\chi^2 = 6.52, p < .05$). Nonetheless, for the bottom-ability students, the significant short-term policy effects were no longer visible: the benefit of the policy effects for those who were treated by the indifferent expectation in grade 3 disappeared by the end of spring 2004, signaled by a loss of .23 ($SE$ = .10, $\chi^2 = 5.06, p < .05$); in the meantime, those who were treated by positive-expectation teachers in grade 3, though suffering from the policy in a short run, caught up with their peers in the control condition by increasing .15 ($SE$ = .05, $\chi^2 = 9.86, p < .01$) during the years.
Summary

In this chapter, I evaluated the effects of the grade-3 test-based retention policy and teacher expectations on different aspects of student academic performance. Below I summarize the findings in two segments: (a) effects of the grade-3 test-based retention and (b) effects of teacher expectations.

Effects of the Grade-3 Test-based Retention Policy

I organize the findings about the effects of the grade-3 test-based retention policy by the four aspects of student academic performance; in each subsection, I first revisit the relevant hypotheses imposed in Chapter 2 and then highlight in bullet form the key findings from the current chapter.

Student overall academic performance. For students under the positive or negative expectation, I hypothesized that the test-based retention policy would not significantly affect their overall learning in tested subjects; but would hurt the learning in non-tested subjects.

• Conditional policy effects: Controlling for teacher expectations, I found that the policy had no influence on student overall math or reading performance, but negatively affected student science performance in general.

• Policy effects by teacher expectations: Under the positive expectation, in conformity with the hypothesis, the policy was ineffective to students’ overall learning in reading and math, but detrimental to the learning in science. Under the negative expectation, the policy did not affect students’ math learning either, but slightly reduced the reading and science learning. Under the indifferent expectation, the policy did not change students’ overall learning in every subject including science.

• Moderating role of teacher expectations: I discovered that teacher expectations played a
significant role in modifying the relationship between the policy and student academic performance in reading and science. Compared to the other expectations, the indifferent expectation could shelter students from the negative policy effects on students’ overall reading and science learning.

- **Between-subject differences in policy effects**: Between-subjects comparisons revealed that as long as teachers held a clear expectation, either positive or negative, of their students’ learning ability, the test-based retention policy would produce uneven learning distribution across the subjects. Only under the indifferent expectation, would the policy produce equally insignificant learning across the three subjects.

**Student mastery of cognitive skills.** Even though the policy effects might be minimal on a composite measure of student learning in tested subjects, I predicted that the test-based retention policy might generate improved learning in the advanced skills with positive-expectation teachers; it might decrease learning in the same skills but increase learning in other lower-level skills with negative-expectation teachers; the effects pattern would be more prominent in math where skills learning is often viewed as hierarchical and sequential.

- **Conditional policy effects**: After taking the differences in teacher expectations into account, I did not find any significant conditional effect of the test-based retention policy on students’ mastery of each math/reading cognitive skill.

- **Policy effects by teacher expectations**: The observed effects pattern by teacher expectations was consistent with the hypothesis. If implemented by negative-expectation teachers, the policy would harm the learning of an advanced level skill, i.e. rate and measurement, but to a certain extent, would foster the learning of skills that are below grade 3 curriculum standard, such as addition/subtraction in math and words comprehension in reading. If implemented by
positive-expectation teachers, the test-based retention policy would benefit students’ learning
of the advanced rate and measure skill, but would fail more students into the lowest reading
proficiency category. If implemented by indifferent-expectation teachers, the high-stakes
policy would be detrimental to the learning of multiplication/division, a focal emphasis of the
grade-3 math curriculum.

- **Moderating role of teacher expectations:** The analysis of the interaction effects indicated that
  the test-based retention policy would be more favored under the negative expectation than
  under the positive expectation if students need to learn basic reading skills such as words
  comprehension. However, the policy would work better under the positive expectation than
  under the negative expectation for students to acquire the advanced rate and measurement
  skill in math. The policy would be less desirable under the indifferent expectation than under
  the negative expectation for learning multiplication/division, a grade-3 level math skill.

- **Across-skills differences in policy effects:** I confirmed that the policy effects were unequal
  across skills; and the differential pattern was more pronounced in math than in reading and
  depended on teacher expectations.

  *Student differential academic performance by prior ability.* I hypothesized that, with
  positive-expectation teachers, the test-based retention policy would have positive influence on
  the learning of the average-ability students, and probably would have negative influence on the
  learning of the lowest-ability students; with negative-expectation teachers, the policy would
  negatively affect the learning of the average-ability students, but would have zero influence on
  the lower-than-average-ability students.

- **Conditional policy effects:** Conditioning on teacher expectations, the test-based retention
  policy appeared to have no significant effects in most of the subjects and on most of the
ability groups except that it improved the math performance of ability 2 and 3 groups.

- **Policy effects by teacher expectations:** The observed effects pattern was consistent with the hypothesis. If operated with the negative expectation, the policy would reduce the reading and science learning of the middle-ability students and would even reduce the science learning of the highest-ability students. If combined with the positive expectation, the policy would benefit the middle-ability students, typically in math and science, and would even help the ability 2 students to improve their math learning; nevertheless, the policy appeared to hurt the bottom-ability students in reading and science. If implemented by the indifferent expectation, the policy would reduce the math learning of the ability 2 students, but would dramatically increase the science learning of the bottom-ability students.

- **Moderating role of teacher expectations:** I found that teacher expectations modified the policy effects only within average- or lower-than-average-ability subpopulation. For the middle-ability students, the test-based retention policy would work better under the positive expectation than under the negative expectation in every subject; the policy would also work better under the indifferent expectation than under the negative expectation in terms of the students’ reading learning. For the ability 2 students, the policy would be more desirable under the positive expectation than under the negative expectation for the math learning. For the bottom-ability students, the policy would be more helpful under the indifferent expectation than under the other expectations for their science learning.

- **Across-ability levels differences in policy effects:** The policy effects were found unevenly distributed across ability levels under positive or negative expectation. Under the negative expectation, the policy hurt the reading learning of the middle-ability students more than most of the other ability groups. Under the positive expectation, the policy appeared to
prioritize the ability 2 and 3 students at the sacrifice of other students, especially of the bottom-ability students in math and science.

**Student long-term academic performance.** I hypothesized that if students were taught by positive-expectation teachers in grade 3, in a long run, they would bounce back from short-term negative policy effects, if there were any. I also hypothesized that if students were taught by negative-expectation teachers, the test-based retention policy would have negative residual effects on student long-term learning.

- **Long-term overall academic performance**: The significant negative effects on the grade-3 overall science learning under the positive expectation disappeared two years after the promotional gate grade, which was congruent with my hypothesis.

- **Long-term differential academic performance**: Through examining the long-term academic performance by student prior ability, I found that under the negative expectation, the learning loss of the middle-ability students due to the grade 3 test-based retention policy were persistent over time. Under the positive expectation, the short-term positive policy effects for the ability 2 and 3 students sustained or even got stronger in a long run while the negative effects for the bottom-ability students were no longer visible after the two years. For students who were once taught by indifferent-expectation teachers, I discovered that the positive policy effect on the science learning of the bottom-ability group faded away by the end of spring 2004; among those who were once exposed to the indifferent expectation and the test-based retention policy, long-term learning gains were observed in reading for the bottom-ability students and in math for the ability 2 students.

**Effects of Teacher Expectations**

Here I summarize the findings about the effects of teacher expectations on (a) student
overall academic performance, (b) their mastery of cognitive skills, and (c) their differential
academic performance by prior ability. As the analysis of student long-term academic
performance was only exploratory in nature, I did not examine the effects of teacher expectations
over time.

**Student overall academic performance.** In either the standardized testing only or test-
based retention condition, the three types of expectations did not differ from one another in terms
of student overall academic performance.

**Student mastery of cognitive skills.** In the standardized testing only condition, the
positive expectation showed an advantage over the indifferent expectation for acquiring the
advanced extrapolation skill in reading. When the test-based incentive was imposed, the negative
expectation outperformed the other two expectations in promoting the learning of below-grade 3
level skills such as words comprehension in reading and addition/subtraction in math; the
positive expectation was more helpful than the negative expectation for learning advanced math
skills such as rate and measurement; compared to the negative or positive expectation, the
indifferent expectation was associated with a lower acquisition rate of multiplication/division, a
grade-3 level math skill.

**Student differential academic performance by prior ability.** Within each ability
subpopulation, situated in the low-stakes context, the negative expectation seemed to be optimal
to the middle-ability students in reading and science than the other expectations. Once the
retention consequence was attached to test results, reading performance of each ability group was
no longer affected by the teacher expectations; the positive expectation became a better choice
for the middle-ability students or even for the ability 2 students in learning math or science.
However, compared to the indifferent expectation, the positive expectation was found harmful to
the science learning of the bottom-ability group in the high-stakes context.

**Notes to Chapter 7**

1 I used the population standard deviation (i.e. 10) of reading, math, or science performance to compute the effect sizes.

2 Indifferent-expectation group also had a smaller sample size and a larger standard error than the positive-expectation group. However I consider that the policy effects under the indifferent expectation were ignorable as the effect size was smaller than 10% in each subject.

3 This study is interested in the consequences of having the threat of retention, but not the retention itself. Hence my analysis of the long-term academic performance followed all the third graders involved in my previous analysis over time, including those who have been retained in grade 3 due to the high-stakes testing policy.
CHAPTER 8
EFFECTS OF TEST-BASED RETENTION POLICY AND TEACHER EXPECTATIONS ON STUDENT SELF-PERCEIVED COMPETENCE AND INTERESTS

Chapter 8 presents the findings on student self-perceived competence and interests in tested subjects (i.e. reading and math). In parallel to my focal analysis of the effects of test-based retention policy and teacher expectations on student academic performance, I investigated three aspects of student self-perceived competence and interests: (a) student overall self-perception, (b) student differential self-perception by prior ability, and (c) student long-term self-perception. The statistical models used in this part of analysis were similar to those for analyzing student academic performance.

**Student Overall Self-perception**

**Student Overall Self-perception in Each Subject**

Did the grade-3 test-based retention policy and teacher expectations jointly affect student self-perceived competence and interests toward tested subjects in grade 3? To answer this question, I first analyzed a three-level weighted univariate model (See Equations 7.1-7.3). The last two rows of Table 8.1 displays the estimated average academic self-perception in reading and math in relative to the six treatment conditions.

**Effects of teacher expectations.** Adopting a Bonferroni adjusted alpha level, in either standardized testing only or test-based retention condition, I compared the differences between the three teacher expectations. As shown in Table 8.1, there was no significant main effects of teacher expectations on student overall self-perception in reading or math.
Table 8.1

Estimated Mean (Standard Error) of Student Academic Self-perception by Prior Ability and Treatment Conditions

<table>
<thead>
<tr>
<th>Ability</th>
<th>Reading Self-perception</th>
<th></th>
<th>Math Self-perception</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Negative Expectation</td>
<td>Indifferent expectation</td>
<td>Positive expectation</td>
<td>Negative Expectation</td>
</tr>
<tr>
<td>Ability 1</td>
<td>Standardized testing only</td>
<td>3.18 (0.08)</td>
<td>3.22 (0.10)</td>
<td>3.22 (0.04)</td>
</tr>
<tr>
<td></td>
<td>Test-based retention</td>
<td>3.04 (0.09)</td>
<td>3.03 (0.15)</td>
<td>3.29 (0.06)</td>
</tr>
<tr>
<td>Ability 2</td>
<td>Standardized testing only</td>
<td>3.29 (0.06)</td>
<td>3.40 (0.07)</td>
<td>3.20 (0.04)</td>
</tr>
<tr>
<td></td>
<td>Test-based retention</td>
<td>3.19 (0.08)</td>
<td>3.27 (0.11)</td>
<td>3.22 (0.10)</td>
</tr>
<tr>
<td>Ability 3</td>
<td>Standardized testing only</td>
<td>3.43 (0.06)</td>
<td>3.27 (0.08)</td>
<td>3.26 (0.04)</td>
</tr>
<tr>
<td></td>
<td>Test-based retention</td>
<td>3.37 (0.09)</td>
<td>3.35 (0.11)</td>
<td>3.28 (0.06)</td>
</tr>
<tr>
<td>Ability 4</td>
<td>Standardized testing only</td>
<td>3.24 (0.08)</td>
<td>3.17 (0.08)</td>
<td>3.30 (0.05)</td>
</tr>
<tr>
<td></td>
<td>Test-based retention</td>
<td>3.46 (0.13)</td>
<td>3.36 (0.11)</td>
<td>3.24 (0.07)</td>
</tr>
<tr>
<td>Ability 5</td>
<td>Standardized testing only</td>
<td>3.44 (0.10)</td>
<td>3.28 (0.13)</td>
<td>3.44 (0.04)</td>
</tr>
<tr>
<td></td>
<td>Test-based retention</td>
<td>3.52 (0.18)</td>
<td>3.11 (0.16)</td>
<td>3.36 (0.10)</td>
</tr>
<tr>
<td>Total</td>
<td>Standardized testing only</td>
<td>3.36 (0.04)</td>
<td>3.28 (0.04)</td>
<td>3.28 (0.02)</td>
</tr>
<tr>
<td></td>
<td>Test-based retention</td>
<td>3.30 (0.05)</td>
<td>3.21 (0.09)</td>
<td>3.28 (0.04)</td>
</tr>
</tbody>
</table>
Effects of the grade-3 test-based retention policy. Through comparing the models with or without the policy predictors, omnibus likelihood ratio tests did not reveal any significant effects of the grade-3 test-based retention policy in reading or math. Constraining the policy effects as equal across the three expectations, I found that the policy led to negligible decreases in students’ self-perception toward reading (coefficient = -.02, \( SE = .04, t = -.61, p > .5 \)) and math (coefficient = -.02, \( SE = .04, t = -.52, p > .5 \)). A closer look at student overall self-perception under each teacher expectation did not suggest any significant effects of the test-based retention policy either (see the bottom rows of Table 8.2).

Moderating role of teacher expectations. According to multivariate hypothesis testing, the policy effects on student overall self-perception did not differ by teacher expectations in reading (\( \chi^2 = 1.15, df = 3, p > .5 \)) and math (\( \chi^2 = 2.24, df = 3, p > .5 \)).

Table 8.2
Estimated Policy Effects on Student Academic Self-perception by Teacher Expectations

<table>
<thead>
<tr>
<th>Ability 1</th>
<th>Reading Self-perception</th>
<th>Math Self-perception</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff.</td>
<td>SE</td>
</tr>
<tr>
<td>Negative expectation</td>
<td>-0.14</td>
<td>0.10</td>
</tr>
<tr>
<td>Indifferent expectation</td>
<td>-0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>Positive expectation</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td>( \chi^2_{(3)} = 4.17 )</td>
<td>( \chi^2_{(3)} = 5.78 )</td>
</tr>
<tr>
<td>Ability 2</td>
<td>Reading Self-perception</td>
<td>Math Self-perception</td>
</tr>
<tr>
<td></td>
<td>Coeff.</td>
<td>SE</td>
</tr>
<tr>
<td>Negative expectation</td>
<td>-0.10</td>
<td>0.09</td>
</tr>
<tr>
<td>Indifferent expectation</td>
<td>-0.12</td>
<td>0.14</td>
</tr>
<tr>
<td>Positive expectation</td>
<td>0.02</td>
<td>0.11</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td>( \chi^2_{(3)} = 2.17 )</td>
<td>( \chi^2_{(3)} = 5.12 )</td>
</tr>
<tr>
<td>Ability 3</td>
<td>Reading Self-perception</td>
<td>Math Self-perception</td>
</tr>
<tr>
<td></td>
<td>Coeff.</td>
<td>SE</td>
</tr>
<tr>
<td>Negative expectation</td>
<td>-0.06</td>
<td>0.12</td>
</tr>
</tbody>
</table>
### Reading Self-perception vs. Math Self-perception

<table>
<thead>
<tr>
<th></th>
<th>Coeff.</th>
<th>SE</th>
<th>Coeff.</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indifferent expectation</td>
<td>0.08</td>
<td>0.14</td>
<td>0.16</td>
<td>0.17</td>
</tr>
<tr>
<td>Positive expectation</td>
<td>0.02</td>
<td>0.07</td>
<td>0.10</td>
<td>0.12</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td>(\chi^2_{(3)} = .59)</td>
<td>(\chi^2_{(3)} = 1.68)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Ability 4**

<table>
<thead>
<tr>
<th></th>
<th>Coeff.</th>
<th>SE</th>
<th>Coeff.</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative expectation</td>
<td>0.21</td>
<td>0.14</td>
<td>0.04</td>
<td>0.20</td>
</tr>
<tr>
<td>Indifferent expectation</td>
<td>0.19</td>
<td>0.14</td>
<td>0.35*</td>
<td>0.15</td>
</tr>
<tr>
<td>Positive expectation</td>
<td>-0.07</td>
<td>0.08</td>
<td>-0.26*</td>
<td>0.13</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td>(\chi^2_{(3)} = 5.53)</td>
<td>(\chi^2_{(3)} = 11.00*)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Ability 5**

<table>
<thead>
<tr>
<th></th>
<th>Coeff.</th>
<th>SE</th>
<th>Coeff.</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative expectation</td>
<td>0.08</td>
<td>0.17</td>
<td>0.12</td>
<td>0.18</td>
</tr>
<tr>
<td>Indifferent expectation</td>
<td>-0.17</td>
<td>0.23</td>
<td>-0.11</td>
<td>0.19</td>
</tr>
<tr>
<td>Positive expectation</td>
<td>-0.09</td>
<td>0.10</td>
<td>-0.02</td>
<td>0.12</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td>(\chi^2_{(3)} = 1.43)</td>
<td>(\chi^2_{(3)} = .89)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Total**

<table>
<thead>
<tr>
<th></th>
<th>Coeff.</th>
<th>SE</th>
<th>Coeff.</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative expectation</td>
<td>-0.05</td>
<td>0.06</td>
<td>-0.03</td>
<td>0.07</td>
</tr>
<tr>
<td>Indifferent expectation</td>
<td>-0.07</td>
<td>0.10</td>
<td>-0.14</td>
<td>0.10</td>
</tr>
<tr>
<td>Positive expectation</td>
<td>-0.00</td>
<td>0.04</td>
<td>0.01</td>
<td>0.04</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td>(\chi^2_{(3)} = 1.15)</td>
<td>(\chi^2_{(3)} = 2.24)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: * \(p<.05\). ** \(p<.01\). *** \(p<.001\).

**Between-subject Comparison of the Policy-by-Teacher Effects**

Then how did the policy-by-teacher effects differ between different subjects? I compared student overall self-perception between reading and math using a three-level multivariate weighted model with a similar level-1 structure as Equation 7.4. The results suggested that under each teacher expectation, the test-based retention policy led to equivalently trivial changes in student self-perceived competence and interests in the both subjects.
**Student Differential Self-perception by Prior Ability**

In view of the zero policy-by-teacher effects on student overall self-perception, I further looked into each ability group and examined whether the same conclusion could be made within each subpopulation.

**Student Differential Self-perception**

I first asked: how did the policy-by-teacher effects differ among students at different prior ability levels? Below I describe the results from a set of weighted univariate analysis (See Equations 7.10 -7.12), including the effects of teacher expectations, the effects of the grade-3 test-based retention policy, and the moderating role of teacher expectations.

**Effects of teacher expectations.** Using a Bonferroni corrected significance level of .017, I conducted pairwise comparisons between teacher expectations in either policy condition. As shown in Table 8.1, in the standardized testing only condition, teacher expectations affected the self-perceived reading competence and interests of the ability 2 group. These students would have a higher self-perception toward reading if assigned to the indifferent expectation rather than to the positive expectation (contrast = .19, $SE = .08$, $\chi^2 = 5.71, p = .016$).

In the test-based retention condition, compared to the negative expectation, the positive expectation was associated with a higher reading self-perception of the bottom-ability student (contrast = .24, $SE = .10$, $\chi^2 = 5.73, p = .016$); but compared to the indifferent expectation, the positive expectation was associated with a lower math self-perception of the ability 4 students (contrast = -.35, $SE = .15$, $\chi^2 = 5.66, p = .016$).

**Effects of the grade-3 test-based retention policy.** I first conducted omnibus likelihood ratio tests comparing models with or without the policy indicators for each ability subpopulation to roughly detect the existence of the policy effects. The results were significant only in math
and for ability 1 ($\chi^2 = 14.25, df = 3, p < .01$) and ability 4 students ($\chi^2 = 10.55, df = 3, p < .05$).

However, after constraining teacher expectations as equal for each ability group, I did not detect any conditional policy effects in any subject.

I then evaluated the differential effects of the test-based retention policy by teacher expectations (see Table 8.2). My evaluation was based not only on the $p$ value of the policy effects, but also on their effects sizes (i.e. to see whether Cohen’s $d$ equal or larger than .20).

Under the negative expectation, the policy appeared not to matter for students’ math self-perception within any ability group; however, it resulted in a moderate decrease in the reading self-perception of the bottom-ability students (coefficient = -.14, $SE = .10$, $t = -1.30$, $p = .19$, $d = .21$), but an increase in that of the ability 4 students (coefficient = .21, $SE = .14$, $t = 1.53$, $p = .13$, $d = .32$).

Under the indifferent expectation, the test-based retention policy appeared to benefit the academic self-perception of the ability 3 and 4 students but hurt that of the other students, depending on subject areas. In reading, the policy effects were not statistically significant. But there seemed to be a considerable increase in the self-perception of the ability 4 students (coefficient = .19, $SE = .14$, $t = 1.34$, $p = .18$, $d = .29$), but a decrease in that of the bottom-ability students (coefficient = -.19, $SE = .19$, $t = -1.03$, $p = .30$, $d = .29$) and in that of the highest-ability students (coefficient = -.17, $SE = .23$, $t = -.73$, $p = .47$, $d = .26$). In math, the high-stakes policy raised the self-perception of the ability 3 (coefficient = .16, $SE = .17$, $t = 2.03$, $p = .33$, $d = .20$) and ability 4 students (coefficient = .35, $SE = .15$, $t = 2.42$, $p < .05$, $d = .44$), but lowered the self-perception of the ability 1 (coefficient = -.68, $SE = .33$, $t = -2.03$, $p < .05$, $d = .86$) and of ability 2 students (coefficient = -.16, $SE = .16$, $t = -.99$, $p = .32$, $d = .20$).

Under the positive expectation, the policy effects were trivial on students’ reading self-
perception, regardless of their ability levels; however, in math, the test-based retention policy negatively affected the self-perception of the ability 4 group (coefficient = -.26, SE = .13, t = -2.01, p < .05, d = .33) while exerting positive influence on that of the ability 2 group (coefficient = .20, SE = .10, t = 2.04, p < .05, d = .25).

**Moderating role of teacher expectations.** Through multivariate hypothesis testing, I found that teacher expectations differentiated the policy effects only in math and only within the ability 4 group (χ² = 11.00, df = 3, p < .05) (see the rows of Hypothesis Testing in Table 8.2); the effect was .61 points larger (SE = .19, χ² = 10.83, p < .01) under the indifferent expectation than under the positive expectation.

**Across-ability levels Comparison of the Policy-by-Teacher Effects**

Did the policy-by-teacher treatments produce differential effects pattern across different ability subpopulations? To explore whether there was any difference in the policy effects for a given expectation, I performed omnibus hypothesis testing across the five ability groups. The results were only significant in math under the indifferent expectation (χ² = 22.32, df = 5, p < .001). Post-hoc tests indicated that under the indifferent expectation, the test-based retention policy led to an increase in the math self-perception of the ability 4 students, which was significantly different from the reduction in the math self-perception of the ability 1 (contrast = 1.03, SE = .32, χ² = 10.64, p < .01), ability 2 (contrast = .51, SE = .22, χ² = 5.53, p < .05), and ability 5 (contrast = .46, SE = .24, χ² = 3.79, p < .05) students; and the positive policy effect in math for the middle-ability students was significantly from the negative effect for the bottom-ability student (contrast = -.84, SE = .34, χ² = 6.15, p < .05).
Student Long-term Self-perception

The last part of analysis, again, seeks suggestive answer to the question: Did the effects sustain till two year later? I first compared student overall self-perception in spring 2002 and 2004, but did not discover any differences between the years – the policy effects remained insignificant on student overall self-perception in both years. I then turned to student differential self-perception by prior ability and found that all the effects and non-effects of the test-based retention policy that I observed in the promotional gate year lasted till two years later; but there were two exceptions within the ability 2 group: this group of student, if treated by the indifferent expectation and test-based retention policy in grade 3, would experience improved self-perception in math over time ($\chi^2 = 5.70, p < .05$); however, if they were taught by negative-expectation teachers, their reading self-perception would deteriorate ($\chi^2 = 4.22, p < .05$) two years after the exposure to the high-stakes testing policy.

Summary

In this chapter, I focused on student self-perceived competence and interests and examined how it was subject to the joint influence of the grade-3 test-based retention policy and teacher expectations. Below is a summary of the findings organized into two sections: (a) effects of the grade-3 test-based retention policy, and (b) effects of teacher expectations.

Effects of the Grade-3 Test-based Retention Policy

I organize the findings about the policy effects by the three aspects of student self-perception. In each subsection, I first revisit the relevant hypothesis presented in Chapter 2 and then highlight the key findings from the current chapter.

Student overall self-perception. I hypothesized that the test-based retention policy would
not improve students’ overall self-perceived competence and interests in the tested subjects.
Consistent with the hypothesis, I found that, conditioning on teacher expectations, the policy had no influence on students’ overall self-perception; in either subject, the policy effect was similarly insignificant across teacher expectations.

**Student differential self-perception by prior ability.** No hypothesis has been established regarding the specific differential self-perception pattern under each expectation.

In this chapter, I did not detect any significant conditional effect of the test-based retention policy for any ability group after controlling for differences in teacher expectations. Nevertheless, the patterns of the policy effects were different under different expectations. Under the negative expectation, the policy benefited the math self-perception of the ability 4 students, but hurt that of the bottom-ability students. Under the positive expectation, the policy exhibited negative effects on the math self-perception of the ability 4 students and positive effects on that of the ability 2 students. Under the indifferent expectation, the high-stakes policy was found helpful to the academic self-perception of the ability 3 and ability 4 students, but destructive to that of the other students, depending on subject areas. Only when teachers were indifferent toward the students’ learning ability, would the policy effects vary across the ability groups.

I discovered that for a given ability group, the policy effects on student math/reading self-perception were mostly similar across teacher expectations. But the policy worked better under the indifferent expectation than under the positive expectation for improving the math self-perception of the ability 4 group.

**Student long-term self-perception.** I predicted that if operated with the negative expectation, the test-based retention policy might have negative residual effects on student self-perceived competence and interests after the promotional gate grade; however, if operated with
the positive expectation, the policy might not have any long-term negative effects on student academic self-perception.

My analysis results were somewhat different from the hypothesis. It was found that most of the effects and lack-of-effects of the test-based retention policy were persistent over time. However, the ability 2 students would experience a significant recovery from the negative short-term policy effects on their math self-perception if being exposed to the indifferent expectation in grade 3; but they would suffer in a long run from a worsening effect of the test-based retention policy on the reading self-perception if once subject to the negative expectation.

**Effects of Teacher Expectations**

The evaluation of student overall self-perception in reading/math did not reveal any effects of teacher expectations in either policy condition. Nonetheless, the analysis of the differential self-perception by prior ability suggested that the effects of teacher expectations varied by policy contexts. In the standardized testing only condition, the positive expectation generated a lower reading self-perception of the ability 2 students than did the indifferent expectation. In the test-based retention condition, the positive expectation was associated with a lower math self-perception of the ability 4 students than was the indifferent expectation, and with a higher reading self-perception of the bottom-ability students than was the negative expectation.

**Note to Chapter 8**

1 Here I used the standard deviation for the whole ECLS grade-3 sample to compute the corresponding effect sizes. According to the ECLS grade-3 psychometric report (2005), the third graders’ academic self-perception had a standard deviation of 0.66 in reading and 0.79 in math.
CHAPTER 9
DISCUSSION

I have focused my dissertation on analyzing the causal effects of the grade-3 test-based retention policy and understanding the role of instructional capacity in modifying the policy effects. In this last chapter, I connect the results from the previous chapters and discuss their conceptual implications for the ongoing high-stakes testing debate, for school improvement under the current test-based accountability reform, and for research and practices of teacher effects. I also explore the methodological implications of my analytic approach for studies of complex educational policy issues. At the end, I discuss the limitations of the current study and propose directions for future research.

Conceptual Implications

Assumptions of High-stakes Testing

High-stake testing perhaps is one of the most controversial issues under the current accountability regime. The advent of the NCLB Act has triggered an explosion of high-stakes testing practices in the current educational accountability system and has brought the related debate to forefront. This dissertation contributes to the debate through evaluating the grade-3 test-based retention policy. Analyzing different teaching and learning outcomes, it allows me to empirically examine theoretical assumptions regarding teachers’ instructional capacity, their teaching practices, and student motivation underlying the use of high-stakes testing in the grade 3 context.

Assumption 1. High-stakes Testing and Instructional Capacity. One implicit assumption that has been ignored by many educational researchers is that teachers have the instructional capacity to produce desired teaching and learning outcomes to meet policy
expectations; if not, they will automatically acquire the capacity if rewards/sanctions are attached to student test performance. The results from this study are in contradiction with this assumption. The descriptive statistics of the measure of teacher expectations in Chapter 4 (see the Measures section) indicate that not all teachers held the same level of beliefs or self-efficacy. Although most of the sampled teachers held a positive expectation, some other teachers did not believe in or were indifferent about their students’ learning ability. My exploration of the construct of teacher expectations found dramatic differences between the three types of teachers; and the differences were not solely attributed to school and classroom contexts. According to the distribution of the final analytic sample (see Table 4.4), even after accounting for the differences in school demographic characteristics and student composition, teachers did not universally hold a positive expectation. As suggested by previous research (Hoy et al, 2008; Hughes et al, 2005; Ready & Wright, 2011; Tiedemann, 2002; Timperly & Philips, 2003; van den Bergh et al., 2010), teacher expectations may reflect some important attributes such as teachers’ knowledge and skills, their beliefs of gender/ethnicity differences, and the self-efficacy. These attributes may be deep-seated in individual teachers and may vary from teacher to teacher. Chapter 5 revealed an insignificant relationship between the grade-3 test-based retention policy and teacher expectations. The result further speaks against the implicit assumption and confirms that simply tying high-stakes incentives to student performance is inadequate for changing teacher expectations. Although more research is needed to examine the relationship between the policy and other aspects of instructional capacity, the current finding partly corroborates the statement of Cohen (1996a, 1996b) and Elmore (2003, 2008) that high-stakes incentives may mobilize existing instructional capacity, but may not produce new capacity.
Assumption 2. High-stakes testing and Teaching. Policymakers assume that high-stakes testing can promote teaching. The findings from this study show that the policy can change teaching, but cannot mandate the direction and nature of the changes. In Chapter 6, I reported that teachers uniformly increased the amount of time for math instruction under the test-based retention policy and kept reading and science time unchanged. Although both reading and math are the focuses of the current accountability system, according to Spillane (2005), primary-grade teachers may have more integrated advisory networks and more professional development opportunities for reading instruction than for math instruction. One potential explanation for the observed differences in the policy effects is that the teachers significantly increased math time but not reading time because they relied more on their individual instructional capacity for teaching math and thus felt greater pressure to improve student math performance on their own. Past survey research indicated that teachers prioritized tested subjects (see review of Jones et al., 2003; Herman, 2004). The results from this study lead to a conclusion that is partially consistent with, but somewhat different from the previous findings. They suggest that high-stakes incentives can direct teachers’ attention to some but not all tested subjects. Judging from the differences between reading and math instruction in primary schools, I suspect that when responding to high-stakes incentive(s), teachers may mostly focus on the tested subject where scarce professional support is available within school. In addition, the results from the between-subjects comparisons of Chapter 6 imply an unintended consequence of the attention shift. As the total amount of instructional time is often fixed in a school, the increased attention to math may come at the expense of other subjects or activities. Although this study did not examine other teaching practices than instructional time allocation, Chapter 7 showed us that depending on teacher expectations, the test-based retention policy led to uneven learning distribution
between subjects, across different cognitive skills within tested subjects, and among students of different prior academic ability. If we view changes in learning as a direct consequence of teaching, the results hint that instructional changes in high-stakes context, albeit not uniform across teacher expectations, may be limited to superficial responses including shifting curricular emphasis and reallocating teaching resources among students, which echoes the previous research findings that high-stakes testing has limited power in influencing teachers’ pedagogy (see review by Herman, 2004).

Assumption 3. High-stakes testing and student motivation. The rise of high-stakes testing may also be based on the assumption that the policy improves student motivation; and the policy effects on student learning are expected to channel through student motivation in addition to through teaching practices. However, this assumption is questionable in a grade-3 context as early- or middle-grade students are still in the process of developing their motivation and may not hold a stable view of their academic competencies. Chapter 8 shows that the test-based retention policy in general had no effects on students’ overall self-perception in reading and math, but exerted different influence on students of different ability levels, according to teacher expectations. If operated with the negative expectation, the test-based retention policy would raise the math self-perception of students with somewhat higher-than-average ability (i.e. ability 4 students), but would lower that of the bottom-ability students. If operated with the positive expectation, the policy would benefit the math self-perception of the ability 2 students who were in the lower end of the ability distribution but hurt that of the ability 4 students. If operated with the indifferent expectation, the policy would be beneficial to the academic self-perception of the middle-ability or ability 4 students, but be detrimental to other students, depending on subject areas. The observed differential effects pattern does not correspond to that of student differential
academic performance described in Chapter 7, suggesting that students’ academic self-
perception is not a strong predictor of their short-term academic performance. The effects pattern
under each expectation is puzzling, which I suspect is due to the composite nature of the
academic self-perception measure; the policy effects on students’ self-perceived competencies
may not be the same as those on the intrinsic motivation. Although much still needs to be learnt,
it is evident that the test-based retention policy would lower some students’ confidence and
interests in certain classroom contexts during their grade-3 school year. Moreover, the results
from the analysis of student long-term self-perception, though non-causal in nature, somewhat
suggest that the high-stakes testing policy may have residual negative effects on some students
(e.g. ability-2 students with negative-expectation teachers) in a long run.

In summary, this study adds to the literature by directly challenging the assumptions
underlying the use of high-stakes testing. The results pinpoint the flaws in the prevalent
assumptions discussed above. The work involved in improving the ultimate policy outcome,
student academic performance, should be far more complicated than what policymakers have
anticipated. The findings speak to the policymakers by calling for more realistic expectations of
the policy functions and for more attention to the unintended consequences of high-stakes
testing. In addition, the study suggests that the test-based retention policy may interfere with the
natural process of motivational development and may hurt some students’ self-perceived
competence and interests in certain grade-3 classrooms. In the current era that emphasizes
educational standards and accountability, quite a few states and districts are moving toward test-
based requirement for promotion at key transitional points including grade 3. My findings
present a need for policymakers to evaluate the appropriateness of pushing the test-based
retention policy to grade 3.
Instructional Capacity, High-stakes Testing, and Student Academic Performance

By highlighting the role of instructional capacity as indicated by teacher expectations, this study also provides a fresh perspective regarding the relationship between high-stakes testing and student academic performance. Chapter 7 identified both similarities and differences in the policy effects on student academic performance between the three types of teacher expectations. It is found that regardless of teacher expectations, the increase in math time revealed in Chapter 5 was not translated into student learning. This finding is in conformity with Cohen et al.’s discussion (2003) that access to resources itself does not cause learning; what really matters may be how teachers and students use the resources. According to the results from the analysis of student mastery of cognitive skills, teachers might have spent the increased math time, at least part of the time, in calibrating their instruction focus based on their personal judgment of student learning ability. Despite the different effects patterns between teacher expectations, changes in student math learning generally occurred as a form of tradeoff between different skills and were not always positive.

In addition to the general trend, this study also showed that students responded differently to the test-based retention policy according to the expectations they received from the grade-3 teachers. In the classrooms of negative-expectation teachers, the policy was ineffective to students’ overall math learning, but somewhat hurt the overall learning in other subjects. As a result of the test-based retention policy, there were significant gains in the learning of reading and math skills that are below grade-3 curriculum standard, but a learning loss in an advanced math skill. The policy was found detrimental to the middle-ability students and even reduced the science learning of the highest-ability students; the negative effects for the middle-ability students seemed to last over time.

In the classrooms of positive-expectation teachers, the test-based retention policy had a
negative influence on student overall science performance. While inducing an increased learning of the advanced math skill of rate and measurement, the policy failed more students taught by positive-expectation teachers into the lowest reading proficiency category. Under the positive expectation, the high-stakes policy benefited the students who stood around the middle of the ability distribution, but did certain harm to the bottom-ability students; the short-term positive effects persisted till two years later, but the short-term negative effects were proved as only temporary.

In the classrooms of indifferent-expectation teachers, students on average did not respond to the test-based retention policy in any subject; but the policy impeded students’ learning of a focal grade-3 math skill (i.e. multiplication/division). Surprisingly, under the indifferent expectation, the high-stakes policy also produced a dramatic improvement in the science learning of the bottom-ability students with a moderate negative effect only on the math learning of the ability 2 students. In a long run, the positive short-term effects under the indifferent expectation for the bottom-ability students faded away; but two years after the exposure to the test-based retention policy, certain improvement was observed for students in the lower end of the ability distribution.

The above findings confirm my theory that teachers respond to high-stakes testing in light of their own instructional capacity. And the observed differential policy effects by student prior ability are consistent with the effects on student mastery of cognitive skills if we consider the compatibility between skill focus and students’ academic ability. Negative-expectation teachers may lower the difficulty of their teaching content due to their low expectation of student learning ability; they may prioritize the bottom-ability students as a way to boost up the average class performance and to avoid high retention rate; however, due to limited instructional capacity
and/or unchanged pedagogy, the triage and the emphasis on lower-level skills are not able to bring out any effective learning of the bottom-ability students, but hurts the learning of the average-ability students, both in the promotional gate grade year and in a long run. Positive-expectation teachers may not intentionally differentiate students by their ability levels in the instruction as the teachers are optimistic about the students’ academic competency. However, positive-expectation teachers may introduce more advanced knowledge, which benefits the middle-ability students but does not gear toward the learning needs of the bottom-ability students. The positive expectation conveyed by the teachers may shield students from reduced intrinsic motivation and thus enables the students to recover from the negative policy effects in a long run. The long-term policy effects under the indifferent expectation are somewhat perplexing, which I consider, may be attributed to students’ learning experience during the unexamined time period between spring 2002 and 2004. However, it seems that in the promotional gate grade, indifferent-expectation teachers do not respond to the policy as vigorously as the other two types of teachers. Unlike the positive-or negative-expectation teachers whose instructional changes put some students’ science learning at a disadvantage, indifferent-expectation teachers can help the bottom-ability students to achieve a short-life gain in science learning. The radical improvement in the science learning of the bottom ability students hint that the indifferent-expectation teachers might resort to other factors than student academic ability in deciding on how to distribute instructional resources among students. However, lack of attention to indifferent-expectation teachers has been a major missing piece in the teacher expectancy literature. More empirical investigation is needed in order to understand what exactly influenced the teaching decision of the indifferent-expectation teachers.

My study shows that the student outcomes of high-stakes testing are framed by teacher
expectations that reflect teachers’ knowledge skills, and/or personal beliefs. Most of the previous research neglected the variation in individual actions and the role of instructional capacity. To demonstrate the potential consequences of this over-simplistic approach, I estimated the effects of the grade-3 test-based retention policy on student mastery of the highest level math skill without taking into consideration teacher expectations. Students’ learning of this skill was found to remain unchanged under the high-stakes testing policy (coefficient = -.03, odds = .97, $SE=.18$, $t = -.17, p > .5$). The results disguised the fact that the policy had a significant negative effect on the learning of rate and measurement under the negative expectation, a positive effect under the positive expectation, and a zero effect under the indifferent expectation (see Table 7.3). Ignoring the differences between classrooms led to the analytic result of little theoretical and practical significance.

O’Day (2004) pointed out that one of the major problems of the school-based accountability is that the school is the unit of intervention, yet the individual is the unit of action. Through multiple lenses, my analysis has illustrated that the ultimate policy effects depend on how teachers enact and interpret the policy and policy context. In order to better understand how high-stakes testing exerts its influence in real classrooms, educational researchers need to shift their focus from the macro-level analysis of institutional environment to the multilevel analysis that connects the policy with classroom practices.

For decades, previous empirical studies on high-stakes testing have been mostly devoted to the debate of the policy effects without providing any effective solutions to school improvement under the high-stakes regime. As Cizek (2001) argued, to monitor education quality, we have to evaluate students and schools with large-scale standardized tests; there is simply no way to avoid making high-stakes decisions in order to hold schools and individuals
accountable. Since this study shows that high-stakes testing is not able to produce large-scale positive changes in student learning and has negative consequences for some students in some contexts, considering the inevitable existence of high-stakes testing, one question naturally emerging is: how can we avoid these negative consequences and improve student learning in high-stakes contexts? This question is crucial to educators and policymakers, especially at present when the NCLB Act has been sweeping almost every corner of the current educational system and has induced various high-stakes testing practices at various levels of the American public school system. Although my study focuses on high-stakes testing policy for student accountability, it has suggested that teachers’ instructional capacity may interfere with the policy implementation process and directly determines the policy outcomes. It also demonstrated that provision of high-stakes incentives itself cannot exert any significant influence on instructional capacity. Therefore, one potential answer to the question may lie in enhancing teachers’ instructional capacity and thus improving teaching practices.

Spillane’s study (2001) has informed us that to reconstruct their practices under institutional pressure, teachers first have to believe in their students’ learning ability. My study has substantiated this argument by showing some benefits of the positive expectations over the other expectations in moderating the policy effect, such as providing more access for students to advanced knowledge, producing higher short-term learning gains of average-ability students, and inducing more resilient student learning over a long term. It stresses the importance of changing teachers’ belief and values about what is possible to do for school improvement in the current test-based accountability system. However, this study has also revealed that having a positive expectation toward student learning alone cannot avoid all negative policy effects, especially the negative effects in nontested subjects. If implemented by a positive-expectation teacher, high-
stakes testing can be detrimental to student overall science learning and to students’ learning of basic reading/math skills and can as well place the bottom-ability students at a disadvantage.

Elmore, Peterson, & McCarthey (1996) pointed out that “to be accomplished practitioners, people have to be committed, enthusiastic, knowledgeable, and skilled…but they also have to have the capacity to reflect on their practice in a community of colleagues who know deeply what the problems are (p. 236)”. Clearly, raising expectations of individual teachers alone is not sufficient for school improvement under the high-stakes regime. The results from this study present a need to seek more attributes of instructional capacity that can help alleviate the negative effects of high-stakes testing and that can lead to large-scale improve in teaching and learning. One such attribute, as Elmore et al. mentioned, is related to the organizational environment in which teachers conduct their work. Researchers should examine how the organizational and individual attributes interact with each other in modifying the high-stakes testing effects. By doing so, we can identify optimal attributes of instructional capacity and find solutions to school improvement under the current test-based accountability reform.

Organizational Contexts for Teacher Effects

This study also suggests that effects of teacher expectations, including the effects on instructional time allocation, on student mastery of cognitive skills and on differential academic outcomes by prior ability, should not be assumed constant across policy contexts, such as high- and low-stakes contexts. Although more knowledge is needed in order to understand the differences in the expectation effects between the standardized testing only condition and the test-based retention condition, the findings challenge a non-situational view of psychological research on teacher expectations. They also speak to research of teacher effects in general and emphasize the importance of organizational contexts in examining teacher effectiveness. My
study results, to a certain extent, support Cohen and Moffitt’s statement (2009) that instructional capacity is relational: “it waxes and wanes in interaction with the aims that policies set, the instruments that they deploy, and the environments in which policy and practices subsist (p. 39). 

In a recent publication of *Educational Researcher*, Kennedy (2010) admonished that previous research have been focusing too much on the characteristics of teacher themselves and have overlooked situational factors that may have a strong bearing on the quality of teaching practices. When seeking personal characteristics that account for effective work of teachers, researchers should also take into consideration the conditions where the work is conducted. Similarly, in designing school professional development, policymakers and educational administrators should notice that in addition to teachers’ knowledge and skills, organizational environment is an equally important attribute that can influence teaching and learning.

**Methodological Implications**

In Chapter 3 I identified several major methodological challenges for studies of high-stakes testing. This study meets these challenges by using low-stakes ECLS-K tests to measure individual students’ academic performance in a nationally representative sample, by explicating the characteristics of the research context and sample, and most importantly, by combining MMW-S with multilevel modeling to remove selection bias associated with a large number of covariates in a multilevel setting. Hong and her colleagues (Hong, 2010a, 2010b, 2011; Hong et al., 2011; Hong & Hong, 2009; Hong et al., under review) have explored the potential of this novel statistical approach for evaluating the causal effects of multi-dosage or multiple concurrent treatments as well as for estimating moderated or mediated treatment effects. In this dissertation, I have extended the MMW-S method to a causal analysis where the class-level treatment assignment is sequential to school-level assignment and have developed a multilevel
stratification and weighting procedure. To demonstrate its effectiveness in removing selection bias, I estimated the policy and teacher effects on student math performance using a multilevel model (see Equations 7.1-7.3) without any statistical adjustment (i.e. naïve model). The naïve analysis results show that the policy effects are positively significant under the positive expectation (coefficient = 1.67, $SE = .70$, $t = 2.38$, $p < .05$) and are insignificant under the negative expectation (coefficient = .30, $SE = 1.17$, $t = .26$, $p > .5$) and indifferent expectation (coefficient = 1.97, $SE = 1.32$, $t = 1.49$, $p = .14$). The results deviate from the weighted analysis results that suggest zero policy effects across teacher expectations (see Table 7.1). As past research mostly neglected the variations in teacher expectations and found that high-stakes testing significantly improved student math performance, to make the comparison more informative, I further conducted both weighted and naïve analysis of the single policy effect on student math performance. Applying the multilevel marginal mean weights, I still failed to identify a significant policy effect (coefficient = .14, $SE = .47$, $t = .30$, $p > .5$). However, without adjustment for selection bias, the test-based retention policy appeared to significantly improve students’ math learning by 1.48 ($SE = .66$, $t = 2.24$, $p < .05$). The naïve analysis generated a conclusion significantly different from the weighted analysis, but similar to the previous research findings, which casts doubts on the efficiency of previous statistical approaches in removing the selection bias in the studies of high-stakes testing.

The multilevel MMW-S approach that I developed in this dissertation has important implications for studies of complex policy issues, especially for the current heated topic of teacher effectiveness. According to Elmore (2008), effective teaching is shaped not only by individual characteristics that teachers and students bring to the instructional core, but also by organizational surroundings. The multilevel MMW-S approach provides a solution for
examining the interplay between the individual and organizational attributes and their causal effects on teacher and student outcomes.

**Limitations of the Current Study**

While this study has sharpened our understanding of the issues of high-stakes testing and/or teacher effectiveness and has important conceptual and methodological implications, I acknowledge the following major limitations of my data and methods:

1. The measure of test-based retention policy was constructed based on school administrators’ self-reports and may be subject to measurement errors. As the related items were only available in the fifth wave of the ECLS school administrator questionnaire, I was not able to validate the school-level treatment measure.

2. The measure of teacher expectations might as well suffer from measurement errors due to teachers’ self-reports; and the relationship between teacher expectations and instructional capacity has not been formally tested. Although I have explored the differences between the three types of teacher expectations and provided certain evidence for its construct validity and for its potential link to instructional capacity, more studies are needed to further validate this measure and to empirically test the link.

3. In the ECLS data, there was no information available regarding the year in which the test-based retention policy came into place in each treated school. Thus it is difficult to take into account the differences between early and late policy implementers; and it is also difficult to draw a reference point for deciding on prior-policy characteristics. I overcame the limitations by using only two types of pretreatment variables: (a) demographic measures of students, classes and schools in grade 3 and (b) student prior academic and social emotional status and learning experience at the kindergarten entry. Nevertheless, there is a chance that this approach may not
impose sufficient control over the selection bias and may lead to inflated or attenuated estimation of the policy and teacher effects.

4. ECLS did not have information on the second grade between the fourth and the fifth waves of data collection and only assessed students during each sampled academic year. To construct the measure of student prior ability, I used a value-added model-based approach to extrapolate the K-1 literacy growth trajectory to the end of grade 2. But I did not find qualified data to validate the estimated measure. The potential measurement errors may interfere in my findings regarding the differential policy effects by student prior ability.

5. At the time of my analysis, ECLS hasn’t released the item level information about student self-perceived competence and interests, which prevented me from separating the outcomes of intrinsic motivation and self-perception of academic ability in analyzing the policy-by-teacher effects. The ambiguous outcome measure may explain the perplexing findings regarding the differential effects on student academic self-perception. Future research will analyze different aspects of early-graders’ motivation using item-level data.

6. Due to the data constraint and study scale, I did not empirically test the three assumptions for causal inference. Below I illustrate with scenarios in which the assumptions may not hold. The assumptions of no interference between schools would be violated, say, if two students from different schools lived in a same neighborhood and competed with each other, which might lead to increased learning of both students. The assumption of intact schools and classes would be violated, for example, if parents could change schools or choose teachers for their kids. In addition, there is a possibility that my selection of pretreatment variables did not exhaust all important confounders of the treatments effects. Suppose that the grade-3 test-based retention policy was coupled with, but was not a cause of, other school initiatives, such as
professional development activities and teacher evaluation policy that might change teachers’ behavior and student learning. Then the ignorability assumption would not hold as I did not control for school climate in grade 3. Unfortunately ECLS data does not allow me to directly test the existence of these scenarios.

7. The lack of grade-4 data in the ECLS data prevents me from performing longitudinal analysis and from generating causal inference regarding the long-term effects. Therefore, the related results are only suggestive in nature and are subject to alternative explanations.

8. The large-scale survey data is not sufficient for revealing the specific teaching strategies and teachers’ decision-making process under the high-stakes testing. My theoretical framework views instructional practices as an important mediator of the policy effects and considers that teacher expectations influenced teachers’ judgment and thus modified their teaching behavior. Although students’ learning outcomes have offered me indirect evidence, I have not directly tested the theory with the ECLS data. Future analysis can explore the ECLS survey items on the teaching practices of individual teachers. But to understand what was actually happening within classrooms and schools, further investigation will need to complement the survey with first-hand data collection methods, such as classroom observation and interviews with teachers, students, and school administrators. First-hand data are also needed in order to answer some questions that emerged from the current study. For example, why was the increase in the math time not translated into learning? Why was the effects pattern of student differential academic performance by prior ability not consistent with the pattern of student differential self-perception? Who are the teachers with each type of expectations, especially who are the indifferent-expectation teachers, and what did they do in respond to the high-stakes testing policy? Why did teacher expectations exert different influence in different policy contexts?
Although I have provided some tentative, these questions should be empirically examined with multiple data sources.

9. A major concern of using propensity score-based methods is the generalizability of the findings. Although I employed a nationally representative dataset, the final analytic sample has different student, teacher, and school composition from the original sample. To maximize the generalizability, I have specified the characteristics of the sample where the findings are situated in. However, the results are yet to be corroborated through replication studies with different datasets and/or different methods.

**Future Research Agenda**

This dissertation study has suggested several lines of work for further policy research. One line of work is to examine the generalizability of the current findings. Future research can extend to different test-based incentives such as high-school graduation exam, teachers’ merit pay, and school probation and investigate the individual or joint effects of these incentives on student and teacher outcomes in different domains. The second line of work is to gain more understanding about teaching behaviors under high-stakes testing and to corroborate the conceptual framework that I developed in Chapter 2. In addition to a further analysis of the ECLS data regarding instructional practices in grade 3, it is also necessary to take a closer look into classroom and to examine the process where teachers interpret and enact the high-stakes testing policy in light of their expectations. Other data sources, such as survey of school administrator and direct observation of their work, may also be useful to understand the environment that fosters the instructional process. The third line of work is to empirically study the relationship between teacher expectations and instructional capacity and to identify optimal attributes of instructional capacity for improving student learning under high-stakes regime. This
experimenting process may require multiple data sources and various approach of inquiry. It is also interesting to see how the attributes of instructional capacity may differ by policy contexts and how different attributes interact with each other in influencing student learning. The last line of work is to seek effective strategies for improving the identified attributes of instructional capacity in a given policy context.

In addition to the four lines of policy work, this dissertation study has also sparked some ideas for future methodological studies. To evaluate the joint effects of the grade-3 test-based retention policy and teacher expectations, I have developed a procedure for applying MMW-S method to estimating the causal effects of multilevel treatments. The procedure was used to approximate a sequential randomized design. Future methodological research should further validate the procedure using simulation data. An alternative approach is to view the treatments as being assigned simultaneously and thus obtain a joint class-level marginal mean weight based on the joint probability for the treatments assignment. It is necessary to compare this alternative approach with the current one in terms of their effectiveness in removing selection bias. Application study is as well needed to explore the potential of the multilevel MMW-S method in investigating various complex policy issues. The approach may also be useful for some of the policy studies that I suggested above.
APPENDIX A

Empirical Identification of Student Prior Academic Ability

To evaluate the effects of the grade 3 test-based retention policy on students’ K-1 reading growth, I analyzed a three-level weighted model. At level 1, I combined OLDS and reading outcomes \( Y_{ik} \) for student \( i \) of school \( k \) at time \( t \) and differentiated them using two dummy indicators, \( D_{READ} \) and \( D_{OLDS} \):

\[
Y_{ikt}^* = D_{READ_{ik}} \times [\pi_{ik} \text{Intercept}_{ikt} + \pi_{2ik} (\text{Dur}_{ikt}^* - 12) + \pi_{3ik} (\text{Dur}_{ikt}^* - 12)^2] + \\
D_{OLDS_{ik}} \times [\pi_{4ik} \text{Intercept}_{ikt} + \pi_{5ik} (\text{Dur}_{ikt}^* - 12) + \pi_{6ik} (\text{Dur}_{ikt}^* - 12)^2] + e_{ikt}^*,
\]

\( e_{ikt}^* \sim N(0, \sigma_{ikt}^2) \) (A.1)

I centered the duration from kindergarten entry to the assessment (\( Dur \)) at the 12th month, i.e. the end of kindergarten, and used its linear and quadratic terms to capture the amount of academic growth that differed between kindergarten and grade 1. To enable a weighted model with homogeneous level-1 variance, I rescaled the outcome and the predictors for OLDS scores using the square root of the ratio between OLDS and reading scores variance at each assessment time \( t \): \( \rho_t = \sqrt{\text{Var(OLDS)}_t / \text{Var(READ)}_t} \). Specifically, for OLDS scores: \( Y_{ikt}^* = Y_{ikt} / \rho_t \), \( \text{Intercept}_{ikt}^* = 1 / \rho_t \), \( \text{Dur}_{ikt}^* = \text{Dur}_{ikt} / \rho_t \), \( \text{Dur}_{ikt}^2 = \text{Dur}_{ikt}^2 / \rho_t \), and \( e_{ikt}^* = e_{ikt} / \rho_t \); for reading scores, the variables remained the same as before: \( Y_{ikt}^* = Y_{ikt} \), \( \text{Intercept}_{ikt}^* = 1 \), \( \text{Dur}_{ikt}^* = \text{Dur}_{ikt} \), \( \text{Dur}_{ikt}^2 = \text{Dur}_{ikt}^2 \), and \( e_{ikt}^* = e_{ikt} \). A level 2, I assumed a same acceleration rate for all individual students and fixed the random effect of the quadratic terms to allow more degrees of freedom.

\[
\pi_{hik} = \beta_{h0k} + \beta_{hk} \sum_{l=1}^{L-1} X_{l ik} + r_{ik}, \text{ for } h = 1, 2, 4, 5, \text{ } r_{ik} \sim N(0, \tau_{\pi})
\]

(A.2)

To improve precision of the estimation, I controlled for \( L = 9 \) important variables of student
demographic characteristics, including student’s gender, SES, language background, age at the kindergarten entry, as well as five dummy indicators of his/her ethnicity. To remove selection bias associated with pre-policy variables, I applied the school-level marginal mean weight at level 3:

\[
\beta_{h0k} = \gamma_{h00} + \gamma_{h01} Z_k + u_{h0k}, \quad \text{for } h = 1, 2, 4, 5, \quad \nu_k \sim N(0, \tau),
\]

\[
\beta_{h0k} = \gamma_{h00} + \gamma_{h01} Z_k, \quad \text{for } h = 3, 6, \quad \beta_{hlk} = \gamma_{h00}
\]

The results from the three-level weighted model suggested that no significant effect of the test-based retention policy \((Z_k)\) on students’ K-1 reading linear growth (coefficient = .00, \(SE = .00, t = .05, p > .5\)) or quadratic growth (coefficient = .00, \(SE = .00, t = -.01, p > .5\)); neither was there any significant policy effect on students’ K-1 OLDS linear growth (coefficient = -.19, \(SE = .18, t = -1.08, p = .28\)) or quadratic growth (coefficient = -.00, \(SE = .01, t = -.21, p > .5\)).

As the policy effects were found negligible, with the same model, I obtained from the level 2 and level 3 residual files the empirical Bayes estimates of the reading status \(\pi_{1ik}\) and reading linear growth \(\pi_{2ik}\) for student \(i\) attending school \(k\) and predicted value of \(\gamma_{300}\). I computed the end-of-grade 2 reading score for each individual student as

\[
\pi_{1ik} + 24 \times \pi_{2ik} + 24^2 \times \gamma_{300}.
\]

The new estimated scores have a normal distribution with a mean of .70 and a standard deviation of .28. Based on the distribution, I categorized the students in the analytic sample into five equal-sized ability groups \((n = 1640\) in each group).
APPENDIX B

List of Pretreatment Variables

Student-level Pretreatment Variables

1. Female
2. Changing school between Spring 2000 and 2002 (variable)
3. Changing school between Spring 2000 and 2002 (missing indicator)
4. Home language as English (variable)
5. Home language as English (missing indicator)
6. Spring 2002 parent having Bachelor’s or higher degree (variable)
7. Spring 2002 parent having Bachelor’s or higher degree (missing indicator)
8. White
9. Hispanic
10. Asian
11. Black
12. Other ethnicity
13. Age at K entry
15. Spring 2002 continuous SES measure
16. Required to take OLDS test at K entry (variable)
17. Required to take OLDS test at K entry (missing indicator)
18. Reading score at K entry (variable)
19. Reading score at K entry (missing indicator)
20. Math score at K entry (variable)
21. Math score at K entry (missing indicator)
22. General knowledge score at K entry (variable)
23. General knowledge score at K entry (missing indicator)
24. OLDS score at K entry (variable)
25. OLDS score at K entry (missing indicator)
26. Lapse of time from K entry to the first ECLS assessment
27. Fall 1998 parent rating of student approaches to learning
28. Fall 1998 parent rating of student self-control
29. Fall 1998 parent rating of student social interaction
30. Fall 1998 parent rating of student being sad/lonely
31. Fall 1998 parent rating of student impulsive/overactive behaviors
32. Fall 1998 teacher rating of student approaches to learning
33. Fall 1998 teacher rating of student self-control
34. Fall 1998 teacher rating of student interpersonal skills
35. Fall 1998 teacher rating of student externalizing problem behaviors
36. Fall 1998 teacher rating of student internalizing problem behaviors
37. % Hispanic in student’s fall 1998 class
38. % black in student’s fall 1998 class
39. % non-Hispanic white students in fall 1998 class
40. % students who can recognize letters in fall 1998 class
41. % students who can read words in fall 1998 class
42. % students who can read sentences in fall 1998 class
43. % LEP (limited English Proficiency) students in fall 1998 class
44. % students of 4 years old or younger in fall 1998 class
45. % students of 6 years old or older in fall 1998 class
46. % girls in fall 1998 class
47. % students with preschool records in fall 1998 class
48. Fall 1998 class with students speaking non-English language (variable)
49. Fall 1998 class with students speaking non-English language (missing indicator)
50. Fall 1998 class size

Class-level Pretreatment Variables

1. Spring 2002 class aggregated measure of current students’ fall 1998 reading scores
2. Spring 2002 class aggregated measure of current students’ fall 1998 math scores
3. Spring 2002 class aggregated measure of current students’ fall 1998 general knowledge scores
4. Spring 2002 class aggregated measure of current students’ fall 1998 OLDS scores
5. Spring 2002 % students not having fall 1998 reading scores in class
6. Spring 2002 % students not having fall 1998 math scores in class
7. Spring 2002 % students not having fall 1998 general knowledge scores in class
8. Spring 2002 % students not having fall 1998 OLDS scores in class
9. Spring 2002 class aggregated measure of current students’ lapse of time from K entry to the first ECLS assessment
10. Spring 2002 class aggregated measure of current students’ approaches to learning rated by parents in fall 1998
11. Spring 2002 class aggregated measure of current students’ self-control rated by parents in fall 1998
12. Spring 2002 class aggregated measure of current students’ social interaction rated by parents in fall 1998
13. Spring 2002 class aggregated measure of current students’ being sad /lonely rated by parents in fall 1998
14. Spring 2002 class aggregated measure of current students’ impulsive/overactive behavior rated by parents in fall 1998
15. Spring 2002 class aggregated measure of current students’ approaches to learning rated by teachers in fall 1998
17. Spring 2002 class aggregated measure of current students’ interpersonal skills rated by teachers in fall 1998
20. Spring 2002 number of students in class
21. Spring 2002 % students below 9-year old in class
22. Spring 2002 % students above 9-year old in class
23. Spring 2002 % Asian in class
24. Spring 2002 % Hispanic in class
25. Spring 2002 % black in class
26. Spring 2002 % non-Hispanic white in class
27. Spring 2002 % girls in class
28. Spring 2002 % eligible students for free or reduced-price breakfast in class
29. Spring 2002 % eligible students for free or reduced-price lunch in class
30. Spring 2002 class with students speaking non-English language (variable)
31. Spring 2002 class with students speaking non-English language (missing indicator)

**School-level Pretreatment Variables**

1. Spring 2002 school in Midwest region
2. Spring 2002 school in South region
3. Spring 2002 school in West region
4. Spring 2002 urban school
5. Spring 2002 suburban school
6. Spring 2002 rural school
7. Spring 2002 public school (variable)
8. Spring 2002 public school (missing indicator)
9. Spring 2002 catholic school
10. Spring 2002 private, other religious affiliation
11. Spring 2002 school of choice
12. Spring 2002 school in USDA breakfast program (variable)
13. Spring 2002 school in USDA breakfast program (missing indicator)
14. Spring 2002 school receiving Title 1 funds (variable)
15. Spring 2002 school receiving Title 1 funds (missing indicator)
16. Spring 2002 school having program for special-needs children
17. Spring 2002 school having prekindergarten
18. Spring 2002 school having 5th grade
19. Spring 2002 school having 6th grade
20. Spring 2002 school having 7 or 8 grade
21. Spring 2002 school having no information on grade levels
22. Spring 2002 school site accommodating less than 500 students (variable)
23. Spring 2002 school site accommodating less than 500 students (missing indicator)
24. Spring 2002 school having service for students with disability (variable)
25. Spring 2002 school having service for students with disability (missing indicator)
26. Spring 2002 school having gifted/talented program (variable)
27. Spring 2002 school having gifted/talented program (missing indicator)
28. Spring 2002 adequacy of school basic facilities, including classrooms, auditoriums, multi-use rooms
29. Spring 2002 adequacy of school other facilities, including art rooms, gym, and music rooms
30. Spring 2002 total number of computers in school
31. Spring 2002 school neighborhood problem
32. Spring 2002 school with heavy traffic problem
33. Spring 2002 number of violence types in school
34. Spring 2002 number of safety measure
35. Spring 2002 total school enrollment
36. Spring 2002 school third grade enrollment
37. Spring 2002 number of students sampled in the school
38. Spring 2002 total number of regular teachers
39. Spring 2002 students/ teacher ratio
40. Spring 2002 % students moving in the school since 10/1/2001
41. Spring 2002 % students moving out of the school since 10/1/2001
42. Spring 2002 % minority students in school
43. Spring 2002 % Hispanic students in school
44. Spring 2002 % black students in school
45. Spring 2002 % LEP students in school
46. Spring 2002 % LEP students in the third grade
47. Spring 2002 school having too few students eligible for breakfast program (variable)
48. Spring 2002 school having too few students eligible for breakfast program (missing indicator)
49. Spring 2002 % students having free or reduced-price lunch in school
50. Spring 2002 school having 50% or more students from low-income family
51. Spring 2002 school aggregated measure of current students’ fall 1998 reading scores
52. Spring 2002 school aggregated measure of current students’ fall 1998 math scores
53. Spring 2002 school aggregated measure of current students’ fall 1998 general knowledge scores
54. Spring 2002 school aggregated measure of current students’ fall OLDS scores
55. Spring 2002 % current students not having fall 1998 reading scores in school
56. Spring 2002 % current students not having fall 1998 math scores in school
57. Spring 2002 % current students not having fall 1998 general knowledge scores in school
58. Spring 2002 % current students not having fall 1998 OLDS scores in school
59. Spring 2002 school aggregated measure of current students’ lapse of time from K entry to the first ECLS assessment
60. Spring 2002 school aggregated measure of current students’ approaches to learning rated by parents in fall 1998
61. Spring 2002 school aggregated measure of current students’ self-control rated by parents in fall 1998
62. Spring 2002 school aggregated measure of current students’ social interaction rated by parents in fall 1998
63. Spring 2002 school aggregated measure of current students’ being sad /lonely rated by parents in fall 1998
64. Spring 2002 school aggregated measure of current students’ impulsive/overactive behavior rated by parents in fall 1998
65. Spring 2002 school aggregated measure of current students’ approaches to learning rated by teachers in fall 1998
67. Spring 2002 school aggregated measure of current students’ interpersonal skills rated by teachers in fall 1998
68. Spring 2002 school aggregated measure of current students’ externalizing problem behaviors rated by teachers in fall 1998
69. Spring 2002 school aggregated measure of current students’ internalizing problem behaviors
rated by teachers in fall 1998

70. Spring 2002 school aggregated measure of number of students in grade-3 classes
71. Spring 2002 school aggregated measure of % students below 9 years old in grade-3 classes
72. Spring 2002 school aggregated measure of % students above 9 years old in grade-3 classes
73. Spring 2002 school aggregated measure of % Asian in grade-3 classes
74. Spring 2002 school aggregated measure of % Hispanic in grade-3 classes
75. Spring 2002 school aggregated measure of % black students in grade-3 classes
76. Spring 2002 school aggregated measure of % Non-Hispanic white in grade-3 classes
77. Spring 2002 school aggregated measure of % girls in grade-3 classes
78. Spring 2002 school aggregated measure % eligible students for free or reduced-price breakfast in grade-3 classes
79. Spring 2002 school aggregated measure % eligible students for free or reduced-price lunch in grade-3 classes
80. Spring 2002 % current classes with students speaking non-English language in school
81. Spring 2002 % current grade-3 teachers not reporting students’ class language(s)
APPENDIX C
Imputation of Multilevel Structured Data

I imputed the pretreatment variables using a multi-stage procedure to take into account the multilevel structure of the missing data. The imputation procedure assumed that the data are missing at random (MAR assumption, Rubin, 1976; Little & Rubin, 1987), i.e. the values of the missing observations are independent of the unobserved data given the observed data. I extended the single level multiple imputation method (Schafer, 1999) to multilevel data. Specifically, I decomposed each student-level variable $X_{ijk}$ for student $i$ of class $j$ from school $k$ as

$$(X_{ijk} - \bar{X}_{jk}) + (\bar{X}_{jk} - \bar{X}_{k}) + \bar{X}_{k}$$

and each class-level variable $W_{jk}$ as $(W_{jk} - \bar{W}_{k}) + \bar{W}_{k}$. Hence at the student level, I have $X_{ijk} - \bar{X}_{jk}$ (i.e. a vector of class-mean centered $X_{ijk}$); at class-level, I have $\bar{X}_{jk} - \bar{X}_{k}$ (i.e. a vector of school-mean centered class aggregate of $X_{ijk}$) and $W_{jk} - \bar{W}_{k}$ (i.e. a vector of school-mean centered $W_{jk}$); and at school level I have $\bar{X}_{k}$, $\bar{W}_{k}$, and $V_{k}$. The essence of the procedure is to first obtain an initial imputation of $X_{ijk}$ and $W_{jk}$ based on multivariate associations among variables at the corresponding levels, and then to use the aggregated values of the initial imputation to conduct the final imputation of $V_{k}$ at the school level followed by the final imputation of $W_{jk}$ at the class level and the final imputation of $X_{ijk}$ at the student level. It is assumed that the initial imputation, once aggregated to a higher level, is less affected by the incomplete covariance structure initially used. The whole procedure involved five iterative steps and was conducted using PROC MI of SAS 9.1.

Step 1. Conduct initial imputation of student-level variables. I imputed student-level variables $X_{ijk}$ only and obtain imputed variables $X'_{ijk}$.

Step 2. Conduct initial imputation of class-level variables using student-level variables...
obtained from step 1. I aggregated the imputed variables $X'_{ijk}$ to class level, and imputed class-level variables $W_{jk}$ with class-aggregated $X^*_{ijk}$ to get $W^*_{jk}$.

Step 3. Impute school-level variables using the student- and class-level variables obtained from steps 1 & 2. Based on the decomposition, at school level I need to impute school-level variables $V_k$ with $X_k$ and $W_k$. I replaced $X_k$ and $W_k$ with school aggregated $X^*_{ijk}$ and $W^*_{jk}$ and obtained final imputed values of $V_k$.

Step 4. Refine the initial imputation of class-level variables. At class level, two components need to be imputed: $X_{jk} - X_k$ and $W_{jk} - W_k$. I centered $W_{jk}$ at the school-aggregated value of $W^*_{jk}$ to impute $W_{jk} - W_k$ and substituted the corresponding form of $X^*_{ijk}$ for imputing $X_{jk} - X_k$. After the imputation, I converted the imputed $W_{jk} - W_k$ back to get final imputed values of $W_{jk}$.

Step 5. Refine the initial imputation of student-level variables. Here I imputed $X_{ijk} - X_{jk}$ by replacing $X_{jk}$ with class-aggregated $X^*_{ijk}$ and then I converted the imputed form to get final values of $X_{ijk}$.

The final imputed dataset included imputed $X_{ijk}$ from step 5, $W_{jk}$ from step 4, and $V_k$ from step 3.

Note: only non-aggregated pretreatment variables were involved in the imputation.
APPENDIX D

Estimating Multilevel Propensity Scores

I estimated the conditional probability that a grade-3 school $k$ would adopt test-based retention policy given the selected school-level covariates $V$, i.e. $\theta_z = Pr(Z = 1 | V)$. Then using a two-level multinomial logistic model, for each cluster, I estimated the propensity of adopting each of the three teacher expectations for class $j$ in school $k$ given the selected school- and class-level covariates, $V$ and $W$, as well as the observed school-level assignment $Z$, i.e.

$$\theta_i = Pr(T = t | Z, W, V).$$

Since propensity scores are the coarsest function of observed covariates, under the ignorability assumption, the class-level assignment to teacher expectation $D(t)$ is independent of the potential outcome $Y(t, z)$ given the propensity score for the treatment assignment $\theta_i$ under a given policy $z$; and the school-level assignment to policy $D(z)$ is independent of the potential outcome $Y(t, z)$ given the propensity score for the policy assignment $\theta_z$. Also, we know that policy assignment $D(z)$ is independent of $\theta_i$ when $\theta_z$ is given. Hence

$$E[Y(t, z) | D(t) = 1, D(z) = 1, \theta_i, \theta_z] = E[Y(t, z) | D(z) = 1, \theta_i, \theta_z] = E[Y(t, z) | \theta_i, \theta_z] \quad (D.1)$$

Literature is not conclusive on types of covariates to be included into propensity models (e.g. Austin, Grootendrst, & Anderson, 2009; Brookhart, Schneeweiss, Rothman, Glynn, Avorn, & Sturmer, 2006; Pekins, Tu, Underhill, Zhou, & Murray, 2000; Rosenbaum & Rubin, 1984; Rubin & Thomas, 1996; Weitzen, Lapane, Toledano, Hume, & Mor, 2004). To have a uniform set of propensity models for evaluating different outcomes, I decided to include only the confounders that were associated with both the outcomes and treatments. In case that this strategy may ignore some important pretreatment variables, in the final model for treatment effects estimation, I combined the MMW-S method with regression adjustment for several
important outcome predictors as suggested by Rubin and Thomas (2000).

Table D1 presents the confounders selected into the school-level propensity model and specifies the outcomes they were associated with. Table D2 lists the confounders for the class-level propensity models. Because MMW-S method is robust to misspecification of the functional form of a propensity model, I did not include any nonlinear terms in the school- and class-level models.
### Table D1

**List of Confounders for the School-level Propensity Model**

<table>
<thead>
<tr>
<th>Confounders</th>
<th>Teacher Expectations</th>
<th>Student Academic Performance</th>
<th>Instructional Time Allocation</th>
<th>Student Self-perception</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring 2002 school in South region</td>
<td>√</td>
<td></td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>Spring 2002 suburban school</td>
<td></td>
<td>√</td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>Spring 2002 public school</td>
<td>√</td>
<td></td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>Spring 2002 total school enrollment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spring 2002 adequacy of school facilities (i.e. art room, gym, and music room)</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spring 2002 number of safety measures in school</td>
<td>√</td>
<td></td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>Spring 2002 % black students in school</td>
<td></td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spring 2002 school aggregated measure of % eligible students for free or reduced-price breakfast in grade-3 classes</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spring 2002 school aggregated measure of current students’ fall 1998 OLDS scores</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spring 2002 school aggregated measure of current students’ approaches to learn rated by parents in fall 1998</td>
<td>√</td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Spring 2002 % current grade-3 teachers not reporting students’ class language(s)</td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
</tr>
</tbody>
</table>
Table D2
List of Confounders for the Class-level Propensity Models

<table>
<thead>
<tr>
<th>Confounders</th>
<th>Outcome(s) Associates with the Confounder</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Student Academic Performance</td>
</tr>
<tr>
<td><strong>School Cluster 1</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Class level</strong></td>
<td></td>
</tr>
<tr>
<td>Spring 2002 class aggregated measure of current students’ approaches to learn rated by teachers in fall 1998</td>
<td>✓</td>
</tr>
<tr>
<td>% current students not taking fall 1998 OLDS test in class</td>
<td>✓</td>
</tr>
<tr>
<td>Spring 2002 % students above 9-year old in class</td>
<td>✓</td>
</tr>
<tr>
<td>Spring 2002 no teacher report of students’ class language(s)</td>
<td>✓</td>
</tr>
<tr>
<td><strong>School level</strong></td>
<td></td>
</tr>
<tr>
<td>Spring 2002 urban school</td>
<td>✓</td>
</tr>
<tr>
<td>Spring 2002 public school</td>
<td>✓</td>
</tr>
<tr>
<td>Spring 2002 school in USDA breakfast program</td>
<td>✓</td>
</tr>
<tr>
<td>Spring 2002 school site accommodating less than 500 students</td>
<td>✓</td>
</tr>
<tr>
<td><strong>School Cluster 2</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Class level</strong></td>
<td></td>
</tr>
<tr>
<td>Spring 2002 class aggregated measure of current students’ fall 1998 OLDS scores</td>
<td>✓</td>
</tr>
<tr>
<td>Spring 2002 class aggregated measure of current students’ externalizing problem behaviors rated by teachers in fall 1998</td>
<td>✓</td>
</tr>
<tr>
<td>Spring 2002 no teacher report of students’ class language(s)</td>
<td>✓</td>
</tr>
<tr>
<td><strong>School level</strong></td>
<td></td>
</tr>
<tr>
<td>Spring 2002 school in South region</td>
<td>✓</td>
</tr>
<tr>
<td>Spring 2002 % students having free or reduced-price breakfast in school</td>
<td>✓</td>
</tr>
<tr>
<td>Spring 2002 school aggregated measure of current students’ fall 1998 math scores</td>
<td>✓</td>
</tr>
<tr>
<td>Confounders</td>
<td>Outcome(s) Associates with the Confounder</td>
</tr>
<tr>
<td>----------------------------------------------------------------------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>Student Academic Performance</td>
</tr>
<tr>
<td>School Cluster 3</td>
<td></td>
</tr>
<tr>
<td><strong>Class level</strong></td>
<td></td>
</tr>
<tr>
<td>class aggregated measure of current students’ social interaction rated by parents in fall 1998</td>
<td>√</td>
</tr>
<tr>
<td>Spring 2002 no teacher report of students’ class language(s)</td>
<td>√</td>
</tr>
<tr>
<td><strong>School level</strong></td>
<td></td>
</tr>
<tr>
<td>Spring 2002 number of violence types in school</td>
<td>√</td>
</tr>
<tr>
<td>Spring 2002 number of safety measures in school</td>
<td>√</td>
</tr>
<tr>
<td>Spring 2002 % current students not having fall 1998 reading scores in school</td>
<td>√</td>
</tr>
<tr>
<td>Spring 2002 school aggregated measure of % eligible students for free or reduced-price lunch in grade-3 classes</td>
<td>√</td>
</tr>
<tr>
<td>Spring 2002 % current classes with students speaking non-English language in school</td>
<td>√</td>
</tr>
<tr>
<td>School Cluster 4</td>
<td></td>
</tr>
<tr>
<td><strong>Class level</strong></td>
<td></td>
</tr>
<tr>
<td>Spring 2002 % black students in class</td>
<td>√</td>
</tr>
<tr>
<td><strong>School level</strong></td>
<td></td>
</tr>
<tr>
<td>Spring 2002 school receiving Title 1 funds</td>
<td>√</td>
</tr>
<tr>
<td>Spring 2002 number of violence types in school</td>
<td>√</td>
</tr>
<tr>
<td>Spring 2002 school with heavy traffic problem</td>
<td>√</td>
</tr>
<tr>
<td>Spring 2002 school aggregated measure of % black students in grade-3 classes</td>
<td>√</td>
</tr>
</tbody>
</table>
APPENDIX E

Strategy for Identifying Common Support

To generate causal inference, it is important for units to have counterfactual information available in the observed data. If treatment groups lack sufficient overlap in the covariate distributions, the effect estimation might be biased and be subject to reduced precision (Cochran, 1957; Hong, 2010; Imbens, 2004; Rubin, 1997). Therefore units that have a zero probability of assignment to a certain treatment and/or units that are hard to find match(es) need to be removed from a causal analysis. If there is only one single-level binary treatment, for example, the test-based retention policy, such units can be detected by comparing the logit of the propensity scores between the two treatment groups. Figure E1 shows the initial distribution of the logit scores for test-based retention policy. On the lower panel are the schools who actually adopted the policy ($Z = 1$) while on the upper panel are those who did not ($Z = 0$). The units that fall on the left of the solid line lacked empirical basis for causal inference and thus were excluded from the analytic sample.

Figure E1. Initial distribution of the school-level logit scores.
For a multi-valued treatment such as teacher expectations, in a single-level setting, common support is to be identified through pairwise comparison of the logit scores between a focal treatment group and the rest of sample. Take the logit scores for the negative expectation, \( \logit(T_{jk} = 1 | W_{jk}, V_k, Z_k) \) as an example, we can compare classes that received a negative-expectation teacher, i.e. \( D(1) = 1 \) with those that did not, i.e. \( D(1) = 0 \) and then exclude unmatchable units. Similarly we can compare \( \logit(T_{jk} = 2 | W_{jk}, V_k, Z_k) \) between \( D(2) = 1 \) and \( D(2) = 0 \) and \( \logit(T_{jk} = 3 | W_{jk}, V_k, Z_k) \) between \( D(3) = 1 \) and \( D(3) = 0 \). Through systematic pairwise comparisons, we can weed out all unqualified units from the analytic sample and identify a common support for causal inference.

However, in the current study of multilevel treatment, pairwise comparisons had to be done in each cluster because class-level stratification was nested within school clusters, i.e. the clusters defined by the dash lines in Figure E1. We know that within the initial school-level clusters, the two policy groups \( Z = 1 \) and \( Z = 0 \) were comparable in the covariate distributions under the ignorable treatment assumption. When removing unmatchable classes across the three teacher expectations, cautions have to be taken in order to maintain the initial balance between the two groups. In other words, for each cluster, if I identify a group of classes that are difficult to be matched across \( T \) within a \( Z \) condition, classes with same pretreatment characteristics should also be removed from the other \( Z \) condition even though they have matches for \( T \) under that \( Z \). For example, suppose that in a cluster, school average composition of Asian students are balanced between the two policy conditions; and at class level, if classes with 25% Asian students selected all \( T \) conditions under \( Z = 0 \), but only selected \( T = 1 \) condition under \( Z = 1 \). In other words, this group of classes have a zero probability of assignment to the other two \( T \) conditions under \( Z = 1 \), but not under \( Z = 0 \). Then all classes with 25% Asian students have to be
removed from both $Z = 1$ and $Z = 0$ conditions. If they are only excluded in $Z = 1$ condition, the
two policy groups will no longer be balanced in terms of the school average Asian composition
and as a result, the weighted estimate of $Z$ effect on outcome $Y$ will be biased.

To remove all unmatchable classes in the multilevel setting poses a challenge for
identifying counterfactual information under each treatment condition. For example, if a unit is
observed in $T = 1$ and $Z = 1$ condition, we not only want to know its propensity for $T = 2$ or 3
under $Z = 1$, but also need to find out it is probability for receiving each of the expectation levels
under $Z = 0$. To circumvent this problem, I employed an empirical Bayes approach by adopting
the same two-level multinomial models for estimating the class-level propensity scores.
Specifically, I denote the probability of adopting each teacher expectation for class $j$ of school $k$
as: $\varphi_{jk} = \Pr(T_{jk} = t)$, where $t$ takes a value of 1 or 2. For the selected school-level covariates $V$
and class-level covariates $W$:

Level 1:

$$\log \left( \frac{\varphi_{jk}}{\varphi_{jk}} \right) = \beta_{0k(i)} + \beta_{1k(i)} W_{jk}$$  \hspace{1cm} (E.1)

Level 2:

$$\beta_{0k(i)} = \gamma_{00(i)} + \gamma_{01(i)} Z_k + \gamma_{02(i)} V_k$$
$$\beta_{1k(i)} = \gamma_{10(i)}.$$  \hspace{1cm} (E.2)

Within school clusters, based on parameter estimates and Empirical Bayes estimates of the
intercepts from the level-2 residual file, two sets of predicted propensity $\Pr(T_{jk} \mid Z_k = 0, W_{jk}, V_k)$
and $\Pr(T_{jk} \mid Z_k = 1, W_{jk}, V_k)$ were computed for all classes regardless of their actual $Z$
assignment. I conducted pairwise comparison for teacher expectation groups and excluded
unmatchable classes based on each set of the predicted scores. A common support was identified
by equating the range of the valid classes across $Z$ in each cluster. To illustrate, Figure E2 displays the logit distributions of $Pr(T_{jk} \mid Z_k = 0, W_{jk}, V_k)$ and $Pr(T_{jk} \mid Z_k = 1, W_{jk}, V_k)$ in cluster 1, the cluster that experienced the largest sample reduction. On each set of logit scores, I contrasted one expectation level (the lower panel) with the rest of the sample (the upper panel) in the cluster. The classes that fall on the two ends of the solid lines did not have matches on at least one set of the logit scores and were not used for causal inference. By doing so, with the new analytic sample, balance between test-based retention and standardized testing conditions was maintained at school level. Within each cluster, there was still no significant difference between the two policies conditions in the school-level logit propensity scores.
Comparison on Logit ($T_\delta = t \mid Z_\delta = 0, W_{\delta k}, V_\delta$)

negative expectation vs. the rest

Figure E2. Distribution of the class-level logit probability scores in cluster 1.
APPENDIX F
Multilevel Marginal Mean Weights (MMW)

School-level MMW

I computed school-level marginal mean weight as:

\[ MMW_{\text{school}} = \frac{n_c}{n_{z,c}} \times Pr\{Z = z\}, \ c=1, \ldots, 4. \]  \hspace{1cm} (F.1)

Here \( n_c \) is the total number of schools in each cluster, \( n_{z,c} \) is the number of treated schools in that cluster, and \( Pr\{Z = z\} \) is the proportion of the treated units in the whole population. Table F1 presents the distribution of schools between the two policy conditions over the four clusters. For instance, the proportion of the schools that adopted the test-based retention policy \( Pr\{Z = 1\} \) is \( 159/539 = 29.50\% \), and within cluster 1, \( n_1 = 151 \) and \( n_{1,1} = 11 \). Therefore, the 11 treated schools in cluster 1 all received a weight \( (151/11) \times 29.50\% = 4.05 \). Likewise I computed weights for schools in standardized testing only condition. After weighting, schools in either condition had a similar pretreatment composition as that of the population as if the schools had been assigned to the test-based retention policy at random with a constant probability of 29.50%.

Table F1

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Logit Propensity Scores Mean (SD)</th>
<th>Unweighted Sample school n</th>
<th>MMW</th>
<th>Weighted Sample school n</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Z=1) (Z=0)</td>
<td>(Z=1) (Z=0)</td>
<td>(Z=1) (Z=0)</td>
<td>(Z=1) (Z=0)</td>
</tr>
<tr>
<td>1</td>
<td>-2.55 (0.34) -2.52 (0.40)</td>
<td>11 140</td>
<td>4.05 0.76</td>
<td>44.54 106.46</td>
</tr>
<tr>
<td>2</td>
<td>-1.44 (0.22) -1.50 (0.24)</td>
<td>17 86</td>
<td>1.79 0.84</td>
<td>30.38 72.62</td>
</tr>
<tr>
<td>3</td>
<td>-0.62 (0.25) -0.67 (0.26)</td>
<td>80 116</td>
<td>0.72 1.19</td>
<td>57.82 138.18</td>
</tr>
<tr>
<td>4</td>
<td>0.20 (0.32) 0.22 (0.25)</td>
<td>51 38</td>
<td>0.51 1.65</td>
<td>26.25 62.75</td>
</tr>
<tr>
<td>Total</td>
<td>-0.58 (0.79) -1.45 (0.99)</td>
<td>159 380</td>
<td>159 380</td>
<td></td>
</tr>
</tbody>
</table>


**Class-level MMW**

To compute class-level marginal mean weight, I treated each school cluster as a pseudo population. Within each pseudo population $c$, I computed the weights as a ratio of total number of classes in each stratum, $n(s|c)$ to the number of classes with certain expectation level in that stratum $n(s|c)$:

$$MMW_{class} = \frac{n(s|c)}{n(t,s|c)} \times Pr\{T = t \mid C = c\}, \quad s = 1, \ldots, L, \quad (F.2)$$

where $Pr\{T = t \mid C = c\}$ denotes the proportion of classes with the focal expectation in the cluster $c$. Table F2 displays the estimated MMW for grade-3 classes in each expectation group within each school cluster. For example, in school cluster 1, for negative expectation group, the goal of weighting is to make treated classes in every stratum to have a same probability of having a negative-expectation teacher as that of the whole treated classes in the cluster, i.e. $57/245 = 23.27\%$. Hence I assigned a weight of 2.28 to stratum 1 where the proportion of the treated classes had been underweighted and gave weights smaller than 1 to stratum 2 and 3 where the proportion had been overweighed. After the adjustment, the weighted composition in cluster 1 resembled the composition of the whole pseudo population.

**Table F2**

**Class-level Marginal Mean Weights for the Final Analytic Sample**

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Negative Expectation (T1)</th>
<th>Indifferent Expectation (T2)</th>
<th>Positive Expectation (T3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n(s</td>
<td>c)$</td>
<td>$n(t,s</td>
</tr>
<tr>
<td>School Cluster 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>137</td>
<td>14</td>
<td>2.28</td>
</tr>
<tr>
<td>2</td>
<td>77</td>
<td>22</td>
<td>0.81</td>
</tr>
<tr>
<td>3</td>
<td>31</td>
<td>21</td>
<td>0.34</td>
</tr>
<tr>
<td>4</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>
Students-level MMW

The strategy for computing the student-level weights is similar as that for class-level weights except that within each cluster I broke down students into 5 groups by their prior ability and computed the weights separately for each ability group. Note that once examined within each ability group, some students exposed to a certain expectation did not have any match in the other two expectation groups due to sample variations across the ability levels. For example, in stratum 1 of cluster 1, all lowest ability students \((n = 9)\) was found to be assigned to indifferent or negative expectation, but not to positive expectation; as a result, all the 9 students had to be
excluded from the previous sample. Hence student-level marginal mean weights were computed
based on the adjusted student distribution (see Table F3 for the student distribution within
stratum). Within a cluster \( c \), for students from ability group \( a \), the weights are computed as a
ratio of the total number of ability \( a \) students in a stratum, i.e. \( n_{(s,c|A=a)} \) to the number of the
students at the same ability level who shared a same teacher expectation \( t \) in that stratum, i.e.
\( n_{(t,s,c|A=a)} \). I write it as:

\[
\text{MMW}_{\text{student}} = \frac{n_{(t,s,c|A=a)}}{n_{(t,s,c|A=a)}} \times Pr \{ T = t \mid A = a, C = c \},
\]  

(F.3)

where \( Pr \{ T = t \mid A = a, C = c \} \) is the proportion of students who received teacher expectation \( t \)
among ability group \( a \) within cluster \( c \). Table F3 summarizes all the student-level marginal mean
weights in this study.
Table F3
Student-level Marginal Mean Weights for the Final Analytic Sample

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Ability 1 (lowest ability)</th>
<th>Ability 2</th>
<th>Ability 3</th>
<th>Ability 4</th>
<th>Ability 5 (highest ability)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T1</td>
<td>T2</td>
<td>T3</td>
<td>T1</td>
<td>T2</td>
</tr>
<tr>
<td>Cluster 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.73</td>
<td>2.39</td>
<td>--</td>
<td>1.74</td>
<td>2.04</td>
</tr>
<tr>
<td></td>
<td>(9)</td>
<td>(6)</td>
<td></td>
<td>(14)</td>
<td>(9)</td>
</tr>
<tr>
<td>2</td>
<td>0.82</td>
<td>0.39</td>
<td>1.04</td>
<td>0.84</td>
<td>0.36</td>
</tr>
<tr>
<td>3</td>
<td>0.39</td>
<td>1.11</td>
<td>1.14</td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>(7)</td>
<td>(2)</td>
<td>(33)</td>
<td>(14)</td>
<td>(8)</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.83</td>
<td>0.93</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(34)</td>
<td>(27)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total n</td>
<td>29</td>
<td>22</td>
<td>98</td>
<td>45</td>
<td>25</td>
</tr>
</tbody>
</table>

Cluster 2

<table>
<thead>
<tr>
<th></th>
<th>Ability 1 (lowest ability)</th>
<th>Ability 2</th>
<th>Ability 3</th>
<th>Ability 4</th>
<th>Ability 5 (highest ability)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T1</td>
<td>T2</td>
<td>T3</td>
<td>T1</td>
<td>T2</td>
</tr>
<tr>
<td>1</td>
<td>1.93</td>
<td>1.19</td>
<td>2.21</td>
<td>1.92</td>
<td>1.34</td>
</tr>
<tr>
<td></td>
<td>(9)</td>
<td>(19)</td>
<td>(7)</td>
<td>(7)</td>
<td>(14)</td>
</tr>
<tr>
<td>2</td>
<td>0.66</td>
<td>0.62</td>
<td>1.47</td>
<td>0.84</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>(8)</td>
<td>(14)</td>
<td>(19)</td>
<td>(6)</td>
<td>(6)</td>
</tr>
<tr>
<td>3</td>
<td>0.56</td>
<td>1.29</td>
<td>0.88</td>
<td>0.39</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>(13)</td>
<td>(6)</td>
<td>(14)</td>
<td>(9)</td>
<td>(5)</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.78</td>
<td>0.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(72)</td>
<td>(47)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total n</td>
<td>30</td>
<td>39</td>
<td>112</td>
<td>22</td>
<td>25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Total n</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>29</td>
<td>(29)</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>30</td>
<td>(30)</td>
</tr>
</tbody>
</table>
### Student-level MMW (n)

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Ability 1 (lowest ability)</th>
<th>Ability 2</th>
<th>Ability 3</th>
<th>Ability 4</th>
<th>Ability 5 (highest ability)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T1</td>
<td>T2</td>
<td>T3</td>
<td>T1</td>
<td>T2</td>
</tr>
<tr>
<td>Clusters 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5.24</td>
<td>1.36</td>
<td>3.21</td>
<td>1.48</td>
<td>2.09</td>
</tr>
<tr>
<td></td>
<td>(3)</td>
<td>(14)</td>
<td>(7)</td>
<td>(8)</td>
<td>(7)</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.07</td>
<td>0.66</td>
<td>1.01</td>
<td>1.05</td>
<td>0.57</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td>1.10</td>
<td>0.75</td>
<td>1.13</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>(17)</td>
<td>(9)</td>
<td>(22)</td>
<td>(4)</td>
<td>(9)</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.44</td>
<td>0.29</td>
<td>0.86</td>
<td>0.25</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>(7)</td>
<td>(6)</td>
<td>(73)</td>
<td>(7)</td>
<td>(3)</td>
</tr>
<tr>
<td>Total n</td>
<td>46</td>
<td>34</td>
<td>116</td>
<td>36</td>
<td>25</td>
</tr>
</tbody>
</table>

| Clusters 4 |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 1       |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|         | 12.22| 1.17| 1.45| 1.59| 4.03| 0.99| 1.92| 2.22| 1.26| 2.51| 2.71| 2.15| 1.00| 1.65| 1.09|
|         | (1) | (5) | (2) | (6) | (2) | (5) | (5) | (5) | (6) | (2) | (3) | (3) | (8) | (2) | (2) |
| 2       |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|         | 0.76| 0.79| 0.98| 1.36| 0.49| 1.16| 0.46| 0.53| 1.18| 0.50| 0.53| 1.04| --  | 0.68| 1.04|
| 3       |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|         | 0.47|-- | 0.39| 0.76| 0.39| 0.79| 0.39| 0.79| 0.24| 0.78|-- | 0.93|-- |     |    |
|         | (18)|-- | (7) | (26)| (4) | (33)| (2) | (21)|-- |     |     | (14)|     |     |     |
| Total n | 26  | 9  | 48  | 15  | 14  | 70  | 13  | 18  | 68  | 7   | 14  | 53  | 8   | 6  | 37  |

Note: -- indicates students have been removed due to lack of match.
APPENDIX G
Weighted Multilevel Balance Checking

Although during stratification, I confirmed no statistically significant difference in the mean logit propensity scores between the school/class treatments within each cluster/strata, a global weighted balance checking was still needed. The goal was to examine how effective the computed school- and class-level weights would be in removing bias associated with the observed covariates and whether the weighted composition of each treatment group resembled that of the entire (pseudo) population. Hence I compared treatment groups at each level both in their logit propensity scores and in the distribution of each observed pretreatment covariate. It should be noted that balance checking was not necessary for student-level weights. By treating student prior ability as a moderator, comparison across different ability levels was considered descriptive rather than causal. Since student-level weights were computed based on class-level stratification, it should help achieve the same degree of balance as the class-level MMW.

Balance on Logit of Propensity Scores

With each set of the school-level marginal mean weights, I conducted a single-level weighted regression and made sure that there was no significant difference between the two policy groups in the logit propensity scores with the final sample (Coefficient = .02, SE = .10, t = .16). To examine whether balance was achieved at class level within school clusters C1-C4, for each expectation level $T = t$, I ran a two-level model with the corresponding logit scores as an outcome and applied class-level marginal mean weights at level 1:

Level 1
\[
\text{logit}(T = t)_{jk} = \beta_{0k} + \beta_{1k} * (T2)_{jk} + \beta_{2k} * (T3)_{jk} + r_{jk} ;
\]  

(G.1)
Level 2

\[ \beta_{hk} = \gamma_{h1} * (C1)_k + \gamma_{h2} * (C2)_k + \gamma_{h3} * (C3)_k + \gamma_{h4} * (C4)_k + u_{hk}, \quad h = 0 - 2 \]  

(G.2)

By testing the null hypothesis \( \gamma_{1c} = \gamma_{2c} = 0 \) for \( c = 1-4 \), I compared the three expectation groups in each set of the class-level logit propensity scores. Table G1 shows the results from both unweighted and weighted analysis, from which we can see that the weighting method substantially remove the differences between the class-level treatment groups in the joint distributions of the observed pretreatment characteristics.

Table G1

<table>
<thead>
<tr>
<th>Treatment groups</th>
<th>Coefficient (SE)</th>
<th>logit ((T=1)) Before</th>
<th>After</th>
<th>logit ((T=2)) Before</th>
<th>After</th>
<th>logit ((T=3)) Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>School Cluster 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative Expectation ((T=1))</td>
<td>-1.5 (0.05)</td>
<td>-1.63</td>
<td>-2.07 (0.03)</td>
<td>-2.06 (0.04)</td>
<td>0.8 (0.04)</td>
<td>0.91 (0.04)</td>
<td></td>
</tr>
<tr>
<td>Indifferent Expectation ((T=2))</td>
<td>-1.67 (0.06)</td>
<td>-1.67</td>
<td>-1.93 (0.04)</td>
<td>-1.98 (0.05)</td>
<td>0.88 (0.05)</td>
<td>0.92 (0.04)</td>
<td></td>
</tr>
<tr>
<td>Positive Expectation ((T=3))</td>
<td>-1.7 (0.05)</td>
<td>-1.69</td>
<td>-1.99 (0.03)</td>
<td>-1.97 (0.04)</td>
<td>0.94 (0.04)</td>
<td>0.91 (0.04)</td>
<td></td>
</tr>
<tr>
<td>Hypothesis testing of mean differences (\chi^2)</td>
<td>17.54***</td>
<td>1.70</td>
<td>10.76**</td>
<td>4.37</td>
<td>11.67**</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>School Cluster 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative Expectation ((T=1))</td>
<td>-1.38 (0.06)</td>
<td>-1.46</td>
<td>-1.84 (0.04)</td>
<td>-1.84 (0.04)</td>
<td>0.63 (0.05)</td>
<td>0.70 (0.04)</td>
<td></td>
</tr>
<tr>
<td>Indifferent Expectation ((T=2))</td>
<td>-1.46 (0.09)</td>
<td>-1.46</td>
<td>-1.72 (0.05)</td>
<td>-1.77 (0.05)</td>
<td>0.63 (0.06)</td>
<td>0.65 (0.06)</td>
<td></td>
</tr>
<tr>
<td>Positive Expectation ((T=3))</td>
<td>-1.57 (0.06)</td>
<td>-1.52</td>
<td>-1.86 (0.04)</td>
<td>-1.84 (0.04)</td>
<td>0.77 (0.04)</td>
<td>0.73 (0.04)</td>
<td></td>
</tr>
<tr>
<td>Hypothesis testing of mean differences (\chi^2)</td>
<td>13.83**</td>
<td>2.81</td>
<td>12.02**</td>
<td>3.23</td>
<td>18.38***</td>
<td>3.40</td>
<td></td>
</tr>
<tr>
<td>School Cluster 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment groups</td>
<td>Coefficient (SE)</td>
<td>logit (T=1)</td>
<td>logit (T=2)</td>
<td>logit (T=3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------</td>
<td>-----------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Before</td>
<td>After</td>
<td>Before</td>
<td>After</td>
<td>Before</td>
<td>After</td>
</tr>
<tr>
<td>Negative Expectation (T=1)</td>
<td>-1.53 (0.04)</td>
<td>-1.79 (0.04)</td>
<td>-1.80 (0.05)</td>
<td>0.67 (0.04)</td>
<td>0.75 (0.05)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indifferent Expectation (T=2)</td>
<td>-1.64 (0.03)</td>
<td>-1.64 (0.07)</td>
<td>-1.73 (0.06)</td>
<td>0.67 (0.05)</td>
<td>0.71 (0.05)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive Expectation (T=3)</td>
<td>-1.68 (0.03)</td>
<td>-1.81 (0.04)</td>
<td>-1.76 (0.04)</td>
<td>0.80 (0.04)</td>
<td>0.74 (0.04)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothesis testing of mean</td>
<td>21.03***</td>
<td>2.45</td>
<td>8.34*</td>
<td>15.67***</td>
<td>0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>differences (χ²)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

School Cluster 4

<table>
<thead>
<tr>
<th>Treatment groups</th>
<th>Coefficient (SE)</th>
<th>logit (T=1)</th>
<th>logit (T=2)</th>
<th>logit (T=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Before</td>
<td>After</td>
<td>Before</td>
</tr>
<tr>
<td>Negative Expectation (T=1)</td>
<td>-1.73 (0.04)</td>
<td>-2.15 (0.09)</td>
<td>-2.06 (0.10)</td>
<td>1.00 (0.05)</td>
</tr>
<tr>
<td>Indifferent Expectation (T=2)</td>
<td>-1.85 (0.03)</td>
<td>-1.84 (0.08)</td>
<td>-1.97 (0.09)</td>
<td>0.90 (0.04)</td>
</tr>
<tr>
<td>Positive Expectation (T=3)</td>
<td>-1.83 (0.03)</td>
<td>-2.07 (0.07)</td>
<td>-2.03 (0.07)</td>
<td>1.00 (0.03)</td>
</tr>
<tr>
<td>Hypothesis testing of mean</td>
<td>18.59***</td>
<td>2.25</td>
<td>13.83**</td>
<td>7.74*</td>
</tr>
<tr>
<td>differences (χ²)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: * p<.05 **p<.01 ***p<.001

**Balance on Individual Covariate**

At school level, I performed a weighted logistic regression for every school-level pretreatment variables \(V\) with treatment \(Z\) as an outcome i.e. \(\log Z = \beta_0 + \beta_1 V\). Balance was found to be achieved on 91% of the 81 school-level covariates.

To check balance between the three teacher expectations under each school policy, by simultaneously applying the two-level weights, I analyzed a two-level weighted multinomial logistic model with teacher expectation \(T\) as the outcome, and school policy \(Z\) and each covariate as predictors. For a class-level covariate \(W\), I specified the weighted multilevel model as:
Level 1
\[ \eta_{mjk} = \log\left(\frac{\phi_{mjk}}{\phi_{\gamma,jk}}\right) = \beta_{0k(m)} + \beta_{1k(m)} * W_jk, \text{ where } \varphi_{mjk} = \Pr(T_{jk} = m) \text{ for } m = 1 \text{ or } 2 \]  

(G.3)

Level 2
\[ \beta_{0k(m)} = \gamma_{00(m)} + \gamma_{01(m)} * Z_k + u_{0k(m)}, \]
\[ \beta_{1k(m)} = \gamma_{10(m)} + \gamma_{11(m)} * Z_k + u_{1k(m)}, \quad u_k \sim N(0, \tau_\beta) \]

(G.4)

If balance was achieved on a class-level covariate \( W \) between different levels of \( T \) under each policy condition, after weighting it is expected that \( \gamma_{10(1)} = \gamma_{10(2)} = 0 \) and \( \gamma_{11(1)} = \gamma_{11(2)} = 0 \). For a school-level covariate \( V \), I specified the balance-checking model as

Level 1
\[ \eta_{mjk} = \log\left(\frac{\phi_{mjk}}{\phi_{\gamma,jk}}\right) = \beta_{0k(m)}, \text{ where } \varphi_{mjk} = \Pr(T_{jk} = m) \text{ for } m = 1 \text{ or } 2, \]  

(G.5)

Level 2
\[ \beta_{0k(m)} = \gamma_{00(m)} + \gamma_{01(m)} * Z_k + \gamma_{02(m)} * V_k + \gamma_{03(m)} * Z_k * V_k + u_{0k(m)}, \]
\[ u_k \sim N(0, \tau_\beta) \]

(G.6)

It is expected that \( \gamma_{02(1)} = \gamma_{02(2)} = 0 \) and \( \gamma_{03(1)} = \gamma_{03(2)} = 0 \). Conducting the two hypothesis tests for every school- or class-level covariate (\( n=112 \)), I found that the three expectation groups resembled each other in approximately 94% of the 112 observed school-and class-covariates under the standardized testing only and in about 91% of the covariates under the test-based retention policy.
APPENDIX H

List of Prognostic Variables in the Final Outcome Models

Table H1

Prognostic Variables in the Final Outcome Models

<table>
<thead>
<tr>
<th>Prognostic Variables</th>
<th>Phase I (relationship between the treatments)</th>
<th>Phase II (effects on instructional time allocation)</th>
<th>Phase III (effects on student academic performance)</th>
<th>Phase VI (effects on student self-perception)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>reading</td>
<td>math</td>
<td>science</td>
<td>Multivariate Model</td>
</tr>
<tr>
<td>Female</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Home language as English (missing indicator)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>White*</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Hispanics*</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Asian*</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Black*</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Other ethnicity*</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Age at K entry</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Spring 2002 continuous SES measure</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Reading score at K entry*</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Reading score at K entry (missing indicator)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Math score at K entry*</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Prognostic Variables</td>
<td>Phase I (relationship between the treatments)</td>
<td>Phase II (effects on instructional time allocation)</td>
<td>Phase III (effects on student academic performance)</td>
<td>Phase VI (effects on student self-perception)</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------------------------------------------</td>
<td>---------------------------------------------------</td>
<td>--------------------------------------------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>reading</td>
<td>math</td>
<td>science</td>
<td>Multivariate Model</td>
</tr>
<tr>
<td>Math score at K entry (missing indicator)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>General knowledge score at K entry*</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>General knowledge score at K entry (missing indicator)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Fall 1998 teacher rating of student approaches to learning</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Fall 1998 teacher rating of student self-control</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Fall 1998 teacher rating of student externalizing problem behaviors</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

**Class Level**

<p>| Spring 2002 class aggregated measure of current students' fall 1998 reading scores | √ |
| Spring 2002 class aggregated measure of current students' approaches to learning rated by teachers in fall 1998 | √ |
| Spring 2002 % black students in class | √ | √ | √ |</p>
<table>
<thead>
<tr>
<th>Prognostic Variables</th>
<th>Phase I (relationship between the treatments)</th>
<th>Phase II (effects on instructional time allocation)</th>
<th>Phase III (effects on student academic performance)</th>
<th>Phase VI (effects on student self-perception)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring 2002 % eligible students for free or reduced-price breakfast</td>
<td>√</td>
<td></td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Spring 2002 % eligible students for free or reduced-price lunch in class</td>
<td></td>
<td>√</td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>Spring 2002 class having no information about whether students speak non-English language</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School Level</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spring 2002 school in MidWest region</td>
<td></td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spring 2002 school in West region</td>
<td></td>
<td>√</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Spring 2002 public school</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spring 2002 school of choice</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spring 2002 % LEP students</td>
<td></td>
<td>√</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Spring 2002 school having no information about USDA breakfast program</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prognostic Variables</td>
<td>Phase I (relationship between the treatments)</td>
<td>Phase II (effects on instructional time allocation)</td>
<td>Phase III (effects on student academic performance)</td>
<td>Phase VI (effects on student self-perception)</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------</td>
<td>----------------------------------------------------</td>
<td>----------------------------------------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>Spring 2002 % students eligible for free or reduced lunch in school</td>
<td>√</td>
<td></td>
<td>√ √ √ √ √ √ √ √ √</td>
<td>√ √ √ √ √ √ √ √ √</td>
</tr>
<tr>
<td>Spring 2002 school having service for students with disability</td>
<td></td>
<td>√ √ √ √ √ √ √ √ √</td>
<td></td>
<td>√ √ √ √ √ √ √ √ √</td>
</tr>
<tr>
<td>Spring 2002 number of safety measures</td>
<td>√</td>
<td></td>
<td></td>
<td>√ √ √ √ √ √ √ √ √</td>
</tr>
<tr>
<td>Spring 2002 school aggregated measure of current students’ fall 1998 math scores</td>
<td>√ √ √ √ √ √ √ √ √</td>
<td></td>
<td>√ √ √ √ √ √ √ √ √</td>
<td>√ √ √ √ √ √ √ √ √</td>
</tr>
<tr>
<td>Spring 2002 school aggregated measure of current students’ fall 1998 general knowledge scores</td>
<td></td>
<td></td>
<td></td>
<td>√ √ √ √ √ √ √ √ √</td>
</tr>
<tr>
<td>Spring 2002 school aggregated measure of current students’ approaches to learning rated by teachers in fall 1998</td>
<td></td>
<td></td>
<td></td>
<td>√ √ √ √ √ √ √ √ √</td>
</tr>
<tr>
<td>Spring 2002 school aggregated measure of current students’ self-control rated by teachers in fall 1998</td>
<td></td>
<td></td>
<td></td>
<td>√ √ √ √ √ √ √ √ √</td>
</tr>
</tbody>
</table>
Prognostic Variables  | Phase I (relationship between the treatments) | Phase II (effects on instructional time allocation) | Phase III (effects on student academic performance) | Phase VI (effects on student self-perception) |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring 2002 school aggregated measure of current students’ externalized problem behavior rated by teachers in fall 1998</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Spring 2002 school aggregated measure of number of students in grade-3 classes</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spring 2002 school aggregated measure of % students below 9 years old in grade-3 classes</td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
</tr>
</tbody>
</table>

Note: Prognostics entered in the final outcome models were selected through systematic model comparisons. All class- and school-level prognostics were not balanced by school- or class-level propensity model(s). Relationship between school- or class-level prognostics and outcome(s) was found mostly linear; hence I did not add in any nonlinear or interaction terms in the final outcome models. For student-level prognostics, I assumed that treatment assignment is independent of individual student pretreatment characteristics given the observed school- and class-level confounders and thus did not add in any nonlinear terms either. For simplicity, prognostic variables used for estimating long-term effect were not specified here; however they were as well selected through model comparison from the variables listed in the table.

* missing value was coded as zero.  -- variable not applicable to the analysis.
REFERENCES


Coburn, C. E. (2006). Framing the problem of reading instruction: Using frame analysis to


Diamond, J. B., Randolph, A., & Spillane, J. P. (2004). Teachers' expectations and sense of
responsibility for student learning: The importance of race, class, and organizational habitus. *Anthropology & Education, 35*(1), 75-98.


Hong, G., Pelletier, J., Hong, Y., & Corter, C. (under review). Does literacy instruction affect kindergartners' externalizing problem behaviors as well as their literacy learning? Taking class manageability into account.


Koretz, D., Barron, S., Mitchell, K., & Stecher, B. (1996). The perceived effects of the Kentucky instructional results information system (KIRIS) (No. MR-792-PCT/FF). Santa Monica,
CA: RAND.


affect students, their parents, teachers, principals, schools, and society. Charlotte, N.C.: Information Age Publishing.


Nicholls, J. G. (1978). The development of the concepts of effort and ability, perception of academic attainment, and understanding that difficult tasks require more ability Child Development, 49, 800-814.


Rubie-Davies, C. M. (2007). Classroom interactions: Exploring the practices of high- and low-


Wheelock, A. (2003). *School awards programs and accountability in Massachusetts*. Cambridge,
MA: Fair Test.