THREE ESSAYS ON ESTIMATING THE EFFECTS OF SCHOOL AND STUDENT IMPROVEMENT INTERVENTIONS

By

Guan Saw

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Measurement and Quantitative Methods – Doctor of Philosophy

2016

ABSTRACT

THREE ESSAYS ON ESTIMATING THE EFFECTS OF SCHOOL AND STUDENT IMPROVEMENT INTERVENTIONS

By

Guan Saw

This dissertation consists of three chapters that examine the effects of school and students improvement interventions. The first chapter investigates whether, for whom, and under which conditions high school mathematics and science course graduation requirements (CGRs) affect student achievement and educational attainment. Drawing on data from the High School Longitudinal Study of 2009 (HSLS:09), fixed effects results show that higher math CGRs have a positive, but modest, impact on student test scores while no impact on college enrollment. Suggestive evidence indicates that higher science CGRs may have a negative, unintended effect on on-time postsecondary attendance. The positive effect of higher math CGRs is largely concentrated among students who are in the lowest and highest end of the math ability distribution, whereas the negative effect of higher science CGRs is primarily driven by students from high-middle socioeconomic families. The effects of CGRs appear to be moderated by institutional contexts in which schools with greater academic and social organizations have the strongest positive impacts.

The second chapter evaluates whether postsecondary remediation influences college persistence, transfer, and attainment, and if effects vary by racial and socioeconomic subgroups. Using data from the National Longitudinal Survey of Youth of 1997 (NLSY97), propensity score analysis results indicate that while remediation in only mathematics or only English has no impact on student outcomes, the effect of remediation in both subjects is positive for students who started postsecondary education in two-year colleges but it is negative for their four-year

college counterparts. Sensitivity tests show that the estimates for four-year colleges are quite robust but they are less so for two-year colleges. Subgroup analyses reveal that in two-year colleges high-socioeconomic students benefited the most from remediation in the long run, whereas in four-year colleges remediation appears to hinder nonwhite and low-socioeconomic students from completing college. Findings suggest that postsecondary remediation plays a critical role in the social stratification process in higher education.

The third chapter, co-authored with Barbara Schneider, Ken Frank, I-Chien Chen, Venessa Keesler, and Joseph Martineau, explores the differential effects of "consequential labeling" versus "non-consequential labeling" in the context of school accountability. Since the No Child Left Behind Act was enacted, grading and labeling low-performing schools has been increasingly used as a means to incentivize failing schools to raise student achievement. Using state-wide high school data from Michigan, our regression discontinuity analyses show that the bottom 5% schools identified as Persistently Lowest Achieving (PLA), which was publicly announced and has imminently threatening accountability, increased their student performance in writing and to a lesser extent in mathematics and social studies. The PLA effect in writing is quite robust, based on various sensitivity analyses. We find no improvement in student achievement for those bottom 6-20% schools labeled as "watch list" that received no actual penalties and little public attention. Our findings suggest that schools respond differently to varying forms of low-performing labeling, depending on the accountability pressure and social stigmatization process.

To Ah Geik Ooi.

ACKNOWLEDGMENTS

First and foremost I truly thank my advisor, Barbara Schneider, who has been very supportive and helpful since the days I began my graduate studies at Michigan State University. I am also extremely grateful to Ken Frank, who has consistently provided guidance and encouragement throughout my research career. I also am indebted to Spyros Konstantopoulos and Joshua Cowen for generously sharing their insightful perspectives on my research work and career development. I have been blessed to have met Li-Yun Wang, Samuel Peng, Tony Tam, and Ping-Yin Kuan while working on my master's degree at National Taiwan Normal University. They inspired and encouraged me to pursue an academic career.

I also would like to thank many colleagues and friends for their invaluable help, support, and encouragement. There are too many to count but I want to mention Tom Almer, Michael Broda, Justin Bruner, Kri Burkander, Jason Burns, Hsun-Yu Chan, Chi Chang, Yin Hong Cheah, I-Chien Chen, Wei-Lin Chen, Yi-Ling Cheng, Michelle Chester, Richie Chester, Meng Han Chin, Ben Creed, Christina Ebmeyer, Jeremy Herliczek, Yu-Han Hung, Enyi Jen, Venessa Keesler, Soobin Kim, Christopher Klager, John Lane, Wei Ching Lee, Cheng-Hsien Li, Kee Choi Lim, Chi-Jui Lu, Joseph Martineau, Elizabeth Covay Minor, Justina Spicer, Anne Traynor, Roxane Weng, Min-Lun Wu, Ran Xu, Danny Yang, Chia-Hsin Yeh, Hsueh-Han Yeh, and Yisu Zhou.

I acknowledge all of the support I have received from my family, including my father, Soon Ho, and my siblings, Guan Ying, Guan Hee, and Ay Lee. Finally, I would like to thank and dedicate this dissertation to my late mother, Ah Geik Ooi, who instilled in me the virtues of hard work, perseverance, courage, and ambition, which have tremendously helped me during my doctoral studies and as a researcher.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

THE IMPACT OF HIGH SCHOOL MATHEMATICS AND SCIENCE COURSE GRADUATIONREQUIREMENTS: SCHOOL STRUCTURAL, ACADEMIC, AND SOCIAL ORGANIZATIONAL FACTORS[1]

1.1 Introduction

Every year about half of high school seniors graduate without the minimal requirements needed to apply to a four-year college (ACT Inc., 2010; Greene & Foster, 2003). Graduates who are low-income, black, and Hispanic are particularly less likely to be academically prepared for postsecondary education (Adelman, 2004; Long, Iatarola, & Conger, 2009). To improve student college readiness while ensuring the opportunity to learn for all, state lawmakers and school leaders have been working to increase high school course graduation requirements (CGRs), especially in mathematics and science (Smerdon & Borman, 2012; The Center for Public Education, & Change the Equation, 2013; see Figure 1.A1 for the increasing trends in high school math and science CGRs in 50 states and District of Columbia from 1980 to 2013). The existing evidence on the effects of high school CGRs, however, is scarce due to the methodological challenges such as lack of reliable and consistent data, and isolating the CGR impact from potential confounding factors with non-experimental data. Few studies have attempted to address these challenges, yet the findings are inconsistent (e.g., Jacob, Dynarski, Frank, Schneider, 2016; Lillard & DeCicca, 2001; Plunk, Tate, Bierut, & Grucza, 2014).

[1] This research was supported by a grant from the American Educational Research Association (AERA) which receives funds for its "AERA Grants Program" from the National Science Foundation (NSF) under Grant #DRL-0941014. Opinions reflect those of the author and do not necessarily reflect those of the granting agencies.

In the limited literature on high school CGRs, there has been little theoretical attention to or empirical research on the school organizations that may influence the CGR effects in schools. This has resulted in part from the fact that policymakers and evaluators tend to be interested in the average treatment effects of high school CGRs while overlooking the potential moderating effects of school context (e.g., teacher composition, school climate). A growing body of sociological literature has demonstrated that school organizational factors play a crucial role in facilitating the implementation of school reform and policy (Hedges & Schneider, 2005; Frank, Zhao, & Borman, 2004; Spillane & Louis, 2005). Thus, examining whether and how schools differ in the influences of high school CGRs on student outcomes can enhance our understanding of how CGR effects, if any, are generated and what interventions needed to be made to improve the policy and practice.

This study contributes to research and policy discussions on high school course graduation requirements in several respects. First, building upon the literature on school effects and social stratification, this study formulates a theoretical framework and develops a series of hypotheses to test the main and differential effects of high school CGRs, which help interpret the seemingly mixed findings in prior studies. Second, in contrast to most previous studies that estimated the effects of CGRs by subject at the district or state level, this study classifies schools into different comparison and treatment groups, which consider the numbers of years of coursework required in both math and science. Defining the treatment of CGRs at the school level allows for examination of differential effects by school context. Finally, this study empirically tests the hypotheses using a nationally representative sample of recent high school students, which yields greater generalizability. Findings suggest that higher CGRs in math and

science have both intended positive and unintended negative consequences on student outcomes and the effects vary by student and school characteristics.

1.2 Theoretical Framework

The effort to hold all high school students to high academic standards has a century old history in American education (Angus & Mirel, 1999). Raising high school CGRs is a recent move in this long tradition of curriculum intensification and standardization. The primary policy rationale for raising CGRs is that all high school students need to have the necessary academic knowledge and skills to succeed either in higher education or in the workplace (US Department of Education, 2004). The policy or practice of high school CGRs, however, is not without controversy. The concern focuses on two issues in relation to educational productivity and equality. The first is whether CGRs have a positive or negative impact on student outcomes. The second is whether CGRs reduce or increase gaps of student outcomes.

The debate on CGRs is fed by assumptions about how graduation requirements influence student coursetaking, which in turn affects achievement and attainment (Chaney, Burgdorf, & Atash, 1997). Citing studies on coursetaking (e.g., Adelman, 2006; Attewell & Domina, 2008; Schneider, Swanson, Catherine, 1998), proponents of high school CGRs argue that students who complete more advanced academic courses have greater achievement gains, and are more likely to graduate from high school and attend a postsecondary institution. The positive correlations between coursetaking and student outcomes, however, may simply reflect self-selection effects. It could be the case that highly-motivated students are more likely to take more advanced course sequences. Thus it is unclear whether pressing all students, particularly those who are academically disadvantaged, toward high academic standards through such policy as high school CGRs would increase student performance.

To some extent, raising high school CGRs for all can be seen as a form of detracking because all students will be exposed to challenging academic courses providing them with more rigorous opportunities to learn (Sorensen, 1970; Domina & Saldana, 2012). In theory, therefore, intensified CGRs could improve overall student performance while narrowing the achievement gap, especially when those traditionally underserved students (i.e., low-income, minority, and low-performing students) who would have been excluded from academic-oriented programs are now taking more advanced classes and potentially attending classrooms with higher skilled peers (Adelman, 1999; Gamoran, 1996; Lee & Smith, 2001). Yet, this conjecture has not been supported by empirical studies. Thus far, no direct evidence shows that systems sorting students into academic tracks have been reduced as a result of high school CGRs (Heck, Price, & Thomas, 2004; Wilson & Rossman, 1993). Additionally, there are concerns that high school CGRs may have unanticipated, negative consequences. For example, if disadvantaged students were forced out of school by such graduation requirements or difficult academic classes, the principle of equal educational opportunity is violated (Mirel & Angus, 1994).

With these considerations and given the policy goals of high school CGRs, this study empirically tests the following hypotheses:

H1a: High school CGRs boost the average achievement and educational attainment of all students (*educational productivity hypothesis*).

H1b: High school CGRs narrow the achievement and educational attainment gaps between racially, socioeconomically, and academically disadvantaged students and their counterparts (*educational equality hypothesis*).

Raising high school CGRs is a commonly used policy instrument to impose certain academic expectations on student coursetaking and performance. Whether institutionalized

expectations are sufficient to improve student outcomes is a research question that has been the focus of many past studies (e.g., Chaney et al., 1997; Jacob et al., 2016; Teitelbaum, 2003). This study extends beyond prior work by examining whether high school CGRs interact with the social structure and activities in schools to produce different outcomes. Drawing upon research literature on school effects and social stratification, a set of hypotheses is developed to test the moderating effects of school context on the relationship between high school CGRs and student outcomes. The concepts of opportunity to learn (OTL) and social organization of schools are used to conceptualize how high school CGRs can be functioning differently at school level and to account for variations in student performance vis-à-vis differences in school structural and organizational factors (Bidwell, 1965; Bryk & Driscoll, 1988; Hedges & Schneider, 2005; McDonnell, 1995; McPartland & Schnieder, 1996; Schmidt & Maier, 2009). Three distinct but overlapping school-based dimensions of OTL are examined: school structure, academic organization, and social organization (see Figure 1.A2 for the hypothesized conceptual framework of school moderating effects).

In the conventional literature, school structural measures that are related to student's OTL include school sector (Bryk, Lee, & Holland, 1993; Coleman, Hoffer, & Kilgore, 1982), urbanicity (Roscigno & Crowley, 2001; Roscigno, Tomaskovic-Devey, & Crowley, 2006), and enrollment size (Lee & Smith, 1995, 1997). Student composition is another key school structural feature. A large body of research has demonstrated that students learn more when surrounded by peers who are academically, socially, or economically advantaged (Coleman et al., 1966; Hanushek, Kain, Markman, & Rivkin, 2003; Muller, Riegle-Crumb, Schiller, Wilkinson, & Frank, 2010; Southworth & Mickelson, 2007). Therefore, it is hypothesized that peer composition in school may moderate the impact of high school CGRs.

H2a: High school CGRs affect student outcomes more positively in schools with higher concentration of advantaged peers (*school structure hypothesis*)

The academic organization is a prime dimension of OTL and to this study. One typical form of school academic organization is curriculum structure which can be represented by components such as course offerings (Gamoran, 1987) and content standards or coverage (Schmidt et al., 2001). Prior research has shown that students perform better in schools with a more constrained curriculum composed mostly of academic courses (Bryk et al., 1993; Lee, Croninger, & Smith, 1997). Apart from curriculum structure, many scholars argue that teachers are the key agents in implementing curricular policies or reforms (Schwille et al., 1983; Spillane & Louis, 2005). This study thus adds instructional capacity as an important feature of academic organization, focusing on teachers' professional qualifications and knowledge that have been documented to be associated with student learning gains (Goldhaber & Brewer, 2000; Hill, Rowan, & Ball, 2005). Therefore, the following hypotheses are proposed:

H2b: High school CGRs affect student outcomes more positively in schools with greater academic/instructional capacity (*academic organization hypothesis*)

Beyond structural and academic contexts, social organizations of schools can also play a crucial role in facilitating school reform and improving student achievement (Bryk & Schneider, 2005; Hedges & Schneider, 2005; Frank et al., 2004). Social organizations of schools can be characterized by measuring social interactions/relationships and shared norms/beliefs within school. Several social organizational factors in schools that are likely to interact with the CGRs effects are academic press as measured by teachers' expectation for student learning (Lee & Smith, 1999; Muller, 1998), school climate (Lee & Bryk, 1989; Thapa, Cohen, Guffey, & Higgins-D'Alessandro, 2013), and student engagement (Finn & Zimmer, 2012; Newmann, 1992).

H2c: High School CGRs affect student outcomes more positively in schools with stronger academic norms/climate (*social organization hypothesis*).

1.3 Prior Findings on the Impact of High School CGRs

Lawmakers and school leaders who initiate policy to raise high school CGRs must expect to see improvement in student academic performance and educational attainment. Yet, there is little empirical evidence supporting this policy rationale. More disturbing, some unintended, negative consequences of high school CGRs on student schooling careers have been revealed in recent research. At present, there are only five causal studies that isolate the impact of high school CGRs from the potentially confounding effects of other factors. One early study by Lillard and DeCicca (2001) found that increased total CGRs (all academic subjects) led to higher dropout rates among high school students. The conclusion is established based on both nationally representative individual data on dropout decisions and state-level aggregate data on dropout/attrition rates over fifteen years. Three recent studies analyzing data at the district, state, and national level also found similar results, showing that intensified CGRs, particularly in math and science, lowered the high school graduation rates (Jacob et al., 2016; Montgomery & Allensworth, 2010; Plunk et al., 2014).

The empirical evidence in regard to the CGR effects on student achievement is not encouraging either. Montgomery and Allensworth (2010), for example, found that science CGRs mandated in the Chicago Public Schools in 1997, which required high school students to take three years of science coursework, had null effect on student science grades. Their analysis suggested that those graduates who completed the required science courses did improve in science learning but the overall effect of the policy is offset by the relatively higher dropout rates among students from the same cohort. Buddin and Croft (2014), following nine cohorts of 11th-

graders (from 2005 to 2013) in all Illinois school districts other than Chicago, found that the 2005 state-mandated math and science CGRs were not associated with student ACT scores in the two subjects. Examining a set of statewide college-bound CGRs enacted in Michigan in 2007, Jacob et al. (2016) concluded that higher CGRs generally have little impact on student test scores, except in science and for lowest achieving students. [2]

With respect to post-secondary outcomes, the empirical findings on the CGR effects are mostly mixed. The study by Buddin and Croft (2014) indicated that while higher math CGRs had a positive effect on college attendance rates, higher science CGRs did not. Montgomery and Allensworth (2010) also presented evidence showing that increased science CGRs was not associated with college-going rates. Furthermore, analyzing a decade of data from the US Census Bureau, Plunk et al. (2014) reported that higher math and science CGRs had no overall impact on both college enrollment and completion, but there are some differential effects across racial subgroups. For Black women and Hispanic men and women, it appears that higher math and science CGRs were negatively associated with the likelihood of college attendance. However, conditioned on college enrollment, intensified math and science CGRs in high school had a positive effect on obtaining a college degree for Black women and Hispanic men and women.

Mixed findings for the CGR effects on student outcomes may be well understood from theoretical and methodological perspectives. Current research efforts tended to assess the effects by defining or identifying high school CGRs at the district or state level, which has limited number of treatment and comparison groups that may not yield sufficient statistical power to detect an effect. Moreover, prior studies have focused almost exclusively on *whether* high school CGRs have an impact on student outcomes. None of them examined the *conditions under which*

[2] Two early correlational studies by Chaney et al. (1997) and Teitelbaum (2003), based on national datasets, also found that math and science CGRs had no effects on student test scores.

high school CGRs might be effective or ineffective in influencing student achievement and schooling careers. While researchers recognize the potential differential effects of high school CGRs, they explored it only by student subgroups (e.g., academic ability, race/ethnicity, or family background) but not by school structural and organizational characteristics, which may be moderating the influences of CGRs. School contextual factors, which are overlooked in prior research, may help explain the inconsistent findings of CGR effects in the literature.

1.4 Methodological Approach

1.4.1 Data and Sample

This study analyzes the restricted data from the High School Longitudinal Study of 2009 (HSLS:09), which is specifically designed to explore policy-relevant issues with respect to school contextual factors that may affect student educational trajectories and outcomes. HSLS:09 is the most recent national high school longitudinal study that can provide the information on high school course graduation requirements in several academic subjects (i.e., math, science, English, social studies, and foreign language) at the school level. Another valuable benefit offered by HSLS:09 is that the rich survey data on students, teachers, and schools allow for constructing a comprehensive set of school contextual factors, which are crucial for determining whether and how schools differ in high school CGRs effects.

The HSLS:09 sample is generated by a two-stage sampling design with schools selected first the students from the target population, ninth graders in the fall semester of the 2009-2010 school year, within those schools selected second. In this study, a number of sample restrictions are imposed on the data. First, the analytic sample is limited to those respondents who participated in both 2009 base-year and 2012 follow-up survey and with valid information on their high school CGRs. Second, this study focuses only on first-time ninth graders, thus

excluding those students who were repeating ninth grade in fall 2009. Third, the analytic sample excludes non-regular high schools, including special education, technical/alternative, and special focus schools (e.g., math or science, arts/talented, gifted education). The remaining sample includes 16,081 students from 843 schools.

1.4.2 Measures

This study focuses on evaluating the impact of high school math and science CGRs. As shown in Figure 1.A1, the majority of states have raised their high school CGRs to at least three years of math and science. Given the importance of college readiness and science, technology, engineering, and mathematics (STEM) education, the current policy debates on CGRs are primarily focusing on whether to increase math and science CGRs from three years to four years. Thus, there are four potential treatment/comparison groups of schools with varying level of CGRs in math and science that are of interest to this study, including schools with CGRs of: (a) three years of math and science (3M3S), (b) three years of math and four years of science (3M4S), (c) four years of math and three years of science (4M3S), and (d) four years of math and science (4M4S). The 3M3S schools have CGRs that are equal to the "new basics" standard, a set of college readiness curriculum, which was recommended by the report *A Nation at Risk*, released more than three decades ago (U.S. Department of Education, 1983). The other three groups of schools (i.e., 3M4S, 4M3S, and 4M4S) have a higher standard of math and/or science CGRs compared to the "new basics" standard.

Two HSLS:09 survey items responded to by school administrators about "how many years of coursework in [math/science] are required to meet high school graduation requirements?" are used to identify the treatment and comparison schools. A majority of schools surveyed by the HSLS:09 can be classified into the four potential treatment/comparison groups: (a) 3M3S (257

schools), (b) 3M4S (2 schools), (c) 4M3S (252 schools), and (d) 4M4S (186 schools). The remaining 146 schools, "all other patterns," have math and science CGRs that are fewer than three years in both or at least one of the two subjects. The "all other patterns" group consists of a very heterogeneous set of schools with various combination of CGRs in math and science. The impact estimates on the high school CGRs for this "all other patterns" group will be difficult to interpret and have little information to offer for policy and practice. Therefore, these "all other patterns" schools, together with the only two schools in 3M4S group, are excluded in the following analysis (deleting 17.7% of students and 17.8% of schools). The final analytic sample includes 13,240 students from 695 schools.

Advocates of increasing high school CGRs believe that pushing students to take and complete more academic courses will improve their academic ability and will well prepare them to transition to and to succeed in postsecondary education. To empirically test this proposition, this study uses both measures on cognitive skills and college enrollment as student outcomes. The cognitive outcome measure is a continuous variable that is constructed based on a standardized assessment test in math for HSLS:09 respondents in the spring of 2012. The postsecondary outcome is a measure of college enrollment status as of November 2013 (roughly after four years since students entered ninth grade). It is a measure of on-time college attendance status, which is available in the most recently released data from the HSLS:09. The college-going measure is coded as a binary variable indicating whether a student attended a four-year college.

Guided by the theoretical framework and hypotheses regarding moderating effects, this study created several school level variables measuring school structural, academic, and social organizational factors. The first set of school moderators on structural characteristics includes: (a)

percentage of poverty students (who received free/reduced lunch), (b) percentage of minority students (who are non-white), and (c) percentage of Advanced Placement (AP) students. The second set of school moderators on academic organizations includes: (a) requiring a math competency test, (b) offering AP courses, (c) math/science full time teacher ratio in school (computed using measures on number of full-time teachers, and number of full-time math and science teachers). The third set of school moderators on social organizations includes: (a) academic press (a composite score of teacher, counselor, and principal expectations as perceived by counselor; three set of seven survey items), (b) school climate (a construct combined 14 survey items such as frequency of physical conflicts in school and student in-class misbehavior, responded by school administrator), and (c) student engagement (a construct combined 14 survey items such as student tardiness and absenteeism, responded to by school administrator).

Students who attend schools with higher CGRs are likely to be systematically different from their counterparts who go to schools with lower CGRs. Similarly, schools that have higher CGRs may systematically differ from those that have lower CGRs. To effectively account for the differences in student and school baseline characteristics when estimating the effects of CGRs, this study includes a series of individual and institutional covariates that are collected in the base-year surveys (see Appendix Table 1.C1 for the descriptive statistics of all covariates). Student covariates include demographics (e.g., gender, race/ethnicity), family background (e.g., socioeconomic status [SES], family structure), initial academic ability (i.e., 9th grade math standardized test score), pre-high school academic preparation (e.g., most advanced math course taken in 8th grade), educational and occupational aspirations (e.g., STEM career), subjective measures on schooling experience (e.g., academic commitment, sense of school belonging) and on math and science (e.g., math self-efficacy, science identity). School covariates are school

sector, geographic region, urbanicity, school type, Adequate Yearly Progress (AYP) status, student body (e.g., percent of English language learners, percent of special education students), teacher composition (e.g., percent of full time teachers), academic and special programs (e.g., having a General Educational Diploma [GED] test preparation program, offering dropout prevention programs), and college-going counseling activities (e.g., organizing student visits to colleges, assisting students with finding financial aid for college).

Table 1.B1 presents descriptive statistics of key analysis variables by high school math and science CGRs as categorized into three major school groups (i.e., 3M3S, 4M3S, and 4M4S). The HSLS:09 data provides new statistics on the distribution of students by high school math and science CGRs of their schools. In fall 2009, about three-quarters of first-time 9th graders (weighted estimates) were attending high schools with math and science CGRs that were equivalent to or higher than "new basics" standard (i.e., three years of math and science, 3M3S). Of these students, 21.5% of them were attending 4M4S schools, 23.7% were in 4M3S schools, and 30.7% were in 3M3S schools. About one-quarter of students (24.1%) were in "all other patterns" schools that had lower math and science CGRs than "new basics." Students who are black, Hispanic, low-socioeconomic, and low-achieving tend to attend schools with higher CGRs. In terms of school structural, academic, and social organizational factors, on average there are little differences across the three school groups analyzed in this study.

1.4.3 Analytic Strategy

To test the productivity hypothesis of CGRs (H1a), this study first estimates the overall impact of high school math and science CGRs on student achievement and educational attainment by employing ordinary least squares (OLS) regression models that take the following

form:

$$Y_{ict} = \beta_0 + \beta_1 4M3S_{ict_0} + \beta_2 4M4S_{ict_0} + \boldsymbol{STU_{ict_0}\Gamma_1'} + \boldsymbol{SCH_{ct_0}\Gamma_2'} + \mu_s + \varepsilon_{ict} \quad (1)$$

where Y_{ist} is the student outcomes (i.e., 11th grade math score, attending a four-year college) for student *i* in school *c*. $4M3S_{ict_0}$ and $4M4S_{ict_0}$ are dummy variables denoting whether a student attended a school with CGRs of (a) four years of math and three years of science, and (b) four years of math and science, in the fall of 2009 (t_0; the omitted school group is 3M3S). $\boldsymbol{STU_{ict_0}}$ is a vector of student characteristics for student *i* in school *c* as measured in time t_0, whereas $\boldsymbol{SCH_{ct_0}}$ is a vector of school factors for school *c* as measured in time t_0. The OLS models include state fixed effects (μ_s) to control for unobserved time-invariant characteristics that may be unique for each state, for example, state-specific differences in education investment in high schools and the likelihood of college-going that are stagnant over time (Appendix Table 1.C2 shows the distribution of HSLS:09 high schools with different math and science CGRs by states). ε_{ict} is assumed a zero mean normally distributed error term. The OLS models are estimated by clustering at the school level in order to obtain robust standard errors. The normalized follow-up panel weights for each outcome (W2W1STU for 11th grade math score and W3W1W2STU for college enrollment status) are applied to the data to adjust for the oversampling of certain groups (i.e., Asian students) while minimizing the effects of large sample sizes on standard errors and tests of statistical significance.

Given the postsecondary enrollment outcome is a binary measure of attending a four-year college, equation (1) is estimated with a linear probability model (LPM). The use of LPMs rather than logit or probit models has several advantages. Unlike logit or probit models, which are non-linear transformations of linear regressions, LPMs require weaker distributional assumptions (Wooldridge, 2010). Nonetheless, additional analyses showed that the primary results in this

study are not sensitive to selection among logit, probit, and linear probability models. While all the three binary response models may violate the heteroskedasticity assumption, the issue can be addressed by computing robust standard errors in LPMs, but not in logit or probit models. More importantly, coefficients of LPMs are easier and more straightforward to interpret, especially when comparing estimates across models. This advantage of LPMs is particularly critical for this study as one of the research goals is to compare the estimated effects of postsecondary remediation across different student subgroups.

When testing the equality hypothesis of CGRs (H1b), equation (1) is re-estimated using subsamples of students restricted to a specific subgroup as defined by racial/ethnic (i.e., white, black, Hispanic, Asian, and multirace), socioeconomic (four quartiles: lowest SES, low-middle SES, high-middle SES, and highest SES), and academic ability status (four quartiles: lowest-ability, low-middle ability, high-middle ability, and highest ability). To explore whether and the extent to which the effects of CGRs are moderated by school factors (hypotheses H2a, H2b, and H2c), equation (1) is re-estimated separately by including interaction terms of the two dummy variables of interest, $4M3S_{ict_0}$ and $4M4S_{ict_0}$, with each of the following school contextual variables: (a) structural characteristics: percentage of poverty students, percentage of minority students, and percentage of AP students; (b) academic organizations: requiring a math competency test, offering AP courses, and full-time math/science teacher ratio in school; and (c) social organizations: academic press, school climate, and student engagement.

1.5 Results

1.5.1 The Estimated Impact of High School Math and Science CGRs

One of the primary goals of this study is to evaluate whether high school CGRs improve the overall outcomes of students (H1a: educational productivity hypothesis). Table 1.B2 presents

findings from covariate adjustment regression models with state fixed effects that estimates the effects of CGRs on 11th grade math standardized test scores and 4-year college enrollment status. Results suggest that higher CGRs have both intended positive and unintended negative effects on student outcomes. Compared with schools requiring students to complete three years of math and science coursework to graduate (3M3S), schools with CGRs of four years of math and three years of science (4M3S) increased student math test scores in 11th grade by 0.6 points (corresponding to 0.06 standard deviation for this outcome measure). The estimate is statistically significant at the critical level of 5 percent. The results lend support to the educational productivity hypothesis that higher CGRs improve student achievement overall. Interestingly, there is no significant difference in 11th grade math scores between students who attended schools that also have CGRs of four years of math, along with even higher standard of science CGRs (4M4S), and their peers in schools with CGRs of only three years of math and science (3M3S).

When examining the postsecondary outcome, results show that the likelihood of enrolling in a four-year college immediately after high school senior year is comparable between students in schools with CGRs of four years of math and three years of science (4M3S) and their counterparts attending schools with CGRs of three years of math and science (3M3S). Surprisingly, schools with the highest CGRs in math and science (4M4S) reduced the probability of enrolling in a four-year college for their students by 5.4 percentage points, compared with those students in schools with CGRs of three years of math and science (3M3S; statistically significant at the critical level of 5 percent). The empirical results from linking CGRs in math and science to college enrollment do not support the educational productivity hypothesis, yet they suggest that higher CGRs both in math and science may lead to unintended adverse

consequences on on-time college attendance among students.

In addition, it is worth noting that the estimates on the association between student/school characteristics and outcomes are very consistent with current literature. For example, compared with white counterparts, black students had significantly lower math achievement while Asian students had significantly higher math test scores. Both student socioeconomic status and math ability as measured in 9th grade were positively correlated with 11th grade math test scores and probability of attending a four-year college. Students who attended schools with higher concentrations of low-income, minority students scored lower in math tests whereas students enrolling in schools with higher levels of academic press performed better in standardized tests and college going measures.

1.5.2 Sensitivity Analyses

1.5.2.1 Sensitivity to Specification and Sample Selection

To assess the robustness of the primary results, the impact estimates of CGRs are re-estimated (a) using an alternative specification without weighting and (b) using a subsample restricted to only public school samples. The primary estimation models use panel weights provided by HSLS:09 to adjust for the oversampling of certain student groups and to achieve consistent estimation in the potential presence of endogenous sampling in which the likelihood of selection varies with the outcome measures even after conditioning on the explanatory variables. However, if the sampling probabilities vary exogenously instead of endogenously, weighting might be inappropriate for consistency and detrimental for precision (Wooldridge, 1999). Therefore, re-estimating the effects of CGRs without weighting can serve as a useful procedure to test possible misspecification of model and/or misunderstanding of the sampling process (Solon, Haider, & Wooldridge, 2015). Panel A in Appendix Table 1.C3 displays the

results. For the estimated positive effect of CGRs on math test scores, both weighted and unweighted estimates are quite consistent for the parameters, suggesting the primary regression models with weighting are correctly specified. However, they are less so for the estimated negative effect of CGRs on postsecondary attendance.

The main analytic sample in this study consists of students attending both public and private high schools (i.e., Catholic and non-religious schools). Prior studies have shown that private high schools tend to have stronger academic and social organizations that might facilitate the effects of higher curriculum standards such as CGRs. To evaluate whether the estimated positive effects of CGRs on student achievement is largely driven by private school samples, additional analyses using only public school samples are conducted. Panel B in Appendix Table 1.C3 reports the findings (both with and without weighting). The results do not qualitatively change the main conclusions of the primary analysis using full samples (the following results on the heterogeneous effects of CGRs by student subgroups and the moderating effects of school contextual factors using full samples are also successfully replicated in additional analyses using only public school samples).

1.5.2.2 Quantifying the Robustness of Fixed Effects Inferences

Although an extensive set of student and school controls are included, the covariate adjustment regression models (with state fixed effects) employed in this study might suffer from the threats of omitted variable bias. To address this potential confounding issue, this study follows the sensitivity analysis procedures outlined by Frank, Maroulis, Duong, and Kelcey (2013) to quantify how much bias there must be in the estimates to invalidate the inferences, focusing only on the key findings—the estimated positive effect of CGRs on test score (i.e., 4M3S vs. 3M3S) and the negative effect of CGRs on attending in a four-year college (i.e., 4M4S

vs. 3M3S). As defined by Frank et al. (2013), the calculation of proportion of bias to make an inference invalid is the following:

% bias necessary to invalidate an inference = 1 – threshold for inference/estimated effect, where the threshold for inference = s.e. × $t_{critical,df}$. Applied to the estimates of this study, to invalidate the inference of the positive effect of higher math CGRs on student math score, bias must have accounted for (1 – 1.96 ×.280/.601 = .087) about 8.7% of the estimated effect. Similar calculations suggest that about 23.8% bias must be present to invalidate the inference of the estimated negative effect of higher math and science CGRs on enrolling in a four-year college. According to Frank et al. (2013), the median level of robustness is about 30% for observational studies in education. Thus, in this study the estimated negative impact of CGRs on postsecondary attendance is quite robust while it is less so for the estimated positive effect of CGRs on student math scores.

1.5.3 Heterogeneous Effects of High School CGRs across Student Subgroups

The second part of this analysis explores whether and to what extent CGRs narrow or widen the gaps in student outcomes by examining the heterogeneous effects of math and science CGRs between advantaged and disadvantaged students (H1b: educational equality hypothesis). The subgroup analyses provide empirical evidence partially supporting the educational equality hypothesis of CGRs, yet some unexpected findings. Table 1.B3 presents the estimated differential impacts of CGRs on math test scores and postsecondary attendance by student subgroups as defined by racial/ethnic (shown in Panel A1 and A2), socioeconomic (shown in Panel B1 and B2), and academic ability status (shown in Panel C1 and C2). Following Cohen (1983), Z tests of the differences between coefficients are conducted. The following discusses only those differences between subgroups that are statistically significant at the critical level of 5

percent.

Schools with higher math CGRs (4M3S vs. 3M3S) boosted the probability of attending a four-year college for multiracial students by about 14 percentage points, as compared with their white and black peers. However, such benefits are not observed for multiracial students attending schools with the highest math and science CGRs (4M4S). For socioeconomic subgroups, the results reveal that the estimated negative effect of 4M4S CGRs on attending a four-year college is largely concentrated among high-middle SES students (a decrease by 14.5 percentage points), who also had significantly lower math test scores (a decrease by 1.158 points, corresponding to 0.12 standard deviation), as compared with their counterparts from lowest, low-middle, and highest SES families. For student academic subgroups as categorized based on 9th grade math ability, the findings uncover that the estimated positive effect of higher math CGRs (4M3S vs. 3M3S) is largely driven by students in the lowest and highest end of the academic ability distribution (about 1.2-1.5 points, corresponding to 0.12-0.15 standard deviations), as compared to low-middle ability students. The above conclusions on the differential effects of CGRs remain unchanged even when students are divided into five quintiles, instead of four quartiles, based on the socioeconomic status and initial math ability measures (see Appendix Table 1.C4).

1.5.4 Moderating Effects of School Contextual Factors

Examining the potential moderating effects of school contextual factors on the association between high school CGRs and student outcomes is of major interest to this study. The results are reported in Table 1.B4 where Panel A, B, and C, respectively, show the estimated interaction effects of the two higher CGRs school groups (i.e., 4M3S and 4M4S) with each of (a) school structural characteristics: percentage of poverty students, percentage of minority students, and percentage of AP students; (b) school academic organizations: requiring a math competency

test, offering AP courses, and full-time math/science teacher ratio in school; and (c) school social organizations: academic press, school climate, and student engagement.

As is revealed in Table 1.B4, the H2a hypothesis that the effects of CGRs are moderated by school structural characteristics finds no empirical support, at least for the three school structural variables used in this study, including percentage of poverty students, percentage of minority students, and percentage of AP students. However, there is some evidence supporting the hypotheses of H2b and H2c that CGRs influence student outcomes more positively in schools with stronger academic and social organizations. The estimated positive effects of higher math CGRs (4M3S vs. 3M3S) is largely concentrated in schools that offer on-site AP courses (an increase by 1.441 points, corresponding to 0.14 standard deviation). On the other hand, for those schools offering no AP courses, the estimate is negative although not statistically significant. For schools with highest math and science CGRs (4M4S), a one-standard deviation increase in the scale of school-level student engagement drives growth up by 0.46 points in student math scores (corresponding 0.05 standard deviation). While several school contextual factors are moderators for a positive relationship between CGRs and student academic performance, such moderation effects are not found on the linkage between CGRs and postsecondary enrollment.

1.6 Discussion

The purpose of this study is to determine whether, for whom, under which conditions high school course graduation requirements influence student educational outcomes. The empirical results from analyzing the HSLS:09 data point to three main conclusions. First, on the one hand higher math CGRs had intended positive effect on student math cognitive improvement and on the other hand higher CGRs both in math and science had no impact on student test scores yet decreased the likelihood of students attending a four-year college in the semester after

high school senior year. The findings seem conflicting. One possible explanation is that while students in 4M3S schools were taking more math courses to improve their math ability, students in 4M4S schools were struggling to cope with the high graduation standard in both academic subjects that are often perceived by students as most difficult subjects. Additional analyses show that students attending 4M4S schools did complete more and higher levels of science courses. However, spending more time and putting more effort to take more than three years of science coursework in high school may not pay off in college admission because typical four-year colleges only require applicants to have completed three years of science courses in high school.

The second set of key findings of this analysis is that students who are multiracial, lowest- and highest-achieving benefited the most by enrolling in schools with higher math and science CGRs while high-middle SES students who attended higher math and science schools had significantly lower math scores and lower probability of enrolling in a four-year college. The first part of the results are consistent with a few previous studies that suggested CGRs typically have little overall impact on student outcomes, but it could have a meaningful positive effect for certain traditionally disadvantaged student groups (e.g., Jacob et al., 2016; Plunk et al., 2014). In contrast to the study by Plunk et al. (2014), which documented that higher math and science CGRs reduced the likelihood of postsecondary enrollment for black and Hispanics students, this study found no differential effects of math and science CGRs on college attendance for the two minority groups. Adding to the high school CGRs literature, findings from this study analyzing the HSLS:09 data show that high-middle socioeconomic families struggled most in schools with highest CGRs both in math and science.

The third important conclusion of this study is that school contexts play a significant role in moderating the effects of CGRs on student performance. Two key school moderators

identified in this analysis are advanced course offering and student engagement in school. There could be many other school contextual factors that may be facilitating or impeding the positive impacts of high school CGRs or other types of curricular reform programs. The findings of moderation effects in this study implies that mixed findings of CGR effects presented across current studies could be well understood if contextual factors are taken into account. It also highlights for policymakers and school leaders the need to invest in and develop greater academic and social organizations in schools. The reason is that only introducing institutionalized expectation policies such as higher CGRs or high school exit exams may ultimately not suffice to improve student outcomes.

The findings presented in this study offer new national longitudinal evidence to the research and policy literature on high school CGRs, yet they need to be interpreted with some cautionary limitations. The information on high school course graduation requirements are reported by school administrators, hence, there could be measurement errors in the key independent variables used in this analysis. Although HSLS:09 provided an extensive set of survey items on students, teacher, counselors, and schools for constructing various school contextual factors, there still are some important school characteristics concerning school structural, academic, and social organizations missing in the data set, which may be potential school moderators for CGR effects (e.g., teacher mobility rate and principal instability). Relatedly, while the covariate adjustment regression with state fixed effects models employed in this study have controlled for as many student and school observables as possible when estimating effects, there could be important confounding factors unmeasured and unaccounted for, which could bias the results.

Despite a number of limitations, this study adds to a growing body of evidence that high

school course graduation requirements could serve as a policy tool to improve student achievement by increasing opportunities to learn for traditionally underserved students. Yet policymakers and school leaders need to be aware of and to address the potential unintended negative consequences on certain student outcomes and for specific student subgroups, as documented in this analysis and others. This study also emphasizes that identifying and studying the moderating role of school contextual factors can be useful to conceptually and empirically better understand the linkage between high school CGRs or similar school reform programs and student outcomes.

APPENDICES

APPENDIX A

FIGURES FOR CHAPTER 1

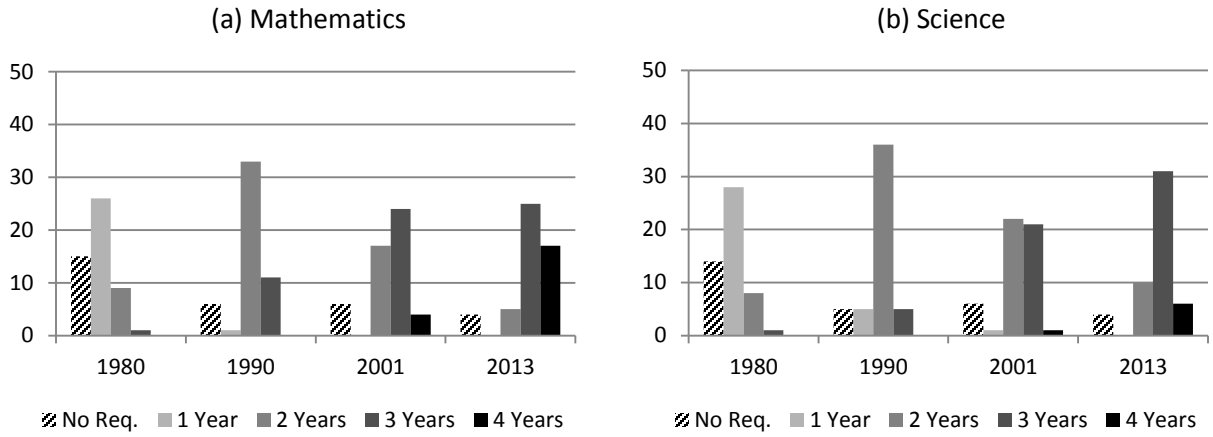


Figure 1.A1 High school mathematics and science course graduation requirements in 50 states and District of Columbia from 1980 to 2013, by year and years of coursework required.
Source. National Center for Education Statistics (2016); Medrich, Brown, Henke, Ross, and McArthur (1992).

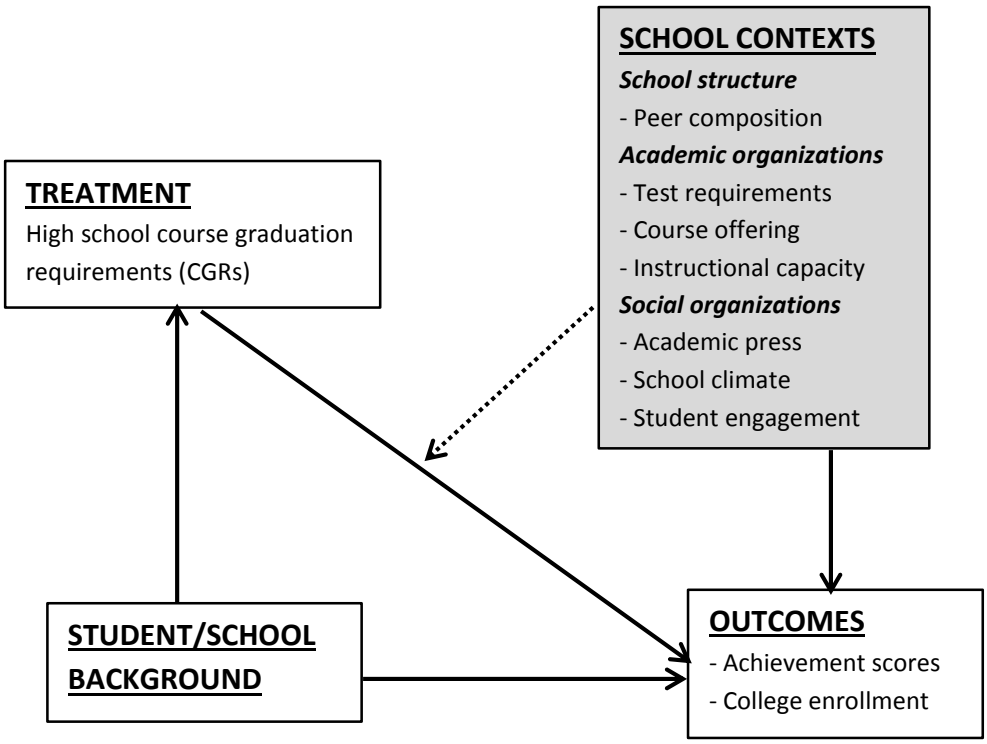


Figure 1.A2 The hypothesized relationship between the course graduation requirements (CGRs) and student outcomes, and the moderating role of school contextual factors.

APPENDIX B

TABLES FOR CHAPTER 1

Table 1.B1 Summary Statistics for Key Variables by High School Mathematics and Science Course Graduation Requirements

		High School Course Graduation Requirements		
	All	3M3S	4M3S	4M4S
Number of students	13,240	4,728	4,905	3,607
Number of schools	695	257	252	186
Race/ethnicity				
White	.598	.659	.627	.479
Black	.129	.102	.114	.184
Hispanic	.164	.129	.147	.232
Asian	.026	.030	.019	.027
Multirace	.074	.072	.076	.074
Other race	.010	.008	.017	.004
Socioeconomic status	.020	.087	-.011	-.040
	(.750)	(.747)	(.755)	(.744)
9th grade math test score	51.101	51.589	50.581	50.976
	(9.659)	(9.765)	(9.863)	(9.238)
School structural organizations				
% of poverty students	.367	.320	.397	.404
	(.243)	(.239)	(.245)	(.237)
% of minority students	.332	.281	.315	.424
	(.290)	(.292)	(.287)	(.269)
% of AP students	.152	.140	.137	.186
	(.126)	(.120)	(.108)	(.145)
School academic organizations				
Requiring a math competency test	.695	.659	.746	.688
Offering AP courses	.899	.898	.905	.896
Math/science teacher ratio	.244	.239	.243	.253
	(.051)	(.049)	(.051)	(.053)
School social organizations				
Academic press	-.037	-.060	-.040	.000
	(1.040)	(1.010)	(1.095)	(1.012)
School climate	-.432	-.469	-.557	-.218
	(.971)	(.923)	(1.003)	(.980)
Student engagement	-.051	-.055	-.178	.104
	(.958)	(.988)	(.900)	(.954)
Student Outcomes				
11th grade math test score	50.452	51.175	50.025	49.894
	(9.866)	(10.036)	(9.900)	(9.516)
Attending a 4-year college	.342	.383	.311	.318

Source. High School Longitudinal Study of 2009 (HSLS:09)

Note. 3M3S = 3 years of math and science; 4M3S = 4 years of math and 3 years of science; 4M4S = 4 years of math and science; AP = Advanced Placement. Sample is restricted to first-time ninth graders in fall 2009, whose schools has valid information on math and science course graduation requirements. Estimates are weighted using base-year student analytic weight (W1STUDENT). Standard deviations appear in the parentheses below means of continuous variables.

Table 1.B2 Effects of High School Math and Science Course Graduation Requirements

	11th grade math test score		Attending a 4-year college	
HS Course graduation requirements (ref: 3 years of math and science; 3M3S)				
4 years of math, 3 years of science (4M3S)	0.601 *	(0.280)	-0.006	(0.018)
4 years of math and science (4M4S)	0.017	(0.347)	-0.054 *	(0.021)
Race/ethnicity (ref: white)				
Black	-0.617 *	(0.278)	-0.001	(0.020)
Hispanic	-0.336	(0.243)	-0.001	(0.017)
Asian	2.226 ***	(0.358)	0.066 **	(0.025)
Multirace	-0.238	(0.319)	-0.023	(0.019)
Other race	-1.605	(0.981)	0.044	(0.063)
Socioeconomic status	0.850 ***	(0.144)	0.125 ***	(0.009)
9th grade math test score	0.518 ***	(0.013)	0.009 ***	(0.001)
School structural organizations				
% of poverty students	-1.683 *	(0.656)	0.012	(0.042)
% of minority students	-0.898 †	(0.470)	0.009	(0.040)
% of AP students	-0.904	(0.785)	-0.036	(0.063)
School academic organizations				
Requiring a math competency test	0.217	(0.203)	-0.015	(0.013)
Offering AP courses	0.723 *	(0.291)	0.024	(0.017)
Math/science teacher ratio	3.763 †	(1.960)	0.231 *	(0.115)
School social organizations				
Academic press	0.221 **	(0.085)	0.015 **	(0.005)
School climate	0.001	(0.109)	0.007	(0.007)
Student engagement	-0.030	(0.133)	0.007	(0.008)
Number of students	13,240		10,509	
Number of schools	695		695	
R-squared	.607		.358	

Note. HS = high school; AP = Advanced Placement. All models estimated with state fixed effects. Models also include student and school covariates presented in Table 1.C1. Standard errors clustered by school are reported in parentheses.
*** p<.001; ** p<.01; * p<.05; †p < .10.

Table 1.B3 Effects of High School Course Graduation Requirements for Student Subgroups

Racial Subgroups	White	Black	Hispanic	Asian	Multirace
Panel A1: 11th Grade Math Test Score					
4 years of math, 3 years of science (4M3S)	0.636*	0.238	1.141 †	1.859 †	1.060
	(0.313)	(0.893)	(0.687)	(0.946)	(0.896)
4 years of math and science (4M4S)	0.245	-0.473	-0.603	-0.576	-0.534
	(0.372)	(1.073)	(0.799)	(1.138)	(1.082)
Number of students	7,822	1,257	1,918	1,006	1,113
Panel A2: Attending a Four-Year College					
4 years of math, 3 years of science (4M3S)	0.002	-0.111 †	0.015	0.120 †	0.142*
	(0.022)	(0.058)	(0.045)	(0.070)	(0.062)
4 years of math and science (4M4S)	-0.046 †	-0.085	-0.004	0.125	-0.077
	(0.028)	(0.069)	(0.055)	(0.083)	(0.072)
Number of students	6,289	960	1,475	826	866

Socioeconomic Subgroups	Lowest SES	Low-Middle SES	High-Middle SES	Highest SES
Panel B1: 11th Grade Math Test Score				
4 years of math, 3 years of science (4M3S)	0.626	0.709	0.448	0.991 *
	(0.527)	(0.547)	(0.447)	(0.436)
4 years of math and science (4M4S)	0.376	0.722	-1.158 *	0.495
	(0.727)	(0.717)	(0.524)	(0.537)
Number of students	3,316	3,304	3,310	3,310
Panel B2: Attending a Four-Year College				
4 years of math, 3 years of science (4M3S)	0.013	0.044	0.011	-0.022
	(0.027)	(0.031)	(0.037)	(0.033)
4 years of math and science (4M4S)	0.028	-0.030	-0.145 **	-0.009
	(0.040)	(0.042)	(0.047)	(0.042)
Number of students	2,437	2,503	2,615	2,954

Math Ability Subgroups	Lowest Ability	Low-Middle Ability	High-Middle Ability	Highest Ability
Panel C1: 11th Grade Math Test Score				
4 years of math, 3 years of science (4M3S)	1.495 **	-0.211	0.738	1.149 **
	(0.457)	(0.511)	(0.480)	(0.415)
4 years of math and science (4M4S)	1.023 †	-0.597	0.372	-0.803
	(0.617)	(0.610)	(0.504)	(0.602)
Number of students	3,310	3,310	3,310	3,310
Panel C2: Attending a Four-Year College				
4 years of math, 3 years of science (4M3S)	-0.015	-0.022	-0.006	0.068 †
	(0.026)	(0.034)	(0.035)	(0.035)
4 years of math and science (4M4S)	-0.042	-0.006	-0.089 *	-0.037
	(0.034)	(0.042)	(0.043)	(0.044)
Number of students	2,419	2,541	2,672	2,877

Note. Each cell in the table shows the estimate on the effect of high school course graduation requirement in math or science (comparison group: 3 years of math and science). All models estimated with state fixed effects. Models include student and school covariates presented in Table 1.C1. Standard errors clustered by school are reported in parentheses.
*** p<.001; ** p<.01; * p<.05; †p < .10 (two-tailed test).

Table 1.B4 Effects of High School Course Graduation Requirements by School Organizations

	11th Grade Math Test Score			Attending a Four-Year College		
Panel A: School Structural Organizations						
4 years of math, 3 years of science (4M3S)	0.408	0.437	0.436	-0.028	-0.015	-0.015
	(0.434)	(0.358)	(0.418)	(0.026)	(0.024)	(0.026)
4 years of math and science (4M4S)	0.191	0.449	0.154	-0.071 **	-0.069 *	-0.069 *
	(0.448)	(0.420)	(0.460)	(0.027)	(0.027)	(0.028)
% of poverty students	-1.930 *			-0.038		
	(0.850)			(0.059)		
% of poverty students x 4M3S	0.812			0.053		
	(0.990)			(0.056)		
% of poverty students x 4M4S	0.357			0.049		
	(0.869)			(0.056)		
% of minority students		-1.281 *			-0.012	
		(0.636)			(0.057)	
% of minority students x 4M3S		1.082			0.027	
		(0.894)			(0.055)	
% of minority students x 4M4S		-0.632			0.035	
		(0.804)			(0.054)	
% of AP students			-1.283			-0.162
			(1.456)			(0.109)
% of AP students x 4M3S			2.327			0.099
			(2.019)			(0.123)
% of AP students x 4M4S			1.087			0.043
			(1.951)			(0.121)
Number of students	12,202	9,715	11,716	7,495	12,239	9,735
Panel B: School Academic Organizations						
4 years of math, 3 years of science (4M3S)	0.608	-0.625	-0.226	-0.008	0.011	0.098
	(0.444)	(0.584)	(1.167)	(0.027)	(0.038)	(0.068)
4 years of math and science (4M4S)	0.114	-0.719	-1.447	-0.034	-0.063	-0.022
	(0.545)	(0.674)	(1.102)	(0.032)	(0.039)	(0.066)
Requiring a math test	0.181			-0.000		
	(0.343)			(0.022)		
Requiring a math test x 4M3S	0.050			-0.003		
	(0.474)			(0.029)		
Requiring a math test x 4M4S	0.011			-0.021		
	(0.492)			(0.031)		
Offering AP courses		-0.009			0.033	
		(0.383)			(0.026)	
Offering AP courses x 4M3S		1.441 **			-0.024	
		(0.554)			(0.037)	
Offering AP courses x 4M4S		1.037			0.007	
		(0.654)			(0.037)	
Math/science teacher ratio			0.046			0.038 *
			(0.302)			(0.019)
Math/science teacher ratio x 4M3S			0.415			-0.038
			(0.450)			(0.026)
Math/science teacher ratio x 4M4S			0.699 †			-0.010
			(0.417)			(0.026)
Number of students	12,154	11,571	11,617	9,640	9,187	9,227

Note. All models estimated with state fixed effects. Models include student and school covariates presented in Table 1.C1. Standard errors clustered by school are reported in parentheses.

*** p<.001; ** p<.01; * p<.05; †p < .10 (two-tailed test).

Table 1.B4 (cont'd)

	11th Grade Math Test Score			Attending a Four-Year College		
Panel C: School Social Organizations						
4 years of math, 3 years of science (4M3S)	0.788 **	0.596 †	0.625 *	-0.005	-0.019	-0.025
	(0.282)	(0.318)	(0.308)	(0.019)	(0.020)	(0.019)
4 years of math and science (4M4S)	0.217	0.265	0.220	-0.032	-0.071 **	-0.057 *
	(0.376)	(0.382)	(0.371)	(0.025)	(0.025)	(0.024)
Academic press	0.232			0.018 †		
	(0.149)			(0.010)		
Academic press x 4M3S	0.035			0.000		
	(0.203)			(0.013)		
Academic press x 4M4S	0.066			-0.022		
	(0.225)			(0.016)		
School climate		-0.161			0.001	
		(0.176)			(0.011)	
School climate x 4M3S		0.061			-0.003	
		(0.202)			(0.012)	
School climate x 4M4S		0.270			0.007	
		(0.218)			(0.014)	
Student engagement			-0.326 †			0.012
			(0.188)			(0.012)
Student engagement x 4M3S			0.250			-0.013
			(0.229)			(0.014)
Student engagement x 4M4S			0.460 *			-0.002
			(0.206)			(0.015)
Number of students	11,771	10,734	11,055	9,332	8,532	8,788

Note. All models estimated with state fixed effects. Models include student and school covariates presented in Table 1.C1. Standard errors clustered by school are reported in parentheses.
*** p<.001; ** p<.01; * p<.05; †p < .10 (two-tailed test).

APPENDIX C

SUPPLEMENTAL TABLES FOR CHAPTER 1

Table 1.C1 Descriptive Statistics of Covariates

Variables	Mean	Standard Deviation	Range
Student characteristics			
Female	.507	.500	0-1
White	.598	.490	0-1
Black	.129	.335	0-1
Hispanic	.164	.370	0-1
Asian	.026	.158	0-1
Multirace	.074	.261	0-1
Other race	.010	.099	0-1
Socioeconomic status (composite)	.020	.750	-1.82-2.88
Family structure			
Intact family	.582	.493	0-1
Two parents/guardians	.178	.383	0-1
Single parent	.222	.415	0-1
Other family structure	.018	.131	0-1
Had a parent worked in STEM fields	.226	.418	0-1
Career aspiration in STEM fields	.336	.472	0-1
Education aspiration			
High school or less	.119	.324	0-1
Associate's degree	.068	.252	0-1
Bachelor's degree	.174	.379	0-1
Advanced degree	.434	.496	0-1
Don't know	.205	.403	0-1
9th grade math test score	51.101	9.659	24.10-82.19
Highest math course taken in 8th grade			
Math 8	.264	.441	0-1
Pre-algebra	.339	.474	0-1
Algebra I	.300	.458	0-1
Algebra II	.055	.229	0-1
Other math	.041	.199	0-1
Number of years of math courses expected to take in HS			
1 or 2 years	.085	.279	0-1
3 years	.245	.430	0-1
4 years	.669	.470	0-1
Knowing the importance of math for applying college	.553	.497	0-1
Knowing the importance of math in college education	.513	.500	0-1
Math identity (composite)	.057	1.002	-1.73-1.76
Math utility (composite)	.003	.996	-3.51-1.31
Math self-efficacy (composite)	.049	.980	-2.92-1.62
Interest in math course (composite)	.054	.996	-2.46-2.08
Science identity (composite)	.043	1.000	-1.57-2.15
Science utility (composite)	.007	.996	-3.10-1.69
Science self-efficacy (composite)	.027	.989	-2.91-1.83
Interest in science course (composite)	.030	.990	-2.59-2.03
Academic commitment (composite)	.014	.988	-4.50-1.60
Sense of school belonging	.071	.975	-4.35-1.59
School engagement (composite)	.078	.958	-3.38-1.39
Hours/week spent on studying/homework	3.158	2.158	.5-16.5

Source. High School Longitudinal Study of 2009 (HSLS:09)

Note. STEM = Science, technology, engineering, and mathematics; HS = high school. Sample is restricted to first-time ninth graders in fall 2009, whose schools has valid information on math and science course graduation requirements. Sample size is 13,240. Estimates are weighted using base-year student analytic weight (W1STUDENT).

Table 1.C1 (cont'd)

Variables	Mean	Standard Deviation	Range
School structural organizations			
% of poverty students	.367	.243	0-1
% of minority students	.332	.290	0-1
% of AP students	.152	.126	0-1
School academic organizations			
Requiring a math competency test	.695	.460	0-1
Offering AP courses	.899	.301	0-1
Math/science teacher ratio	.244	.051	0-.45
School social organizations			
Academic press	-.037	1.040	-6.02-1.43
School climate	-.432	.971	-4.22-1.97
Student engagement	-.051	.958	-2.92-1.81
Other school covariates			
School sector			
Public	.911	.284	0-1
Catholic	.052	.221	0-1
Other private	.037	.190	0-1
Geographic region			
Northeast	.188	.391	0-1
Midwest	.241	.428	0-1
South	.096	.295	0-1
West	.475	.499	0-1
Urbanicity			
City	.244	.430	0-1
Suburb	.344	.475	0-1
Town	.134	.341	0-1
Rural	.277	.447	0-1
Grade span			
PK-12	.054	.225	0-1
6-12	.055	.227	0-1
9-12	.892	.311	0-1
Single-sex school	.023	.151	0-1
Magnet school	.001	.034	0-1
Charter school	.030	.170	0-1
Participated in school choice program	.285	.451	0-1
Average instruction hours per day	6.127	.601	3.8-8
Adequate Yearly Progress (AYP) status			
Met AYP	.695	.460	0-1
Year 1 school improvement	.101	.301	0-1
Year 2 school improvement	.110	.313	0-1
Year 3-5 school improvement	.094	.291	0-1
% of English language learner students	.044	.079	0-.70
% of special education students	.126	.069	0-.45
% of students enrolled in an alternative program	.024	.033	0-.30
% of students enrolled in a dropout prevention program	.016	.035	0-.40
Average daily attendance percentage for HS students	.939	.027	.80-.99

Source. High School Longitudinal Study of 2009 (HSLS:09)

Note. STEM = Science, technology, engineering, and mathematics; HS = high school. Sample is restricted to first-time ninth graders in fall 2009, whose schools has valid information on math and science course graduation requirements. Sample size is 13,240. Estimates are weighted using base-year student analytic weight (W1STUDENT).

Table 1.C1 (cont'd)

Variables	Mean	Standard Deviation	Range
% of 9th graders enrolled in prior year returned	.916	.141	0-1
% of 9th graders who are repeating 9th grade	.049	.068	0-.72
% of graduates attending 2-year colleges	.265	.152	0-1
% of graduates attending 4-year colleges	.499	.238	0-1
Total number of teachers	86.78	49.32	4-260
% of full time teachers	.953	.067	.47-1
% of HS teachers absent on an average day	.034	.025	0-.22
Offering advanced math courses	.942	.234	0-1
Offering alternative programs	.331	.471	0-1
Offering dropout prevention programs	.342	.474	0-1
Having formal GED test preparation program	.149	.356	0-1
Organizing math/science extracurricular programs (composite)	.108	.948	-2.02-2.14
Program encouraging underrepresented students in STEM	.293	.455	0-1
Program informing parents about STEM higher ed/careers	.412	.492	0-1
School counseling program's most emphasized goal			0-1
Help students prepare for postsecondary schooling	.503	.500	0-1
Help students improve achievement in HS	.324	.468	0-1
Other goal (e.g., preparing for work, personal growth)	.174	.379	0-1
Having counselor designated for college-going	.617	.486	0-1
Average caseload for school's counselors	359.515	125.903	4-950
Students are required to have an education plan	.810	.392	0-1
Program to encourage student not considering college to do so	.764	.424	0-1
Consulting with college officers about qualifications	.961	.194	0-1
Organizing student visits to colleges	.702	.457	0-1
Offering Upward Bound/GEAR UP/AVID/MESA	.481	.500	0-1
Holding info session on college transitions for students/parents	.950	.219	0-1
Assisting students with finding financial aid for college	.960	.196	0-1
Providing opportunities for dual/concurrent enrollment	.916	.278	0-1
Taking other steps to assist with HS to college transition	.367	.482	0-1

Source. High School Longitudinal Study of 2009 (HSLS:09)

Note. STEM = Science, technology, engineering, and mathematics; HS = high school. Sample is restricted to first-time ninth graders in fall 2009, whose schools has valid information on math and science course graduation requirements. Sample size is 13,240. Estimates are weighted using base-year student analytic weight (W1STUDENT).

Table 1.C2 Distribution of High School Math and Science Course Graduation Requirements By States

	All	3 Years of Math and Science	4 Years of Math and 3 Years of Science	4 Years of Math and Science
Alabama	14	1	0	13
Alaska	1	0	0	1
Arizona	15	1	13	1
Arkansas	5	0	4	1
California	8	7	0	1
Colorado	8	4	3	1
Connecticut	2	1	1	0
Delaware	4	1	3	0
Florida	43	4	31	8
Georgia	46	2	2	42
Idaho	2	2	0	0
Illinois	19	18	0	1
Indiana	19	13	5	1
Iowa	4	4	0	0
Kansas	7	5	2	0
Kentucky	11	3	8	0
Louisiana	16	1	2	13
Maine	2	1	1	0
Maryland	11	4	7	0
Massachusetts	11	7	3	1
Michigan	42	3	37	2
Minnesota	13	12	1	0
Mississippi	6	1	2	3
Missouri	15	11	2	2
Montana	1	1	0	0
Nebraska	5	5	0	0
Nevada	3	2	0	1
New Hampshire	3	2	1	0
New Jersey	21	16	3	2
New Mexico	4	1	3	0
New York	32	26	2	4
North Carolina	40	3	30	7
North Dakota	2	2	0	0
Ohio	51	23	24	4
Oklahoma	5	4	1	0
Oregon	2	2	0	0
Pennsylvania	49	25	9	15
Rhode Island	2	0	1	1
South Carolina	15	1	12	2
South Dakota	2	2	0	0
Tennessee	44	1	36	7
Texas	48	2	0	46
Utah	1	1	0	0
Vermont	3	3	0	0
Virginia	17	12	1	4
Washington	7	7	0	0
West Virginia	4	0	2	2
Wisconsin	7	7	0	0
Wyoming	2	2	0	0
Total	695	257	252	186

Source. High School Longitudinal Study of 2009 (HSLS:09)

Table 1.C3 Sensitivity Analysis: Effects of High School Math and Science Course Graduation Requirements Using Alternative Specification

Course graduation requirements (ref: 3 years of math and science)	11th grade math test score		Attending a 4-year college	
	Weighted	Unweighted	Weighted	Unweighted
Panel A: All Schools (n=695)				
4 years of math, 3 years of science (4M3S)	0.601 *	0.521 *	-0.006	0.007
	(0.280)	(0.222)	(0.018)	(0.014)
4 years of math and science (4M4S)	0.017	-0.071	-0.054 *	-0.021
	(0.347)	(0.274)	(0.021)	(0.017)
Number of students	13,240		10,509	
Panel B: Only Public Schools (n=558)				
4 years of math, 3 years of science (4M3S)	0.609 †	0.465 †	-0.009	0.010
	(0.319)	(0.264)	(0.020)	(0.017)
4 years of math and science (4M4S)	0.141	0.139	-0.065 **	-0.030
	(0.386)	(0.328)	(0.025)	(0.021)
Number of students	10,530		8,210	

Note. n = number of schools. All models estimated with state fixed effects. Models also include student and school covariates presented in Table 1.C1. Standard errors clustered by school are reported in parentheses.
*** p<.001; ** p<.01; * p<.05; †p < .10.

Table 1.C4 Sensitivity Analysis: Effects of High School Course Graduation Requirements for Student Subgroups Using Alternative Cutoffs

	Socioeconomic Subgroups				
	Lowest	Low-Middle	Middle	High-Middle	Highest
Panel A1: 11th Grade Math Test Score					
4 years of math, 3 years of science (4M3S)	0.345 (0.581)	1.654 ** (0.551)	-0.172 (0.540)	0.366 (0.523)	0.868 † (0.480)
4 years of math and science (4M4S)	-0.038 (0.794)	2.729 *** (0.741)	-1.061 † (0.643)	-1.036 † (0.608)	0.307 (0.625)
Number of students	2,648	2,666	2,630	2,650	2,646
Panel A2: Attending a Four-Year College					
4 years of math, 3 years of science (4M3S)	0.012 (0.029)	-0.026 (0.033)	0.057 (0.042)	-0.059 (0.043)	0.040 (0.037)
4 years of math and science (4M4S)	0.019 (0.041)	-0.076 † (0.046)	-0.005 (0.052)	-0.213 *** (0.049)	0.009 (0.045)
Number of students	1,950	1,992	2,002	2,182	2,383
	Math Ability Subgroups				
	Lowest	Low-Middle	Middle	High-Middle	Highest
Panel B1: 11th Grade Math Test Score					
4 years of math, 3 years of science (4M3S)	1.376 ** (0.501)	0.387 (0.619)	0.062 (0.546)	0.224 (0.533)	1.082 * (0.459)
4 years of math and science (4M4S)	1.180 † (0.685)	-0.704 (0.723)	0.052 (0.665)	-0.372 (0.567)	-0.810 (0.594)
Number of students	2,648	2,648	2,648	2,648	2,648
Panel B2: Attending a Four-Year College					
4 years of math, 3 years of science (4M3S)	-0.005 (0.027)	0.024 (0.032)	-0.058 (0.043)	0.004 (0.040)	0.023 (0.037)
4 years of math and science (4M4S)	-0.003 (0.034)	-0.066 (0.042)	-0.069 (0.046)	-0.120 * (0.051)	-0.027 (0.044)
Number of students	1,927	2,008	2,071	2,169	2,334

Note. Each cell in the table shows the estimate on the effect of high school course graduation requirement in math or science (comparison group: 3 years of math and science). All models estimated with state fixed effects. Models include student and school covariates presented in Table 1.C1. Standard errors clustered by school are reported in parentheses.
*** p<.001; ** p<.01; * p<.05; †p < .10 (two-tailed test).

REFERENCES

REFERENCES

Adelman, C. (1999). *Answers in the toolbox: Academic intensity, attendance patterns, and bachelor's degree attainment*. Washington, DC: US Department of Education.

Adelman, C. (2004). *Principal indicators of student academic histories in postsecondary education, 1972–2000*. Washington, DC: US Department of Education.

Adelman, C. (2006). *The toolbox revisited: Paths to degree completion from high school through college.* Washington, DC: US Department of Education.

American College Testing. (2010). *The condition of college & career readiness 2010*. Iowa City, IA: ACT, Inc.

Angus, D. L., & Mirel, J. E. (1999). *The failed promise of the American high school, 1890-1995*. New York, NY: Teachers College Press.

Attewell, P., & Domina, T. (2008). Raising the bar: Curricular intensity and academic performance. *Educational Evaluation and Policy Analysis, 30*(1), 51-71.

Bidwell, C. E. (1965). The school as a formal organization. In J. G. March (Ed.), *Handbook of organizations* (pp. 972-1022). Chicago, IL: Rand-McNally.

Buddin, R., & Croft, M. (2014). *Do stricter high school graduation requirements improve college readiness?* (ACT Working Paper Series). Retrieved from http://www.act.org/research/-papers/pdf/wp-2014-1.pdf

Bryk, A. S., & Driscoll, M. E. (1988). *The school as community: Theoretical foundations, contextual influences, and consequences for students and teachers*. Madison, WI: National Center on Effective Secondary Schools, University of Wisconsin.

Bryk, A. S., Lee, V. E., & Holland, P. B. (1993). *Catholic schools and the common good.* Cambridge, MA: Harvard University Press.

Bryk, A. S., & Schneider, B. (2002). *Trust in schools: A core resource for improvement*. New York, NY: Russell Sage Foundation.

Chaney, B., Burgdorf, K., & Atash, N. (1997). Influencing achievement through high school graduation requirements. *Educational Evaluation and Policy Analysis, 19*(3), 229-244.

Cohen, A. (1983). Comparing regression coefficients across subsamples: A study of the statistical test. *Sociological Methods & Research, 12*(1), 77-94.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, F., Mood, A. M., Weinfeld, F. D., et al. (1966). *Equality of educational opportunity*. Washington, DC: US Government Printing Office.

Coleman, J. S., Hoffer, T., & Kilgore, S. (1982). *High school achievement*: *Public and private schools*. New York, NY: Basic Books.

The Center for Public Education, & Change the Equation. (2013). *Out of sync: Many common core states have yet to define a common core-worthy diploma*. The Center for Public Education. Retrieved from http://changetheequation.org/sites/default/files/GradReqs_v5.-pdf

Domina, T., & Saldana, J. (2012). Does raising the bar level the playing field? Mathematics curricular intensification and inequality in American high schools, 1982–2004. *American Educational Research Journal, 49*(4), 685-708.

Finn, J. D., & Zimmer, K. S. (2012). Student engagement: What is it? Why does it matter? in S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of student engagement* (pp. 97-131). New York, NY: Springer.

Frank, K. A., Maroulis, S., Duong, M., & Kelcey, B. (2013). What would it take to change an inference? Using Rubin's Causal Model to interpret the robustness of causal inferences. *Education Evaluation and Policy Analysis, 35*(4), 437-460.

Frank, K. A., Zhao, Y., & Borman, K. (2004). Social capital and the diffusion of innovations within organizations: The case of computer technology in schools. *Sociology of Education, 77*(2), 148-171.

Gamoran, A. (1987). The stratification of high school learning opportunities. *Sociology of Education, 60*(3), 135-155.

Gamoran, A. (1996). Curriculum standardization and equality of opportunity in Scottish secondary education: 1984-1990. *Sociology of Education, 69*(1), 1-21.

Goldhaber, D. & Brewer, D. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*(2), 129-145.

Greene, J., & Foster, G. (2003*). Public high school graduation and college readiness rates in the United States* (Manhattan Institute, Center for Civic Information, Education Working Paper, No. 3). New York, NY: Manhattan Institute.

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics, 18*(5), 527-544.

Heck, R. H., Price, C. L., & Thomas, S. L. (2004). Tracks as emergent structures: A network analysis of student differentiation in a high school. *American Journal of Education, 110*(4), 331-353.

Hedges, L. V., & Schneider, B. (Eds.) (2005). *The social organization of schooling*. New York, NY: The Russell Sage Foundation.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371-406.

Jacob, B., Dynarski, S., Frank, K, & Schneider, B. (2016). *Are expectations alone enough? Estimating the effect of a mandatory college-prep curriculum in Michigan*. (NBER Working Paper No. 22013). Cambridge, MA: National Bureau of Economic Research.

Lee, V. E., & Bryk, A. S. (1989). A multilevel model of the social distribution of high school achievement. *Sociology of Education, 62*(3), 172-192.

Lee, V. E., Croninger, R. G., & Smith, J. B. (1997). Course taking, equity, and mathematics learning: Testing the constrained curriculum hypothesis in U.S. secondary schools. *Educational Evaluation and Policy Analysis, 19*(2), 99-121.

Lee, V. E., & Smith, J. B. (1995). Effects of high school restructuring and size on gains in achievement and engagement for early secondary school students. *Sociology of Education, 68*(4), 241-270.

Lee, V. E., & Smith, J. B. (1997). High school size: Which works best, and for whom? *Educational Evaluation and Policy Analysis, 19*(3), 205-227.

Lee, V. E., & Smith, J. B. (1999). Social support and achievement for young adolescents in Chicago: The role of school academic press. *American Educational Research Journal, 36*(4), 907-945.

Lee, V. E., & Smith, J. B. (2001). *High school restructuring and student achievement*. New York, NY: Teachers College Press.

Lillard, D. R., & DeCicca, P. P. (2001). Higher standards, more dropouts? Evidence within and across time. *Economics of Education Review, 20*(5), 459-473.

Long, M. C., Iatarola, P., & Conger, D. (2009). Explaining gaps in readiness for college-level math: The role of high school courses. *Education Finance and Policy, 4*(1), 1-33.

McDonnell, L. (1995). Opportunity to learn as a research concept and a policy instrument. *Educational Evaluation and Policy Analysis, 17*(3), 305-322.

McPartland, J. M., & Schneider, B. (1996). Opportunities to learn and student diversity: Prospects and pitfalls of a common core curriculum. *Sociology of Education, 69*(2), 66-81.

Medrich, E. A., Brown, C. L., Henke, R. R., Ross, L., & McArthur, E. (1992). *Overview and inventory of state requirements for school coursework and attendance*. Washington, DC: US Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Mirel, J., & Angus, D. (1994). High standards for all? The struggle for equality in American high school curriculum, 1890-1990. *American Educator, 18*(2), 4-9, 40-42.

Montgomery, N., & Allensworth, E. M. (2010). *Passing through science: The effects of raising graduation requirements in science on course-taking and academic achievement in Chicago*. Chicago, IL: Consortium on Chicago School Research.

Muller, C. (1998). The minimum competency exam requirement, teachers' and students' expectations and academic performance. *Social Psychology of Education, 2*(2), 199-216.

Muller, C., Riegle-Crumb, C., Schiller, K.S., Wilkinson, L., & Frank, K. A. (2010). Race and academic achievement in racially diverse high schools: Opportunity and stratification. *Teachers College Record, 112*(4), 4-5.

National Center for Education Statistics. (2016). *Table 234.30. Course credit requirements and exit exam requirements for a standard high school diploma and the use of other high school completion credentials, by state: 2013*. Retrieved from https://nces.ed.gov/programs/digest/d14/tables/dt14_234.30.asp?current=yes

Newmann, F. (Ed.) (1992). *Student engagement and achievement in American secondary schools*. New York, NY: Teachers College Press.

Plunk, A. D., Tate, W. F., Bierut, L. J., & Grucza, R. A. (2014). Intended and unintended effects of state-mandated high school science and mathematics course graduation requirements on educational attainment. *Educational Researcher, 43*(5), 230-241.

Roscigno, V. J., & Crowley, M. (2001). Rurality, institutional disadvantage, and achievement/attainment. *Rural Sociology, 66*(2), 268-292.

Roscigno, V. J., Tomaskovic-Devey, D., & Crowley, M. L. (2006). Education and the inequalities of place. *Social Forces, 84*(4), 2121-2145.

Schmidt, W., & Maier, A. (2009). Opportunity to learn. In G. Sykes, B. Schneider, & D. Plank (Eds.), *Handbook of education policy research* (pp. 541-559). New York, NY: Routledge.

Schmidt, W. H., McKnight, c. c., Houang, R. T., Wang, H. C., Wiley, D. E., Cogan, L. S., Wolfe, R. G. (2001). *Why schools matter: A cross-national comparison of curriculum and learning*. San Francisco, CA: Jossey-Bass Publishing.

Schneider, B., Swanson, C. B., & Riegle-Crumb, C. (1998). Opportunities for learning: Course sequences and positional advantages. *Social Psychology of Education, 2*(1), 25-53.

Schwille, J., Porter, A., Belli, G., Floden, R., Freeman, D., Knappen, L., Kuhs, T., & Schmidt, W. (1983). Teachers as policy brokers in the content of elementary school mathematics. In L. S. Shulman & G. Sykes (Eds.), *Handbook of teaching and policy* (pp. 370-391). New York, NY: Longman.

Solon, G., Haider, S. J., & Wooldridge, J. M. (2015). What are we weighting for? *Journal of Human resources, 50*(2), 301-316.

Sorensen, A. B. (1970). Organizational differentiation of students and educational opportunity. *Sociology of Education, 43*(4), 355-376.

Southworth, S., & Mickelson, R. (2007). The interactive effects of race, gender, and school composition on college track placement. *Social Forces, 86*(2), 497-523.

Smerdon, B., & Borman, K. M. (2012). *Pressing forward: Increasing and expanding rigor and relevance in America's high schools*. Charlotte, NC: Information Age.

Spillane, J. P., & Louis, K. S. (2005). School improvement processes and practices: Professional learning for building instructional capacity. *Yearbook of the National Society for the Study of Education, 101*(1), 83-104.

Teitelbaum, P. (2003). The influence of high school graduation requirement policies in mathematics and science on student coursetaking patterns and achievement. *Educational Evaluation and Policy Analysis, 25*(1), 31-57.

Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A. (2013). A review of school climate research. *Review of Educational Research, 83*(3), 357-385.

U.S. Department of Education, National Commission on Excellence in Education (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: Author.

Wilson, B. L., & Rossman, G. B. (1993). *Mandating academic excellence: High school responses to state curriculum reform*. New York, NY: Teachers College Press.

Wooldridge, J. M. (1999). Asymptotic properties of weighted M-estimators for variable probability samples. *Econometrica, 67*(6), 1385-1406.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: MIT Press.

CHAPTER 2

REDUCING OR REINFORCING INEQUALITY? EVALUATING THE IMPACT OF POSTSECONDARY REMEDIATION ON COLLEGE OUTCOMES[3]

2.1 Introduction

Each year, many students enter postsecondary institutions academically underprepared with respect to their numeracy and literacy skills (Parsad, Lewis, & Greene, 2003; Sparks & Malkus, 2013). Recent national statistics indicate that nearly half of undergraduates while enrolled in a two- or four-year college will take at least one remedial course in mathematics or English (Radford & Horn, 2012). Remedial education in college is now a growing concern among educators and policymakers as it costs an estimated $5.6-$7.0 billion a year nationwide (Alliance for Excellent Education, 2011; Scott-Clayton, Crosta, & Belfield, 2014). With so many students and enormous resources involved, the question is whether postsecondary remediation increases the likelihood that students will persist and graduate.

Remediation was initially established to increase college access and success for those academically ill-prepared students who otherwise would have stopped out after high school (Day & McCabe, 1997; Roueche & Roueche, 1999). Statistics consistently show that students who are low-income, black, and Hispanic are disproportionately enrolled in remediation programs across higher education institutions (Adelman, 2004; Radford & Horn, 2012; Sparks & Malkus, 2013). While the benefits of remediation remain unclear, some scholars have argued that such practices,

[3] This research was supported by the NLSY 1997 Postsecondary Research Network funded by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under award number 5R01HD061551-02 and by the Population Research Center at the University of Texas at Austin, which receives core support from the National Institute of Child Health and Human Development under the award number 5 R24 HD042849. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

instead of mitigating differences in college preparation, act as an institutional mechanism for resorting and "cooling out" low-performing disadvantaged students (Bettinger & Long, 2004; Deil-Amen & Rosenbaum, 2002). The next question, then, is whether remediation reduces or reinforces the educational inequality between socially and economically advantaged and disadvantaged student groups?

A large body of empirical research has examined the remediation effects on a variety of college outcomes (e.g., grade, credits earned, persistence, and degree completion) using student samples from two-year and/or four-year institutions. Two reviews of the remediation literature found that earlier studies with observational data were severely flawed methodologically (O'Hear & MacDonald, 1995; Boylan & Saxon, 1999), in particular failing to account for confounding effects arising from non-random remediation participation of students (Bettinger, Boatman, & Long, 2013). Several researchers recently attempted to address the issue of selection bias by employing experimental and quasi-experimental designs. These research efforts produced results that were mixed at best. While some studies found that remediation has a positive but modest impact for two-year college students (Moss & Yeaton, 2013; Moss, Yeaton, & Lloyd, 2014), the others reported an effect that is neutral or negative (Clotfelter, Ladd, Muschkin, & Vigdor, 2015; Martorell & McFarlin, 2011). Similarly, inconsistent estimates are observed for remediation among four-year college students (Attewell, Lavin, Domina, & Levey, 2006; Bettinger & Long, 2009; Martorell & McFarlin, 2011).

In addition to conflicting findings, these previous studies were subject to several important limitations. Most researchers focused only on estimating the independent effect of remediation for different remedial subjects (e.g., math, reading, or writing) separately. These subject-focused studies fail to recognize the fact that a sizeable number of undergraduates are

exposed to multiple subjects at the same time (Bahr, 2007, 2010; Parsad et al., 2003). Most empirical analyses also heavily relied on student data from a single institution or state, which has limited generalizability to other settings and student populations. Moreover, current research has little to provide on the heterogeneous effects of remediation for various socio-demographic groups, especially with nationally representative samples.

This study contributes to this literature by offering new longitudinal empirical evidence on the remediation effect and its association with educational inequality. Unlike prior studies that are overwhelmingly subject-focused, in this analysis students are categorized into different "remedial treatment groups" which consider both the subject areas and the number of enrolled remedial courses. Effects of single remediation (i.e., math or English) and dual remediation (i.e., both math and English) are tested with a nationally generalizable sample of a recent cohort of college students from two- and four-year institutions. A propensity score-based technique with stratification and inverse probability weighting is used to estimate the remediation effects while controlling for potential selection bias. Findings suggest that postsecondary remediation has differential effects for two- and four-year college students and it plays a critical role in the social stratification process in higher education system.

2.2 Background

2.2.1 How Might Postsecondary Remediation Help or Hinder Student Success?

Remediation is perhaps the most common, large-scale intervention that postsecondary institutions use to address academic deficiencies among college students with poor preparation, many of whom are racial and ethnic minorities and from low-income families. Remedial education consists of courses and other learning programs aimed at improving students' academic abilities and preparing them to take college-level coursework. Assignment to a

remedial course typically occurs when undergraduates first arrive on campus. However, it is not uncommon to find students enrolling in remedial classes after their first semester or year in college (Adelman, 1999; Bahr, 2013; Bailey, Jeong, & Cho, 2010; Parsad et al., 2003). Nationally, larger percentages of students attending two-year colleges reported enrollment in remedial coursework than did those attending four-year colleges (among 2003-04 first-time postsecondary students: 68.2% vs. 39.4%; Radford & Horn, 2012).

In recent literature, there are three major arguments as to how postsecondary remediation can influence student college pathways and outcomes in different ways. The first, and a positive view of remediation, is that postsecondary remedial education, often called "developmental education," should help students to foster or strengthen their skills in certain academic areas (e.g., math, reading, and writing) that are critical for success in college (Bettinger et al., 2013). The benefits of remediation, therefore, should include an increased probability of college persistence and completion. This positive developmental effect, however, may be weaker for those students who enrolled in but did not complete a remedial course as they would only receive partial treatment thus limited benefits of remediation (Bettinger & Long, 2004).

The second perspective argues that remediation could negatively affect students in unintended ways. Failing a placement test upon entering college and being assigned to a remedial course can send a signal to the students that they are not "college material" (Clotfelter et al., 2015; Deil-Amen & Rosenbaum, 2002; Scott-Clayton & Rodriguez, 2015). Such a stigma may discourage students from continuing their postsecondary studies as they might feel that they are not succeeding, not up to the task, and subsequently they may not do well in college. This stigma is likely to be more pronounced for remedial students attending a selective college where a larger proportion of their peers are not involved in any remedial programs (Bettinger & Long,

2004; Deil-Amen & Rosenbaum, 2002). If it happens early in a postsecondary career then the students may view themselves as inadequate and not try in other courses—which may explain why many studies show low college persistence and graduation rates for those who take remediation early in college (Adelman, 2004; Bahr, 2013; Radford & Horn, 2012).

The third view of remediation suggests that enrolling in remedial classes may have a disrupting effect on schooling progression, regardless of its impact on student learning or performance. For the vast majority of higher education institutions, it is a common practice that academically underprepared students are required to take and pass noncredit remedial courses prior to taking college-level coursework (Parsad et al., 2003). Students with greater academic deficiencies would even take more than one or two semesters to meet the prerequisite requirements (Bahr, 2013; Bailey et al., 2010; Parsad et al., 2003). This particular enrollment restriction is likely to disrupt coursetaking patterns and impose extra financial burdens on remedial students in college, especially those students who need to take multiple remedial subjects. If the disruption effect is large, the remediated students are expected to take more semesters or years to graduate or to be less likely to obtain a degree as compared with their non-remediated peers (Scott-Clayton & Rodriguez, 2015).

2.2.2 How Might Postsecondary Remediation Affect Students Differently?

One consistent finding in educational statistics is that low-socioeconomic and minority students account for a larger share of enrollment in postsecondary remediation (Adelman, 2004; Radford & Horn, 2012; Sparks & Malkus, 2013). Part of the reason is that polices such as the open admissions of community colleges have increased the college participation rates among socially and economically disadvantaged groups of students who typically have poorer pre-college academic preparation (Lavin, Alba, & Silberstein, 1981; Lavin & Hyllegard, 1996). If

remediation has a strong positive developmental effect, it should generally help increase the proportion of low-income and minority students who persist in college and eventually obtain a degree. However, if the negative effects of remediation outweigh its positive functions, it will widen the gaps in college persistence and completion between advantaged and disadvantaged groups.

The interpretation of the role of remediation in educational inequality could be misleading if it is only derived from results from estimating the main effects of remediation. It is important to recognize that heterogeneity among remedial students based on racial and socioeconomic status may interact with remediation to produce differential effects on college outcomes. A growing body of sociological literature has shown that racial and ethnic minorities in postsecondary institutions could suffer from stereotype threat, a situation in which negative-ability stereotypes increase their cognitive psychological load and reduce their academic effort (Massey & Fischer, 2005; Owens & Lynch, 2012; Steele, 1995). Participating in a remedial class, therefore, is likely to trigger the stereotype threat effect among minority students as they are exposed to the signal of being a member of low-achieving groups in colleges. According to stereotype threat perspective, we would expect to find that remediation affects black and Hispanic students more negatively. The negative effect of stereotype threat could be stronger in selective or four-year colleges that enrolled predominantly white students while it might not be the case in less selective or two-year colleges.

Similar to racial minority groups, low-socioeconomic students tend to be low-achieving and underrepresented in higher education campuses. Yet, they are overrepresented in remedial classes. It is likely that students whose parents had no college education may experience negative effects from enrolling in remediation to a greater extent than do their counterparts whose parents

had some college education. Prior research has documented that first-generation college students (those who are the first in their families to attend a postsecondary institution) struggle to adjust to new cultures or learning styles in the college environment, which are more closely aligned with the cultural capital and academic experience possessed by non-first-generation students (e.g., Amstrong & Hamilton, 2013; London, 1989; Pascarella, Pierson, Wolniak, & Terenzini, 2004). Lacking a sense of "fitting in" or social belonging, as has been shown, can undermine not only subjective well-being but also intellectual performance in college (Walton & Cohen, 2007, 2011). Remedial students with first-generation status thus are likely to be at a higher risk of dropping out of college as being placed in a remedial class may evoke and exaggerate the feeling of uncertainty about their belonging in postsecondary institutions.

2.3 Prior Evidence on Postsecondary Remediation

Several empirical studies have sought to isolate the causal impact of postsecondary remediation by employing experimental and quasi-experimental designs. The only three random assignment studies that sampled students from a single or multiple institutions (mostly two-year colleges) generally found positive or neutral effects of remediation (Barnett et al., 2012; Moss et al., 2014; Visher, Weiss, Weissman, Rudd, & Wathington, 2012). Moss et al. (2014) reported that remediation generated small positive effects on grades in college-level math courses. Barnett et al. (2012) also documented that remediation has a positive effect on completion of college-level courses in math but no impact in reading, whereas Visher et al. (2012) found modest positive effects of remediation on course credits in subjects other than math. Although the experimental studies provide credible causal estimates, their findings are not generalizable across various demographic and geographical backgrounds. Moreover, none of these randomized

controlled trials provide evidence on longer-term college outcomes such as institutional transfer or degree attainment.

A number of researchers have used statewide administrative data with quasi-experimental longitudinal designs to test the remediation effects. The empirical results for two-year college student samples from different states are mostly mixed and far from conclusive. For instance, Bettinger and Long (2005) and Calcagno and Long (2008) documented that enrollment in a remedial math course at two-year colleges in Ohio and Florida increased the number of earned credits and the probability of persistence and transferring up to a four-year college although the positive effect is not found for English remediation or when examining long-term attainment outcomes such as degree completion. In contrast, Clotfelter et al. (2015) found that in North Carolina remediation enrollment in either math or English in two-year colleges significantly decreased the likelihood of passing a college-level math or English course and the likelihood of college success in terms of obtaining a degree or diploma.

The available, but limited, findings on remediation effects for four-year college students from different states are also inconsistent. Drawing on statewide four-year college data from Ohio, Bettinger and Long (2009) found that remediation enrollment significantly increased the probability of persistence, not transferring down to a two-year college, and degree receipt. However, Boatman and Long (2010) analyzed two- and four-year college student data from Tennessee and reported that remediation in math, reading, or writing generally has a negative effect on total credits completed, persistence, and graduation. In another study using statewide data from Texas, which includes both two- and four-year student samples, Martorell and McFarlin (2011) showed that assignment to a math or English remedial class had no impact on earned credits, transfer behavior, and degree completion.

The variation in findings for remediation effects may be well explained from research design and methodological perspectives. First, previous quasi-experimental studies tended to analyze student data from a single state. Although the results are informative in evaluating specific remediation programs or policies at the state level, they may not be generalizable to other parts of the country. Second, a crucial limitation of these studies is that researchers heavily relied on analytic strategies such as regression discontinuity (RD) designs and instrumental variables (IV). While the uses of RD and IV techniques can provide strong internal validity in non-experimental settings, they only yield local average treatment effect (LATE) estimates, which have limited external validity. Studies employing RD and IV approaches essentially focus only on "marginal" students (e.g., by comparing students just above and just below cutoffs of remedial placement tests), which means they produce estimates only for a subgroup, not for the whole sample population.

There is a scarcity of research on postsecondary remediation effects based on a nationally representative sample. The only available study was conducted by Attewell et al. (2006), which used data from the National Education Longitudinal Study of 1988 (NELS:88) and employed propensity score matching methods. The analysis reported mixed results of remediation effects in math and English for both two- and four-year college students. Despite informative findings, their study is limited in several respects. First, the generalizability of their results is difficult to gauge as the final matched samples across different models range from about 53% to as low as 8% of the original defined study population (12 out of 20 models used less than one-third of the student sample). Second, remediation variables in NELS are created based on participants' transcripts throughout their entire postsecondary careers, thus making it impossible to determine the timing of remediation enrollment. As a result, Attewell et al. (2006) focused only on

exploring how remediation predicts degree completion while producing no evidence on other important college outcomes such as persistence and upward/downward transfer. Finally, Attewell et al. (2006) did not examine the heterogeneous effects for student subgroups that bear directly on the question concerning the role of remediation in educational stratification process in higher education.

2.4 The Study

This study builds on previous work to investigate whether and the extent to which postsecondary remediation participation affects student persistence, transfer, and graduation, and how the effects vary by socio-demographic subgroups. Prior studies have primarily focused on estimating the effects of enrollment in a specific remedial subject (e.g., math, reading, and writing). *Subject-focused* analysis works best for evaluating subject-specific domain outcomes such as cognitive improvement or passing a college-level course in a given subject area. However, when considering attainment outcomes such as college persistence and completion, subject-focused estimates provide limited insight as many students would take multiple remedial subjects and the joint effect of multi-subject remediation is likely not simply the sum of the estimates on each remedial subject. The present study takes a *person-focused* approach to understand how remediation influence students' college attainments by examining the outcomes of students in different "remedial treatment groups" which are characterized based on individuals' coursetaking patterns in remediation.

2.4.1 Data and Sample

A systematic examination of remediation effects requires detailed information on students' coursetaking, background, and college experiences. The National Longitudinal Survey of Youth of 1997 (NLSY97) and its postsecondary transcript (PSTRAN) data are used in this

analysis. The NLSY97 is a U.S. nationally representative sample of 8,984 individuals (including 6,748 respondents in the cross-sectional sample and 2,236 respondents in the minority supplemental sample), who were 12-16 years old as of December 31, 1996 and then aged 27-31 in 2011. The NLSY97 survey conducted annually since 1997 and the most recent released data serve the purposes of the present study well because it contains postsecondary transcript data with term-by-term and course-by-course records in regard to college enrollment, coursetaking, and completion for respondents who reported attendance in a postsecondary degree program during any of the annual interviews from 1997 to 2011 (rounds 1 through 15).

This study focuses on estimating the effects of remediation enrollment in the first term of students' college careers as it is the period in which most academically underprepared students would be required or choose to take remedial courses. To construct an analysis sample for this study, a number of restrictions are imposed on the data. First, this analysis focuses only on the individuals whose postsecondary transcript data are available, who account for 3,818 of the original sample in the NLSY97. Second, the analytic sample is limited to those NLSY97 participants who started their postsecondary schooling career in either a two-year or four-year college and had valid information on their postsecondary enrollment status and coursework (deleting 6.7% of students from the baseline transcript data).

Third, the analytic sample excludes students who (a) enrolled in only one course (further deleting 14.0% of students) and (b) enrolled in more than two remedial courses (further deleting 5.4% of students), during their first term in college. In doing so, those students who were taking only one college course but likely not enrolling in any degree program and who were freshmen with the greatest academic deficiencies are dropped from this analysis. Including these two groups of students in the analysis will create difficulty in achieving balanced comparison groups

as they possess very different characteristics in terms of postsecondary attendance and academic background. Fourth, a small proportion of students whose number of courses enrolled during first term of college is more than seven are excluded from the analysis as these observations are likely to be data errors (further deleting 4.3% of students). The remaining sample includes 2,773 students. Among them, 1,191 individuals were first-time college students in a two-year college whereas 1,582 individuals were students in a four-year college.

Distinguishing between two- and four-year college students is an important procedure in investigating postsecondary remediation effects on college outcomes. Students beginning their postsecondary career in a two-year college are likely to possess certain characteristics that are systematically different from their four-year college peers. As shown in prior empirical studies, remediation effects may vary by college selectivity (e.g., Attewell et al., 2006; Bettinger & Long, 2007; Boatman & Long, 2010). Conceptually, the student demographic composition and the share of remedial students in two- and four-year institutions can be substantially different, which means that the potential stigma effects of remediation could be functioning at different magnitudes. Furthermore, when evaluating the impact of remediation, there are separate outcome measures that can be used for two-year (e.g., transferring up to a four-year college, earning an associate's degree [AA]) and four-year college students (e.g., transferring down to a two-year college, earning a bachelor's degree [BA]).

2.4.2 Measures

Prior studies have suggested that postsecondary remediation could affect student outcomes differently, depending on the subject area (Attewell et al., 2006; Bettinger & Long, 2005) and the number of remedial courses taken (Attewell et al., 2006). In this study, students are grouped into distinct remediation categories, taking into account both the subject type and the

number of remedial courses enrolled during their first term of college. By using individuals' course-level information recorded in the transcript data, remedial courses were identified using the 2010 College Course Map (CCM), which provides a taxonomy system for classifying postsecondary classes.[4] All students in the sample are further classified into four "remedial treatment groups": (a) only-math remediation, (b) only-English remediation, (c) both-subject remediation (who were simultaneously enrolled in a math and an English remedial course), and (d) no remediation (as comparison group). Such grouping covers 97.5% of total students in the sample. The remaining 2.5% of students are those who enrolled in two math or two English remedial courses (14 and 55 students respectively). The number of these students is too small to form a new treatment group that has sufficient statistical power to detect meaningful differences in remediation effects. Therefore, they are excluded from this analysis. The final sample includes 2,704 students (1,144 students in two-year colleges and 1,560 students in four-year colleges).

The dependent variables for this study are student outcomes as measured by college persistence, transferring between two- and four-year institutions, and postsecondary attainment. As shown in previous studies (e.g., Attewell et al., 2006; Martorell & McFarlin, 2011), not only is the type of educational outcome important in understanding the impact of postsecondary remediation, equally important is the timing of attaining these outcomes (e.g., on-time or delayed graduation), especially for identifying the potential disruption effect of remediation. In this analysis, detailed chronological, term-level information about students' enrollment across postsecondary institutions and degree completion reported in the transcript data is used to create a series of time-relevant college outcomes. The first set of outcomes on college persistence includes (a) persisted into the 2nd year, and (b) persisted into the 3rd year. The second set of outcomes focuses on institutional transfer between two- and four-year colleges. For two-year

[4] The CCM code for remedial math is "32.0104" and for remedial English "32.0108."

college students, the positive transfer outcomes include (a) transferred upward in 2nd year, and (b) transferred upward in 3rd year; for four-year college students, reverse transfer is regarded as a negative transfer outcome, including (a) transferred downward in 2nd year, and (b) transferred downward in 3rd year. The third set of outcomes on degree completion includes: for two-year college students, (a) earned an AA within 3 years, (b) earned an AA within 4 years, (c) earned a BA within 6 years, and (d) earned a BA within 8 years; and for four-year college students, (a) earned a BA within 4 years, (b) earned a BA within 6 years, and (c) earned a BA within 8 years.[5]

As suggested in the literature, postsecondary remediation effects can largely be explained by student backgrounds, academic preparation, and institutional characteristics (Bettinger et al., 2013). To effectively capture the selection process into postsecondary remediation in math and English, this study uses a collection of individual and contextual covariates (Appendix Table 2.B1 presents the descriptive statistics of covariates). Several important individual and demographic characteristics are: gender, race/ethnicity, cognitive score, parent's highest education, poverty level, intact family status, mother's age at first birth, census region of residence when at age 16, and age when starting college. This analysis also draws on an extensive set of pre-college and college measures with respect to high school academic preparation (including school sector, academic program, gifted education, math pipeline, science pipeline, total earned academic credits, overall grade point average [GPA], math GPA, and English GPA), pre-college schooling experience (including late for school, absent from school, retention, percent of peers who cut classes or school, and percent of peers who plan to go to

[5] The outcome measure of earning a BA degree within four years is not used as a dependent variable for two-year college students because it is very infrequent in the data (see Table 2.A1). Similarly, the outcome measures of obtaining an AA degree within three and four years are not used as dependent variables for four-year college students.

college), and postsecondary attendance (including on-time college enrollment, college sector, college major, number of courses taken in first term).

Table 2.A1 presents descriptive statistics of key variables by remediation status during students' first term in a postsecondary degree program for two- and four-year college students separately. The NLSY97 data provide updated and new national estimates on postsecondary remediation patterns. First, consistent with other national statistics (e.g., Radford & Horn, 2012; Sparks & Malkus, 2013), remediation rates among NLSY97 respondents were higher for two-year college students than for their four-year college peers (64.9% vs. 53.3%). Among remedial students in two-year colleges, remediation enrollment in both subjects was more common (46.1%) than it was in only English (29.3%) or only math (24.6%), whereas in four-year colleges, remediation enrollment in the single subject of English was the majority (46.7%), followed by enrollment in both subjects (28.8%) and in math only (24.5%).

Across baseline characteristics, there are clear similarities and differences by remediation status for two- and four-year college students.[6] Compared with their nonremediated peers, remedial students tend to be younger when first enrolling, on-time college entrants, and taking more courses during their first term across both types of postsecondary institutions. Black students enrolled in any remediation at a higher rate in four-year colleges but not in two-year colleges. Individuals who have lower ability, poorer academic preparation, and were from lower socio-economic families were more likely to be enrolled in remedial courses at four-year universities but were less likely to take remedial courses at two-year institutions. Some of these descriptive statistics on the academic and family background of remedial students in two-year colleges are counterintuitive. Common sense would suggest that remedial students tend to be

[6] Only differences between "any remediation" and "no remediation" groups that are statistically significant in t-test are discussed in text.

those who are academically and socioeconomically disadvantaged. Additional analyses comparing samples of two-year college students in NLSY97 and in the Education Longitudinal Study of 2002 (ELS:2002; the most recent national high school longitudinal study of a cohort of U.S. students whose postsecondary transcript data are available) suggest that students with higher ability and higher socio-economic status are overrepresented in the remedial groups in NLSY97 (details are available upon request). Therefore, the use of two-year college student sample from NLSY97 may be a limitation of this study in terms of the generalizability of findings. The issues of external and internal validity of analyzing the two-year college student sample from NLSY97 are further discussed in the limitations section.

2.4.3 Analytic Strategy

Estimating the impact of postsecondary remediation with non-experimental data such as NLSY97 is challenging since students who participate in remediation tend to be systematically different from their non-remedial peers. Two important selection mechanisms are likely to affect remediation enrollment. Students who have weaker academic skills tend to be placed in a remedial class upon entering college. It also could be that students who aspire to succeed in college enroll in these remedial courses and work and try harder than students who do not take these courses. Simply comparing the outcomes of remediated students to their non-remediated counterparts will yield estimates of remediation effects that are biased downward in the first instance and biased upward in the second instance. To minimize the potential selection bias when estimating the postsecondary remediation effects with the observational data from NLSY97, this study employs the marginal means weighting through stratification (MMW-S) method, which can effectively remove selection bias associated with observed pretreatment covariates, under the strong ignorability assumption that the treatment assignment was

independent of the unobserved characteristics given the measured covariates (Rosenbaum & Rubin, 1983, 1984; Rubin, 2005).

The MMW-S was recently developed for evaluating the effects of multivalued and multiple concurrent treatments with non-experimental data (Hong, 2010, 2012). It is particularly useful for this analysis because the remediation treatment is defined by a categorical variable with four groups: "only-math remediation," "only-English remediation," "both-subject remediation," and "no remediation (as reference group)." The MMW-S combines the principle of inverse probability of treatment weights (IPTW; Robins, 1999; Rosenbaum, 1987) with stratification on the propensity score (Rosenbaum & Rubin, 1984). It functions in the spirit of approximating a randomized controlled experiment by facilitating direct comparisons between individuals in a "treatment group" and those in a "control group," both of whom have very similar chances of being assigned to the treatment.

To implement the MMW-S method for estimating the postsecondary remediation effects, this study follows the procedures laid out by Hong (2010, 2012). First, there are a number of observed pretreatment covariates that are identified as being correlated with remediation treatment status (for two-year college analysis: 31 covariates; for four-year college analysis: 37 covariates).[7] For covariates with missing values, indicators of missing data are created to account for different missing patterns. Then, a multinomial logistic regression model at the individual level is estimated for obtaining three propensity scores per individual for three of the four remediation treatment statuses. Those individuals who have a nonzero probability of being assigned in each of the four remediation treatment statuses are included in the analytic sample

[7] All covariates listed in Appendix Table 2.B1 are used to predict remedial treatment status in a series of bivariate multinomial logistics regression models. For two-year college students, 31 covariates are significantly correlated with remedial treatment conditions (at the critical level of 10%), except race/ethnicity, late for school, absent from school, percent of peers who cut classes, high school sector, and high school science pipeline; for four-year college analysis, 37 covariates, except gender and high school program.

for causal inference, which consists of 1,091 students for two-year college analysis (after dropping 4.6% of students) and 1,451 students for four-year analysis (after dropping 7.0% of students). In comparison with the students included in the analytic sample, those excluded from the analytic sample tend to be individuals who are less likely to share similar characteristics with their peers across treatment conditions. For example, the "no remediation" students excluded from the four-year college sample on average have cognitive scores that are about one-half standard deviation above the mean. They are likely to be high-ability students for whom it is difficult to find "matched" cases in any remediation treatment groups.

With the estimated propensity scores, the analytic sample is stratified into either five or six strata for each of the four remediation treatment conditions. According to Cochran (1968), dividing a sample into five strata typically removes at least 90% of the bias associated with a pretreatment covariate, reducing the potential selection bias to a great extent. The next step was to compute a marginal mean weight for each individual in each treated group. After repeating the same procedure for all four treatment groups, the weighted sample, in theory, approximated a randomized experiment within each stratum of students under the assumption of no unmeasured confounders. The covariate balance between the treated students and untreated students in the distribution of all pretreatment covariates is further empirically examined. None of the 39 pretreatment covariates used in this study shows significant differences across the weighted remediation treatment groups, suggesting that the four weighted treatment groups became comparable.

Finally, to estimate the postsecondary remediation effects on college attainments, a weighted regression model is specified:

$$Y_i = \beta_0 + \beta_1 Math_i + \beta_2 English_i + \beta_3 Both\_Subject_i + \varepsilon \quad (1)$$

where Y_i denotes the attainment outcomes (i.e., college persistence, transfer, and completion) for student i. $Math_i, English_i,$ and $Both\_Subject_i$ are dummy variables for whether student i enrolled in (a) only-math remediation, (b) only-English remediation, and (c) both-subject remediation (the omitted group is "no remediation") respectively. ε represents an error term. Since the dependent variables are dichotomous, Equation (1) is estimated with a linear probability model (LPM).[8] When examining the heterogeneous effects of postsecondary remediation by student subgroups, the above MMW-S procedures are replicated using subsamples restricted to a given racial or social subgroup (i.e., minority students in two-year institutions, low-socioeconomic students in four-year institutions).[9]

2.5 Results

2.5.1 The Impact of Postsecondary Remediation on Attainment Outcomes

The empirical findings in this study indicate that postsecondary remediation has differential effects on two- and four-year college students and the impacts vary across remediation enrollment patterns. Table 2.A2 displays results from linear probability models estimating the remediation effects on attainment outcomes, including college persistence, transfer, and completion that are measured at various time points. The left panel of Table 2.A2 presents the impact estimates of three remedial treatment groups (i.e., only-math, only-English, and both-subject remediation) for two-year college students, whereas the right panel of Table 2.A2 presents estimates for four-year college students.

[8] Additional analyses showed that the average marginal effects computed from logistics regression models are essentially very similar to those estimates from linear probability models.

[9] Across the eight subgroup analyses, the proportion of covariates remaining significantly different among the remediation treatment groups ranges from 0% to approximately 5%, after the steps of propensity score stratification and weighting. In theory, 5% of the covariates could show statistical imbalance among the treatment groups at the significant level of .05, even in a properly implemented experimental study.

For two-year college students, enrolling in any remediation generally has no impact on a variety of student outcomes as measured by college persistence, transfer, and degree receipt. The estimates are either essentially zero or smaller than their standard errors, with few notable exceptions. One is that participating in only-math remediation increased the probability of persisting into the second and third year of college by 12.2 and 12.9 percentage points respectively. The other is that when comparing differences in college completion between students who simultaneously enrolled in a math and an English remedial course and their non-remediated counterparts, the probability for both-subject remediation students to transfer upward to a four-year college in their third year increased by 5.7 percentage points and to obtain a bachelor degree within six years increased by 6.6 percentage points and by 6.8 percentage points within eight years.

Similar to the two-year college results, when comparing four-year college students who took any remediation with their non-remediated peers, there are no significant differences in the majority of college outcomes, specifically short-term outcomes including college persistence and transfer in second and third year of college. However, there is a clear pattern that remediation in math or in both subjects hinders students from attaining a bachelor degree. Enrolling in only-math remediation and both-subject decreased the likelihood for an undergraduate to graduate on time (within four years) by 9.6 and 14.9 percentage points respectively. When considering longer duration before graduation, the estimated probabilities of earning a four-year college degree within six and eight years for only-math remedial students are much smaller and not significant anymore, whereas for both-subject remedial students are about the same (a decrease by 14.1 and 14.4 percentage points respectively).

2.5.2 Sensitivity Analyses

2.5.2.1 Sensitivity to Sample Selection

To evaluate the robustness of the primary results, remediation impacts are re-estimated using several alternative samples. The main analytic sample in this study is restricted to students who enrolled in at least two courses but not more than seven courses during their first term of college. The sample involves both traditional and non-traditional college students, and oversampled minority students in NLSY97. To assess the sensitivity of the estimates to the sample selection, a series of models are estimated using subsamples that are (a) limited to students who were enrolled in at least three courses but not exceeding six courses in their first term of college, (b) traditional students who were aged 18-20, first-time, on-time (enrolled within a year from high school graduation), full-time college students, and (c) restricted to NLSY97 cross-sectional participants, excluding the oversampled minorities.

Appendix Table 2.B2 and 2.B3 show estimates from sensitivity tests for two- and four-year college analyses respectively. The results do not qualitatively alter the main conclusions. Two particular findings deserve mention. The estimated effects of only-math remediation on persistence for two-year college students (positive effects) and on on-time graduation for four-year college students (negative effects) are quite sensitive to choice of sample. The effects disappear in models using subsamples restricted to only traditional college students and they are occasionally not significant in the other two subsamples. Another observation is that the estimated effects of both-subject remediation on college completion, which for two-year college students are positive whereas for four-year college students are negative, are quite robust to a variety of sensitivity tests using different subsamples.

2.5.2.2 Quantifying the Robustness of MMW-S Inferences

A propensity score-based analysis such as MMW-S employed in this study can produce unbiased estimates of treatment effects, under the assumption of strongly ignorable treatment assignment, which requires all relevant covariates to have been measured (Rosenbaum & Rubin, 1983, 1984; Rubin, 2005). Although this analysis includes a rich set of individual and contextual measures that are correlated with remediation treatment conditions, there may be important unobserved confounding factors. To address this potential confounding issue, this study follows the sensitivity analysis procedures suggested by Frank, Maroulis, Duong, and Kelcey (2013) to quantify how much bias there must be in the estimates to invalidate my inferences, focusing only on the key findings—the estimated effects of both-subject remediation on college transfer (for two-year college students) and completion (for both two- and four-year college students). As defined by Frank et al. (2013), the calculation of proportion of bias to make an inference invalid is the following:

% bias necessary to invalidate an inference = 1 – threshold for inference/estimated effect, where the threshold for inference = s.e. × $t_{critical,df}$. Applied to the estimates of this study, to invalidate the inference of the effect of both-subject remediation on transferring upward to a four-year college in third year for two-year college students, bias must have accounted for (1 – 1.96 ×.024/.057 = .174) about 17% of the estimated effect. Similar calculations suggest that about 22-26% bias must be present to invalidate the inference of the estimated effect of both-subject remediation on obtaining a bachelor degree within six or eight years for two-year college students. Further, for the analysis of four-year college students, the bias necessary to invalidate the inferences of the estimated effect of both-subject remediation on obtaining a bachelor degree within four, six, or eight years is about 22-34%. According to Frank et al. (2013), the median

level of robustness is about 30% for observational studies in education. Thus, in this analysis the estimated impacts of both-subject remediation on college degree attainment for the four-year college students are quite robust while they are less so for the two-year college students.

2.5.3 Heterogeneous Effects of Remediation across Student Subgroups

Having used the full sample to identify whether and to what extent an association among various remediation conditions and college attainments exists, the second part of this analysis focuses on exploring the heterogeneous effects of postsecondary remediation across social subgroups. For racial subgroup analysis, white and nonwhite students are the two contrasting groups within two- and four-year colleges respectively. For socioeconomic subgroup analysis, this study divides students based on their parental education differently by college types. For two-year colleges, the comparison groups are students whose parent had no college education (or first generation college students) versus students whose parent had at least some college education (or non-first generation college students), whereas for four-year colleges, students whose parent had some college education or below versus students whose parent had a bachelor degree or above. In doing so, the socioeconomic subgroup comparison is between students who are pursuing more education than their parents, who likely lack the necessary college knowledge and experience, and their counterparts whose parents had at least the same level of higher education.

The subgroup analyses offer empirical evidence suggesting that postsecondary remediation has differential effects on white and nonwhite students, and on low-socioeconomic and high-socioeconomic students. The heterogeneous effects of postsecondary remediation also vary across two- and four-year institutions. The left and middle panel of Table 2.A3 and 2.A4 displays the estimated coefficients of heterogeneous impacts of remediation for two- and four-

year college students, respectively. Robust standard errors for each coefficient are reported in each corresponding cell in Appendix Table 2.B4 and 2.B5. Following Cohen (1983), *Z* tests of the differences between coefficients are conducted. The computed differences in remediation effects between subgroups are reported in the right panel of Table 2.A3 and 2.A4. Corresponding *Z*-scores are displayed in the right panel of Table 2.B4 and 2.B5 in Appendix.[10]

In two-year colleges, postsecondary remediation appears to help minority students in terms of short-term outcomes, specifically college persistence. The estimated positive effects of only-math remediation on persisting into the second or third year of college are largely concentrated among nonwhite students, while the effects were generally negligible for white students (the differences are 17.2 to 19.2 percentage points respectively and they are statistically significant at the critical level of 10%). On the other hand, when examining long-term attainment outcomes, there are no significant differences between white and nonwhite students; however, high-socioeconomic students benefited from remediation more than their low-socioeconomic counterparts. In particular, only-English and both-subject remediation significantly boosted the probability of obtaining a bachelor degree within six or eight years for non-first generation college students, but not for those first generation college students (the differences range from 10.5 to 14.8 percentage points and they are statistically significant at the critical level of 1%).

Not only in two-year colleges but also in four-year colleges does remediation affect students differently based on racial and socioeconomic status. In four-year colleges, taking only a remedial course in English significantly improved the likelihood of persisting into the third year of college for white students, while it had a negative marginal impact on nonwhite students (the difference is 17.4 percentage points). The racial difference in the effect of only-English

[10] In additional analyses, this study also tests if the remediation effects vary by racial and socioeconomic subgroups by adding interaction terms into the full estimation model. The results (not reported here) indicate that there are significant interaction effects between remediation and racial/socioeconomic groups.

remediation carried over into the longer term student outcomes measured by college completion within six and eight years (the differences are 24.7 and 22.8 percentage points respectively). In addition, the estimated negative impact of only-math remediation on on-time graduation is largely concentrated among undergraduates whose parental education level was less than a bachelor degree, whereas the effect is basically null for students whose parent had a bachelor degree or above (the difference is 17.3 percentage points). Taken together, the subgroup analyses provide evidence that remediation has differential effects for socially and economically advantaged and disadvantaged student groups.

2.6 Discussion

The goal of this study was to investigate whether and for whom postsecondary remediation has an impact on college outcomes. The analysis of the NLSY97 data arrives at conclusions that both validate and extend previous work on postsecondary remediation. As with most prior studies, this study finds insignificant effects of remediation on a series of short-term student outcomes, particularly college persistence and transfer. However, when examining long-run college attainment, two clear findings emerged. This study identifies particularly strong negative effects of postsecondary remediation on graduation for four-year college students who enrolled in both math and English remedial courses during their first term of college. On the other hand, both-subject remediation has a positive but modest impact on transferring up to a four-year college and eventually earning a bachelor degree for two-year college students although the estimated effects are less robust to potential omitted variable bias.

Why does enrolling in both math and English remedial course during first term of college positively affect two-year college students on college transfer and completion, but negatively affect graduation rates for four-year college students? One possible explanation is that the

negative function of dual remediation triggered by stigma might outweigh its positive developmental function in four-year colleges but not in two-year colleges. In four-year colleges, those students who simultaneously took a math and an English remedial course may feel more of a stigma as they are an outnumbered group. The NLSY97 data reveal that in four-year colleges both-subject remediation students account for only about 15.4% of the total student population whereas the number is almost doubled in two-year colleges (29.9%; see Table 2.A1).

One key difference between the findings of this analysis and those from prior subject-focused studies is that NLSY97 data indicate that remediation in a single subject, either in math or English, generally has no effect on a variety of college outcomes in terms of persistence, transfer, and completion measured at multiple points in time. Instead, enrolling in both math and English remedial course has a negative effect on degree receipt for four-year college students but a positive effect for two-year college students. Furthermore, the results of this study demonstrate that the effect of dual-subject remediation, which is a common remedial scenario among college freshmen, is not simply the sum of effects of individual remedial subjects. Thus, the person-focused approach employed in this study, which characterizes the remediation treatment groups based on both the subject areas and the number of remedial courses taken, yields empirical evidence that can better explain the remediation effects, especially when examining attainment outcomes of students. Such person-focused estimates are also more informative for designing student-centered educational programs for increasing college persistence and completion rates, specifically early interventions targeted at those four-year college entrants who take both math and English remedial courses during their first semester in college.

In terms of subgroup analysis, the findings on the heterogeneous effects of remediation in this study contribute to the understanding of the role of postsecondary remediation in educational

inequality. One major finding is that in two-year colleges only-math remediation could help minority students in continuing their postsecondary schooling careers in the short term. However, the positive effect of remediation on college persistence for minority students did not translate into reducing the racial gap in degree attainment in the long run. The NLSY97 data further uncover that two-year college students from socioeconomically privileged groups benefited the most from taking remedial courses. Enrolling in only-English or in both-subject remediation appear to increase likelihood of obtaining a bachelor degree for students whose parent had at least some college education, but not for their first-generation college peers.

In four-year colleges, while there is not a consistent pattern demonstrating that remediation benefits white or high-socioeconomic students, it appears to hinder nonwhite and low-socioeconomic students from completing college. Such negative effects of remediation on disadvantaged students, however, are not observed in two-year colleges. It might be the case that the relative rarity of remediation in four-year colleges, along with the over-representation of racial minority and low-income students in the remedial courses, increases the risk of stigma that elicits stereotype threat and belonging uncertainty for these underprivileged students placed in remediation.

Racial and socioeconomic gaps in college persistence and graduation have been a disturbing feature of the higher education system in the United States for decades. Policymakers and researchers have long been interested in identifying the sources of the differences in college persistence and completion between socially and economically advantaged and disadvantaged groups. This national study offers some suggestive evidence to explain part of the racial and socioeconomic gaps in college attainment, demonstrating that postsecondary remediation as an

institutional mechanism plays a significant role in reinforcing educational inequality and in perpetuating the social stratification in higher education.

The findings presented in this study have important substantive and methodological implications for research on postsecondary remediation, yet they must be interpreted with several limitations in mind. First, this study does not directly test the hypotheses concerning the remediation functions including cognitive development or stigma effect. Similar to most previous studies, the estimated effects of remediation in this study could be interpreted as joint effects of all possible combinations of remediation functions. Another limitation is that although the novel person-focused approach points to the importance of characterizing remediation patterns based on the subject areas and the number of remedial courses taken during the first term in college, the four remedial treatment groups in this study (i.e., only-math, only-English, both-subject, and no remediation) may not represent all possible remediation scenarios in higher education. Many students, especially those who are disadvantaged students attending a community college would need to take more than one semester to complete the sequence of required remedial courses (Attewell et al., 2006; Bailey et al., 2010; Parsad et al., 2003). Future research should extend the person-focused approach to examine the outcomes of students who enrolled in remediation in more than one semester.

An additional limitation is that with the observational data from NLSY97, this present research had to make the assumption of strongly ignorable treatment assignment. This analysis might not include all relevant pretreatment covariates to hold the assumption. There could be some confounding variables that potentially bias the findings. Nonetheless, by quantifying possible bias to invalidate an inference in this study, it is reasonable to conclude that the impact estimates of remediation on four-year college students are quite robust while they are less so on

two-year college students. A further limitation is that the two-year student sample in NLSY97 may be problematic in terms of representativeness as it appears that academically and socioeconomically advantaged students are overrepresented in the remedial groups. The ability to generalize from this two-year college student sample may be somewhat limited, however, the analytic strategy employed in this study, which can effectively mitigate selection bias by controlling for students' academic and family background, should yield estimates with strong internal validity that can be informative for theoretical and policy debates. To improve the external validity of remediation effects for two-year college students, future research efforts should replicate the current study using other nationally representative samples of two-year college students.

Despite these limitations, this analysis draws attention to the general and differential effects of postsecondary remediation on college attainment across college types and student subgroups. The present study broadens the operationalized definition of postsecondary remediation to include both the subject areas and the number of remedial courses enrolled. Drawing on recent data from NLSY97, this article provides updated national estimates on postsecondary remediation patterns, reporting that a considerable number of remedial students in both two- and four-year colleges enrolled in single as well as multiple remedial subjects during their first term in college. This study is also the first large-scale national analysis to offer impact estimates of single- and dual-subject remediation on postsecondary outcomes as measured by persistence, transfer, and completion at multiple time points. Finally, by presenting the heterogeneous effects of remediation for social subgroups, this analysis highlights that postsecondary remediation plays a critical role in reinforcing racial and social inequality in college attainment.

APPENDICES

APPENDIX A

TABLES FOR CHAPTER 2

Table 2.A1 Sample Means for Key Analysis Variables by Postsecondary Remediation Status

	Two-Year College			Four-Year College		
	All	Any Remediation	No Remediation	All	Any Remediation	No Remediation
Remediation						
Any remediation	.649	1.000	.000	.533	1.000	.000
Only math	.160	.246	.000	.131	.245	.000
Only English	.190	.293	.000	.249	.467	.000
Both math & English	.299	.461	.000	.154	.288	.000
College outcomes						
Persisted into the 2nd year	.614	.647	.555	.851	.851	.852
Persisted into the 3rd year	.413	.447	.351	.754	.761	.747
Transferred up in 2nd year	.064	.074	.046	--	--	--
Transferred up in 3rd year	.104	.130	.055	--	--	--
Transferred down in 2nd year	--	--	--	.084	.094	.073
Transferred down in 3rd year	--	--	--	.080	.098	.058
Earned AA within 3 years	.087	.092	.076	.038	.045	.030
Earned AA within 4 years	.123	.130	.110	.053	.065	.040
Earned BA within 4 years	.014	.012	.017	.331	.263	.408
Earned BA within 6 years	.081	.101	.045	.571	.532	.616
Earned BA within 8 years	.119	.144	.072	.619	.581	.663
Demographics						
Female	.498	.520	.456	.539	.538	.541
White	.650	.653	.646	.755	.708	.809
Black	.145	.149	.138	.113	.156	.063
Hispanic	.144	.134	.164	.082	.089	.074
Other race	.060	.064	.052	.051	.047	.054
Parent's highest education	13.41	13.63	13.00	14.81	14.48	15.19
	(3.44)	(3.32)	(3.62)	(3.41)	(3.45)	(3.33)
Cognitive score (ASVAB)	42.59	43.99	40.01	60.87	55.79	66.66
	(29.89)	(29.51)	(30.46)	(31.46)	(31.10)	(30.87)
High school math - High	.200	.222	.160	.522	.481	.569
Age when starting college	20.11	19.92	20.46	19.09	18.92	19.28
	(2.88)	(2.76)	(3.07)	(1.92)	(1.58)	(2.23)
On-time college attendance	.494	.561	.371	.802	.862	.733
Number of courses (1st-term)	3.74	3.92	3.42	4.77	4.97	4.54
	(1.33)	(1.28)	(1.37)	(1.33)	(1.24)	(1.40)
Number of students	1,144	732	412	1,560	862	698

Source. National Longitudinal Survey of Youth 1997 (NLSY97)

Note. AA = associate's degree; BA = bachelor's degree; AVSAB = Armed Services Vocational Aptitude Battery. Sample is restricted to first-time college students who had valid information on their college's institutional level. Sample size is 2,704. Data are weighted to be generalizable to the population of youth aged 12-16 in 1996 in the United States. Standard deviations appear in the parentheses below means of continuous variables.

Table 2.A2 Effects of Postsecondary Remediation for Two- and Four-year College Students

	Remediation Enrollment					
	Two-Year College			Four-Year College		
Dependent Variable	Only Math	Only English	Both Subjects	Only Math	Only English	Both Subjects
Persistence						
Persisted into the 2nd year	0.122 *	-0.057	0.022	-0.008	0.006	-0.026
	(0.049)	(0.050)	(0.042)	(0.035)	(0.026)	(0.041)
Persisted into the 3rd year	0.129 *	-0.013	0.079 †	0.036	0.043	-0.019
	(0.053)	(0.046)	(0.041)	(0.040)	(0.032)	(0.051)
Transferring						
Transferred up in 2nd year	-0.013	0.011	0.023	--	--	--
	(0.020)	(0.021)	(0.023)			
Transferred up in 3rd year	0.033	0.038 †	0.057 *	--	--	--
	(0.025)	(0.023)	(0.024)			
Transferred down in 2nd year	--	--	--	0.002	-0.001	0.003
				(0.021)	(0.019)	(0.036)
Transferred down in 3rd year	--	--	--	0.039	0.007	0.043
				(0.024)	(0.018)	(0.037)
College Attainment						
Earned AA within 3 years	-0.022	-0.001	-0.037	--	--	--
	(0.032)	(0.028)	(0.023)			
Earned AA within 4 years	-0.022	-0.030	-0.032	--	--	--
	(0.038)	(0.030)	(0.029)			
Earned BA within 4 years	--	--	--	-0.096 *	-0.034	-0.149 **
				(0.043)	(0.033)	(0.050)
Earned BA within 6 years	0.005	0.015	0.066 **	-0.031	-0.010	-0.141 *
	(0.026)	(0.021)	(0.025)	(0.047)	(0.037)	(0.056)
Earned BA within 8 years	0.015	0.018	0.068 *	-0.049	-0.010	-0.144 **
	(0.030)	(0.025)	(0.027)	(0.047)	(0.036)	(0.056)
Number of observations		1,091			1,451	

Note. AA = associate's degree; BA = bachelor's degree. Each cell in the table shows the estimate on a dummy variable indicating the effect of a specific type of remediation enrollment by subject, using "no remediation" as the reference group. Robust standard errors are reported in parentheses.

*** p<.001; ** p<.01; * p<.05; †p < .10 (two-tailed test).

Table 2.A3 Heterogeneous Effects of Postsecondary Remediation for Two-Year College Students

	(1)	(2)	(3)	(4)	(5)	(6)	(1) - (4)	(2) - (5)	(3) - (6)
	Only Math	Only English	Both Subjects	Only Math	Only English	Both Subjects	Only Math	Only English	Both Subjects
	Nonwhite			White			Difference between subgroups		
Persisted into the 2nd year	0.217 **	-0.003	0.028	0.045	0.021	0.048	0.172 †	-0.024	-0.020
Persisted into the 3rd year	0.217 **	0.047	0.075	0.025	-0.028	0.122 †	0.192 †	0.075	-0.047
Transferred up in 2nd year	0.024	0.011	0.014	-0.027	0.006	0.037	0.051	0.005	-0.023
Transferred up in 3rd year	0.047	0.024	0.024	0.081	0.056	0.091 †	-0.034	-0.032	-0.067
Earned AA within 3 years	0.039	0.041	0.025	-0.045	-0.053	-0.086*	0.084	0.094	0.111 *
Earned AA within 4 years	0.046	0.011	-0.008	-0.046	-0.063	-0.026	0.092	0.074	0.018
Earned BA within 6 years	0.018	0.023	0.085*	0.023	0.026	0.076 †	-0.005	-0.003	0.009
Earned BA within 8 years	0.024	0.004	0.073	0.027	0.049	0.091*	-0.003	-0.045	-0.018
Number of observations		525			457				
	Parents' highest education: HS or below			Parents' highest education: Above HS			Difference between subgroups		
Persisted into the 2nd year	0.073	-0.105	0.010	0.028	0.037	0.060	0.045	-0.142	-0.050
Persisted into the 3rd year	0.097	0.017	0.074	0.045	-0.056	0.042	0.052	0.073	0.032
Transferred up in 2nd year	-0.014	0.020	-0.000	-0.058 *	0.001	0.001	0.044	0.019	-0.001
Transferred up in 3rd year	0.010	-0.024	-0.020	-0.015	0.037	0.073 †	0.025	-0.061	-0.093 †
Earned AA within 3 years	-0.036	-0.041	-0.070 *	-0.039	0.019	-0.044	0.003	-0.060	-0.026
Earned AA within 4 years	-0.059	-0.063	-0.008	-0.014	-0.004	-0.048	-0.045	-0.059	0.040
Earned BA within 6 years	-0.015	-0.043	-0.016	0.036	0.062 *	0.114 ***	-0.051	-0.105 **	-0.130 **
Earned BA within 8 years	-0.011	-0.060 *	-0.026	0.048	0.080 *	0.122 ***	-0.059	-0.140 **	-0.148 **
Number of observations		449			533				

Note. AA = associate's degree; BA = bachelor's degree; HS = high school. "No remediation" is the reference group. Robust standard errors for each coefficient (in column [1] to [4]) and z-scores for each test of significant difference (in column [5] and [6]) are reported in Appendix Table 2.B4.
*** p<.001; ** p<.01; * p<.05; †p < .10 (two-tailed test).

Table 2.A4 Heterogeneous Effects of Postsecondary Remediation for Four-Year College Students

	(1)	(2)	(3)	(4)	(5)	(6)	(1) - (4)	(2) - (5)	(3) - (6)
	Only Math	Only English	Both Subjects	Only Math	Only English	Both Subjects	Only Math	Only English	Both Subjects
	Nonwhite			White			Difference between subgroups		
Persisted into the 2nd year	-0.066	-0.087	-0.025	-0.005	0.030	-0.080	-0.061	-0.117 †	0.055
Persisted into the 3rd year	-0.009	-0.085	-0.067	0.023	0.089 *	-0.015	-0.032	-0.174 *	-0.052
Transferred down in 2nd year	0.008	0.013	-0.027	-0.002	0.004	-0.030	0.010	0.009	0.003
Transferred down in 3rd year	-0.005	0.021	-0.017	0.059 †	0.012	0.039	-0.064	0.009	-0.056
Earned BA within 4 years	-0.091	-0.064	-0.089	-0.030	-0.009	-0.193 **	-0.061	-0.055	0.104
Earned BA within 6 years	-0.008	-0.160 *	-0.201 **	-0.020	0.087 *	-0.138 †	0.012	-0.247 **	-0.063
Earned BA within 8 years	0.006	-0.158 *	-0.189 *	-0.059	0.070 †	-0.141 †	0.065	-0.228 **	-0.048
Number of observations		469			824				
	Parents' highest education: Below BA			Parents' highest education: BA or above			Difference between subgroups		
Persisted into the 2nd year	-0.017	0.019	-0.069	0.015	-0.030	0.025	-0.032	0.049	-0.094
Persisted into the 3rd year	-0.018	0.027	-0.075	0.071	0.034	0.024	-0.089	-0.007	-0.099
Transferred down in 2nd year	0.034	0.036	-0.001	-0.009	-0.042	-0.038	0.043	0.078 †	0.037
Transferred down in 3rd year	0.062	0.039	0.015	0.100	-0.028	0.041	-0.038	0.067 †	-0.026
Earned BA within 4 years	-0.163 ***	-0.023	-0.131 *	0.010	-0.041	-0.213 **	-0.173 †	0.018	0.082
Earned BA within 6 years	-0.063	-0.057	-0.193 **	0.046	0.041	-0.213 *	-0.109	-0.098	0.020
Earned BA within 8 years	-0.075	-0.049	-0.175 **	0.042	0.001	-0.228 *	-0.117	-0.050	0.053
Number of observations		661			533				

Note. BA = bachelor's degree; HS = high school. "No remediation" is the reference group. Robust standard errors for each coefficient (in column [1] to [4]) and z-scores for each test of significant difference (in column [5] and [6]) are reported in Appendix Table 2.B5.
*** p<.001; ** p<.01; * p<.05; †p < .10 (two-tailed test).

APPENDIX B

SUPPLEMENTAL TABLES FOR CHAPTER 2

Table 2.B1 Descriptive Statistics of Covariates

Variables	Mean	Standard Deviation	Range
Demographics			
Female	.523	.500	0-1
White (base group)	.712	.453	0-1
Black	.126	.332	0-1
Hispanic	.107	.309	0-1
Other race	.054	.226	0-1
Age when starting college	19.501	2.410	16.000-31.333
Cognitive score (ASVAB)	62.121	25.712	.118-100
Parent's highest education (years)	14.554	2.823	1-20
Ratio of household income to poverty level	391.713	356.076	0-3227
Mother's age at 1st birth	24.149	4.646	11-44
Intact family (at age 17; dummy)	.588	.492	0-1
Census region of residence (at age 17)			
Northeastern (base group)	.171	.376	0-1
North Central	.283	.451	0-1
South	.334	.472	0-1
West	.212	.409	0-1
Pre-college schooling experience			
Ever repeated a grade (dummy)	.086	.281	0-1
Never late for school (in 1996; dummy)	.657	.475	0-1
Never absent from school (in fall 1996; dummy)	.212	.408	0-1
25% or less peers who cut classes or school (in 1997; dummy)	.639	.480	0-1
75% or more peers who plan to go to college (in 1997; dummy)	.627	.484	0-1
High school academic preparation			
School sector			
Public (base group)	.921	.270	0-1
Catholic	.047	.212	0-1
Other private	.032	.176	0-1
Participated in gifted course program	.262	.440	0-1
Academic program			
College preparatory	.452	.498	0-1
General education (base group)	.320	.467	0-1
Vocational education	.227	.419	0-1
Math pipeline – high level (dummy)	.492	.500	0-1
Science pipeline – high level (dummy)	.373	.484	0-1
Total academic credits	18.153	5.318	0-32
Overall grade point average (GPA)	3.019	.644	0-4.08
English GPA	2.885	.828	0-4.19
Math GPA	2.727	.808	0.4.08
College Attendance			
On-time college-going (dummy)	.728	.445	0-1
College sector			
Public (base group)	.770	.421	0-1
Private non-profit	.172	.377	0-1
Private for-profit	.056	.230	0-1

Table 2.B1 (Cont'd)

Variables	Mean	Standard Deviation	Range
College major			0-1
Science, technology, engineering, and mathematics (STEM)	.184	.388	
Social/Behavioral sciences	.139	.346	0-1
Art and humanities	.075	.263	0-1
Law and business	.137	.344	0-1
Education	.059	.236	0-1
Health sciences	.090	.286	0-1
Other vocational (e.g., mechanic, transportation, etc.)	.004	.063	0-1
Not declared yet (base group)	.312	.463	0-1
Number of courses taken in first term	4.355	1.425	2-7

Source. National Longitudinal Survey of Youth 1997 (NLSY97)

Note. n = sample size; AVSAB = Armed Services Vocational Aptitude Battery. Sample is restricted to first-time college students who had valid information on their college's institutional level. Data are weighted to be generalizable to the population of youth aged 12-16 in 1996 in the United States.

Table 2.B2 Sensitivity Tests of Estimated Effects of Postsecondary Remediation for Two-Year College Students

	Excluding students taking <2 or >6 courses in 1st-term			Traditional students only			Excluding oversampled minority students		
Dependent Variable	Only Math	Only English	Both Subjects	Only Math	Only English	Both Subjects	Only Math	Only English	Both Subjects
Persistence									
Persisted into the 2nd year	0.123 †	0.045	0.067	0.007	-0.145	-0.045	0.079	-0.070	0.024
	(0.064)	(0.070)	(0.052)	(0.095)	(0.097)	(0.080)	(0.064)	(0.058)	(0.049)
Persisted into the 3rd year	0.064	-0.020	0.062	-0.009	-0.058	0.029	0.151 *	-0.022	0.082 †
	(0.065)	(0.066)	(0.052)	(0.115)	(0.100)	(0.087)	(0.067)	(0.053)	(0.048)
Transferring									
Transferred up in 2nd year	-0.009	0.013	0.005	-0.033	0.007	0.011	-0.021	0.018	0.017
	(0.021)	(0.027)	(0.022)	(0.037)	(0.044)	(0.042)	(0.027)	(0.028)	(0.028)
Transferred up in 3rd year	0.050	0.022	0.056 *	0.104	0.061	0.042	0.060	0.050 †	0.076 *
	(0.036)	(0.026)	(0.027)	(0.071)	(0.053)	(0.045)	(0.038)	(0.029)	(0.030)
College Attainment									
Earned AA within 3 years	-0.037	-0.012	-0.023	-0.015	0.024	-0.035	-0.011	-0.018	-0.050 †
	(0.036)	(0.037)	(0.033)	(0.059)	(0.057)	(0.042)	(0.041)	(0.033)	(0.030)
Earned AA within 4 years	-0.048	-0.047	-0.027	-0.049	-0.031	-0.097	0.019	-0.028	-0.027
	(0.043)	(0.042)	(0.040)	(0.083)	(0.074)	(0.062)	(0.052)	(0.035)	(0.034)
Earned BA within 6 years	-0.001	0.006	0.074 *	0.019	0.072	0.090 *	0.020	0.015	0.066 *
	(0.027)	(0.026)	(0.030)	(0.036)	(0.047)	(0.039)	(0.035)	(0.026)	(0.030)
Earned BA within 8 years	0.027	0.014	0.076 *	0.059	0.062	0.112 *	0.029	0.036	0.068 *
	(0.035)	(0.031)	(0.032)	(0.057)	(0.049)	(0.044)	(0.040)	(0.033)	(0.033)
Number of observations		715			338			811	

Note. AA = associate's degree; BA = bachelor's degree. Each cell in the table shows the estimate on a dummy variable indicating the effect of a specific type of remediation enrollment by subject, using "no remediation" as the reference group. Robust standard errors are reported in parentheses.
*** p<.001; ** p<.01; * p<.05; †p < .10 (two-tailed test).

Table 2.B3 Sensitivity Tests of Estimated Effects of Postsecondary Remediation for Four-Year College Students

	Excluding students taking <2 or >6 courses in 1st-term			Traditional students only			Excluding oversampled minority students		
Dependent Variable	Only Math	Only English	Both Subjects	Only Math	Only English	Both Subjects	Only Math	Only English	Both Subjects
Persistence									
Persisted into the 2nd year	0.006	0.016	-0.022	0.010	0.002	-0.009	-0.026	0.006	-0.072
	(0.037)	(0.030)	(0.040)	(0.032)	(0.029)	(0.033)	(0.037)	(0.026)	(0.057)
Persisted into the 3rd year	0.008	0.024	0.012	0.039	0.043	-0.023	0.003	0.061 †	-0.063
	(0.046)	(0.037)	(0.046)	(0.043)	(0.037)	(0.048)	(0.044)	(0.033)	(0.069)
Transferring									
Transferred down in 2nd year	0.028	0.004	0.001	-0.011	-0.028	-0.030	-0.014	-0.006	-0.029
	(0.031)	(0.023)	(0.033)	(0.032)	(0.028)	(0.035)	(0.026)	(0.023)	(0.026)
Transferred down in 3rd year	0.059 †	0.008	0.052	0.049	0.005	0.025	0.033	0.001	0.020
	(0.031)	(0.021)	(0.034)	(0.032)	(0.025)	(0.034)	(0.029)	(0.021)	(0.029)
College Attainment									
Earned BA within 4 years	-0.055	-0.026	-0.157 ***	-0.032	0.017	-0.124 *	-0.087 †	-0.031	-0.223 ***
	(0.049)	(0.037)	(0.044)	(0.057)	(0.043)	(0.053)	(0.047)	(0.037)	(0.039)
Earned BA within 6 years	-0.039	-0.021	-0.170 **	-0.008	0.034	-0.142 *	-0.063	0.017	-0.219 ***
	(0.051)	(0.042)	(0.054)	(0.055)	(0.045)	(0.059)	(0.051)	(0.040)	(0.059)
Earned BA within 8 years	-0.050	-0.026	-0.155 **	-0.017	0.014	-0.122 *	-0.086 †	0.012	-0.223 ***
	(0.051)	(0.041)	(0.055)	(0.054)	(0.044)	(0.058)	(0.051)	(0.039)	(0.062)
Number of observations		1,102			906			1,187	

Note. BA = bachelor's degree. Each cell in the table shows the estimate on a dummy variable indicating the effect of a specific type of remediation enrollment by subject, using "no remediation" as the reference group. Robust standard errors are reported in parentheses.
*** p<.001; ** p<.01; * p<.05; †p < .10 (two-tailed test).

Table 2.B4 Robust Standard Errors for Heterogeneous Effects of Postsecondary Remediation and Z-Scores for Tests of Significant Difference: Two-Year College Students

	Robust Standard Errors						Z-Scores		
	(1)	(2)	(3)	(4)	(5)	(6)	(1) - (4)	(2) - (5)	(3) - (6)
	Only Math	Only English	Both Subjects	Only Math	Only English	Both Subjects	Only Math	Only English	Both Subjects
	Nonwhite			White			Difference between subgroups		
Persisted into the 2nd year	0.067	0.071	0.065	0.079	0.085	0.069	1.660	-0.217	-0.211
Persisted into the 3rd year	0.074	0.068	0.063	0.079	0.084	0.070	1.774	0.694	-0.499
Transferred up in 2nd year	0.036	0.032	0.032	0.030	0.033	0.047	1.088	0.109	-0.405
Transferred up in 3rd year	0.040	0.033	0.033	0.053	0.038	0.047	-0.512	-0.636	-1.167
Earned AA within 3 years	0.040	0.039	0.026	0.050	0.045	0.041	1.312	1.579	2.286
Earned AA within 4 years	0.050	0.044	0.032	0.053	0.047	0.053	1.263	1.149	0.291
Earned BA within 6 years	0.031	0.031	0.042	0.039	0.034	0.042	-0.100	-0.065	0.152
Earned BA within 8 years	0.038	0.035	0.046	0.043	0.042	0.046	-0.052	-0.823	-0.277
	Parents' highest education: HS or below			Parents' highest education: Above HS			Difference between subgroups		
Persisted into the 2nd year	0.082	0.080	0.074	0.082	0.074	0.061	0.388	-1.303	-0.521
Persisted into the 3rd year	0.080	0.072	0.075	0.082	0.073	0.063	0.454	0.712	0.327
Transferred up in 2nd year	0.027	0.051	0.026	0.029	0.035	0.038	1.110	0.307	-0.022
Transferred up in 3rd year	0.037	0.030	0.029	0.033	0.037	0.040	0.504	-1.281	-1.882
Earned AA within 3 years	0.047	0.039	0.032	0.045	0.050	0.039	0.046	-0.946	-0.515
Earned AA within 4 years	0.050	0.042	0.069	0.057	0.052	0.043	-0.593	-0.883	0.492
Earned BA within 6 years	0.042	0.027	0.031	0.029	0.028	0.034	-0.999	-2.699	-2.825
Earned BA within 8 years	0.046	0.029	0.034	0.035	0.035	0.037	-1.021	-3.080	-2.945

Note. AA = associate's degree; BA = bachelor's degree; HS = high school. "No remediation" is the reference group.

Table 2.B5 Robust Standard Errors for Heterogeneous Effects of Postsecondary Remediation and Z-Scores for Tests of Significant Difference: Four-Year College Students

	Robust Standard Errors						Z-Scores		
	(1)	(2)	(3)	(4)	(5)	(6)	(1) - (4)	(2) - (5)	(3) - (6)
	Only Math	Only English	Both Subjects	Only Math	Only English	Both Subjects	Only Math	Only English	Both Subjects
	Nonwhite			White			Difference between subgroups		
Persisted into the 2nd year	0.063	0.059	0.060	0.043	0.030	0.067	-0.800	-1.768	0.612
Persisted into the 3rd year	0.073	0.069	0.078	0.055	0.037	0.070	-0.350	-2.222	-0.496
Transferred down in 2nd year	0.038	0.036	0.027	0.030	0.027	0.027	0.207	0.200	0.079
Transferred down in 3rd year	0.044	0.046	0.040	0.033	0.022	0.031	-1.164	0.177	-1.107
Earned BA within 4 years	0.068	0.060	0.059	0.062	0.045	0.071	-0.663	-0.733	1.127
Earned BA within 6 years	0.081	0.071	0.072	0.062	0.044	0.077	0.118	-2.957	-0.598
Earned BA within 8 years	0.081	0.073	0.075	0.062	0.042	0.078	0.637	-2.707	-0.444
	Parents' highest education: Below BA			Parents' highest education: BA or above			Difference between subgroups		
Persisted into the 2nd year	0.056	0.044	0.063	0.031	0.035	0.026	-0.500	0.872	-1.379
Persisted into the 3rd year	0.064	0.051	0.068	0.046	0.044	0.070	-1.129	-0.104	-1.014
Transferred down in 2nd year	0.029	0.027	0.023	0.051	0.032	0.039	0.733	1.863	0.817
Transferred down in 3rd year	0.040	0.028	0.026	0.063	0.027	0.044	-0.509	1.722	-0.509
Earned BA within 4 years	0.046	0.048	0.055	0.077	0.059	0.079	-1.929	0.237	0.852
Earned BA within 6 years	0.065	0.055	0.062	0.065	0.056	0.102	-1.186	-1.249	0.168
Earned BA within 8 years	0.065	0.056	0.064	0.060	0.054	0.105	-1.323	-0.643	0.431

Note. AA = associate's degree; BA = bachelor's degree; HS = high school. "No remediation" is the reference group.

REFERENCES

REFERENCES

Adelman, C. (1999). *Answers in the tool box: Academic intensity, attendance patterns, and bachelor's degree attainment*. Washington, DC: US Department of Education.

Adelman, C. (2004). *Principal indicators of student academic histories in postsecondary education, 1972–2000*. Washington, DC: US Department of Education.

Amstrong, E. A., & Hamilton, L. T. (2013). *Paying for the party: How college maintains inequality*. Cambridge, MA: Harvard University Press.

Attewell, P., Lavin, D., Domina, T., & Levey, T. (2006). New evidence on college remediation. *Journal of Higher Education, 77*(5), 886-924.

Bahr, P. R. (2007). Double jeopardy: Testing the effects of multiple basic skill deficiencies on successful remediation. *Research in Higher Education, 48*(6), 695-725.

Bahr, P. R. (2010). Revisiting the efficacy of post-secondary remediation: The moderating effects of depth/breadth of deficiency. *Review of Higher Education, 33*(2), 177-205.

Bahr, P. R. (2013). The aftermath of remedial math: Investigating the low rate of certificate completion among remedial math students. *Research in Higher Education, 54*(2), 171-200.

Bailey, T., Jeong, D. W., & Cho, S. W. (2010). Referral, enrollment and completion in developmental education sequences in community colleges. *Economics of Education Review, 29*(2), 255-270.

Bettinger, E. P., Boatman, A., & Long, B. T. (2013). Student supports: Developmental education and other academic programs. *The Future of Children, 23*(1), 93-115.

Bettinger, E., & Long, B. T. (2004). *Shape up or ship out: The effects of remediation on students at four-year colleges.* (NBER Working Paper No. 10369). Cambridge, MA: National Bureau of Economic Research.

Bettinger, E., & Long, B. T. (2005). Remediation at the community college: Student participation and outcomes. *New Directions for Community Colleges, 129*, 17-26.

Bettinger, E., & Long, B. T. (2007). Institutional responses to reduce inequalities in college outcomes: Remedial and developmental courses in higher education. In S. Dickert-Conlin & R. Rubenstein (Eds.), *Economic inequality and higher education: Access, persistence, and success* (pp. 69-100). New York, NY: Russell Sage Foundation.

Bettinger, E., & Long, B. T. (2009). Addressing the needs of underprepared students in higher education: Does college remediation work? *Journal of Human Resources, 44*(3), 736-771.

Boatman, A., & Long, B. T. (2010). *Does remediation work for all students? How the effects of postsecondary remedial and developmental courses vary by level of academic preparation.* (NCPR Working Paper). New York, NY: National Center for Postsecondary Research.

Boylan, H. R., & Saxon, D. P. (1999). *What works in remediation: Lessons from 30 years of research*. Boone, NC: National Center for Developmental Education.

Calcagno, J. C., & Long, B. T. (2008). *The impact of postsecondary remediation using a regression discontinuity approach: Addressing endogenous sorting and noncompliance* (NBER Working Paper. No. 14194). Cambridge, MA: National Bureau of Economic Research.

Clotfelter, C. T., Ladd, H. F., Muschkin, C. G., & Vigdor, J. L. (2015). Developmental education in North Carolina community colleges. *Educational Evaluation and Policy Analysis, 37*(3), 354-375.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24,* 295-313.

Cohen, A. (1983). Comparing regression coefficients across subsamples: A study of the statistical test. *Sociological Methods & Research, 12*(1), 77-94.

Day, P. R., Jr., & McCabe, R. H. (1997). *Remedial education: A social and economic imperative*. Washington, DC: American Association of Community Colleges.

Deli-Amen, R., & Rosenbaum, J. E. (2002). The unintended consequences of stigma-free remediation. *Sociology of Education, 75*(3), 249-268.

Frank, K. A., Maroulis, S., Duong, M., & Kelcey, B. (2013). What would it take to change an inference? Using Rubin's Causal Model to interpret the robustness of causal inferences. *Education Evaluation and Policy Analysis, 35*(4), 437-460.

Gardner, J. N., Siegel, M. J., & Cutright, M. (2001). Focusing on the first-year student. *Priorities, 17*, 1-17.

GoIdrick-Rab, S., Carter, D. F., & Wagner, R. W. (2007). What higher education has to say about the transition to college. *Teachers College Record, 109*(10), 2444-2481.

Hong, G. (2010). Marginal mean weighting through stratification: Adjustment for selection bias in multilevel data. *Journal of Educational and Behavioral Statistics, 35*(5), 499-531.

Hong, G. (2012). Marginal mean weighting through stratification: A generalized method for evaluating multi-valued and multiple treatments with non-experimental data. *Psychological Methods, 17*(1), 44-60.

Lavin, D. E., Alba, R. D., & Silberstein, R. A. (1981). *Right versus privilege: The open-admissions experiment at the City University of New York*. New York, NY: Free Press.

Lavin, D. E., & Hyllegard, D. (1996). *Changing the odds: Open admissions and the life chances of the disadvantaged*. New Haven, CT: Yale University Press.

London, H. B. (1989). Breaking away: A study of first-generation college students and their families. *American Journal of Education, 97*(2), 144–170.

Martorell, P., & McFarlin, I. J. (2011). Help or hindrance? The effects of college remediation on academic and labor market outcomes. *Review of Economics and Statistics, 93*(2), 436-454.

Massey, D. S., & Fischer, M. J. (2005). Stereotype threat and academic performance: New findings from a racially diverse sample of college freshmen. *Du Bois Review, 2*(1), 45-67.

O'Hear, M. F., & MacDonald, R. B. (1995). A critical review of research in developmental education: Part I. *Journal of Developmental Education, 19*(2), 2-6.

Owens, J., & Lynch, S. M. (2012). Black and Hispanic immigrants' resilience against negative-ability racial stereotypes at selective colleges and universities in the United States. *Sociology of Education, 85*(4), 303-328.

Parsad, B., Lewis, L., & Greene, B. (2003). *Remedial education at degree-granting postsecondary institutions in Fall 2000* (NCES 2004-010). Washington, DC: US Department of Education.

Pascarella, E. T., Pierson, C. T., Wolniak, G. C., & Terenzini, P. T. (2004). First-generation college students: Additional evidence on college experiences and outcomes. *Journal of Higher Education, 75*(3), 249−284.

Radford, A.W., & Horn, L. (2012). *An overview of classes taken and credits earned by beginning postsecondary students. Web tables* (NCES 2013-151rev). Washington, DC: US Department of Education.

Robins, J. M. (1999). Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran & D. Berry (Eds.), *Statistical models in epidemiology, the environment, and clinical trials* (pp. 95-134). New York, NY: Springer.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association, 82*, 387-394.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika,70*(1), 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*, 516-524.

Roueche, J. E., & Roueche, S. D. (1999). *High stakes, high performance: Making remedial education work.* Washington, DC: American Association of Community Colleges.

Scott-Clayton, J., Crosta, P. M., & Belfield, C. R. (2014). Improving the targeting of treatment: Evidence from college remediation. *Educational Evaluation and Policy Analysis, 36*(3), 371-393.

Scott-Clayton, J., & Rodriguez, O. (2015). Development, discouragement, or diversion? New evidence on the effects of college remediation policy. *Education Finance and* Policy, 10(1), 4-45.

Sparks, D., Malkus, N. (2013). *First-year undergraduate remedial coursetaking: 1999–2000, 2003–04, 2007–08.* (NCES 2013-013). Washington, DC: US Department of Education.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797-811.

Walton, G. M. & Cohen, G. L. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology, 92*(1), 82-96.

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science, 331*, 1447-1451.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: MIT Press.

CHAPTER 3

THE IMPACT OF BEING LABELED AS A PERSISTENTLY LOWEST ACHIEVING SCHOOL: REGRESSION DISCONTINUITY EVIDENCE ON SCHOOL LABELING[11]

3.1 Introduction

Since the implementation of the *No Child Left Behind Act of 2001* (NCLB) both federal and state governments have employed various sanctions for schools that fail to meet academic standards. By 2012, thirty-two states had imposed specific sanctioning policies on low-performing schools regardless of their Title 1 status (National Center for Education Statistics [NCES] 2016a). One common practice among these sanctioning policies is identifying and labeling low-performing schools within the state. But not all low-performing school labels are equal. Some are merely cautionary labels such as being placed on a watch list. Others involve potential resource-based sanctions, such as withholding funds, assigning vouchers to students to exercise choice, or school closure (NCES 2016b). In this study we will estimate the differential effects of "non-consequential labeling" versus "consequential labeling" that has more immediate accountability pressures involving supervision, requirements, and possible reallocation of resources. Our goal is to identify those labeling practices (with or without implied consequences) that motivate schools to change, and in what areas.

Grading and labeling schools has been increasingly used as a means to hold schools accountable for student achievement. One underlying assumption of labeling and publicizing low-achieving schools is

[11] This research was is supported by the Hannah Chair Partnership (HCP), Michigan State University (MSU), Michigan Department of Education (MDE), and the Institute of Education Sciences (IES), U.S. Department of Education, through Grant No. #R305E100008, Principal Investigators listed in alphabetical order Susan Dynarski, Kenneth Frank, Tom Howell, Brian A. Jacob, Venessa Kessler, Joseph Martineau, and Barbara Schneider. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the organizations. Results, information and opinions solely represent the analysis, information and opinions of the author(s) and are not endorsed by, or reflect the views or positions of MSU, MDE, and IES or any employee thereof.

that schools would respond to the external pressures of accountability that stem from an undesirable label or stigma. It could also be that low-performing school labels may raise attention and community pressure that can be effective in creating organizational change within the schools and can lead to improvements in school performance. A small empirical literature has found the positive impact of school labeling on student academic achievement (e.g., Chakrabarti 2013a; Figlio and Rouse 2006; Winters and Cowen 2012). Yet, the question remains as to how schools would respond to low-performing labels that attract public attention and have immediate accountability requirements compared to those that are relatively minor such as being put on a watch list with little attention and no direct punitive consequences.

In this study, we use state-wide data from Michigan at the school-level to analyze the effects of being identified as a persistently lowest achieving (PLA) school or on a cautionary *watch list* for PLA in 2010. The PLA list is a consequential school label accompanied by accountability requirements and public attention, whereas the watch list is a non-consequential school label with no sanction threats or even media coverage. To test the effects of being on a PLA or watch list, we employ a sharp regression discontinuity (RD) design which allows us to explore the discontinuity around the threshold of being placed on one of the lists (e.g., being on the PLA list because of scoring in the bottom 5 percent of schools). Thus, by using statewide achievement-based school percentile rankings, we leverage the RD design to obtain unbiased and precise measures of the treatments (being placed on a list).

With the introduction of the federal School Improvement Grant (SIG) program in 2009, all 50 states and the District of Columbia started to identify and publicize the lowest achieving 5 percent schools annually (Hurlburt et al. 2011; U.S. Department of Education 2015).[12] Hence, our empirical findings on the PLA list in Michigan suggest broad and timely applications for other school labeling or sanction systems. Our RD analyses indicate that the bottom 5 percent of schools on the PLA list increased their student achievement scores in writing, with marginal evidence in mathematics and social studies, and no

[12] The US Department of Education requires states to use three common criteria to determine the persistently lowest-performing schools: (a) a school's overall academic achievement level, (b) whether there is a "lack of progress" in the school, and (3) for high schools, whether the school has a graduation rate below 60 percent (U.S. Department of Education, 2009). Within each of these criteria, however, there is variation across states (Hurlburt et al. 2011).

evidence of increases in reading and science in year one, compared with those schools which were just above the cutoff for being on the list. The positive effect for writing is robust with respect to alternative specification and potential bias in estimation. We find no impact on student achievement across academic subjects for those bottom 6 to 20 percent schools labeled as a watch list school receiving no actual sanctions (compared to the bottom 21 to 35 percent of schools). Our findings suggest that schools are likely to respond differently to varying forms of low-performing labeling, depending on the accountability pressure and social stigmatization process.

3.2 Background

Labeling has been used for about twenty years to motivate low-performing schools to make substantive academic improvements. For example, in Texas schools were categorized as either "exemplary," "recognized," "acceptable" or "low-performing," while Florida uses an "A" through "F" grading scheme (Florida Department of Education 2015; Texas Education Agency 2015). These rating categories carried no immediate punitive actions as they were simply cautionary labels. However, the NCLB Act in 2002 took a dramatically different path with respect to sanctions by penalizing those schools labeled as failing (NCLB 2002). Beginning with NCLB, all states were required to conduct annual assessments and turn the results of these assessments into a label of "made AYP" or "did not make AYP." Schools failing to make AYP for multiple years were subject to a set of penalties, such as offering students the choice to leave their school and free transportation to another public school. In the past few years, the NCLB criteria have blended with states' identification of schools as in the lowest 5 percent. For example, states seeking waivers from certain federal sanctions were required to identify the lowest 5 percent of schools (labeled Priority Schools) in their state (U.S. Department of Education 2012).

The motivation for identifying and publicizing low-performing schools is based in part on the notion that schools may respond to the accountability pressures or stigma of being publicly labeled as low-performing (Chakrabarti 2013a; Chiang 2009; Figlio and Rouse 2006; Mintrop 2004). Through accountability requirements, supervision, and monitoring, consequential labeling could force low-achieving schools to reallocate their organizational and instructional resources to improve in order to

avoid further more severe sanctions. Labeling could also press schools to make changes by publicly releasing results and using transparency to elicit both social stigma and community pressures. Individuals (e.g., educators, parents) attached to schools being identified as low-performing might be very embarrassed by the labels and will therefore attempt to improve. It is also often assumed that parents and teachers in schools identified as low-performing or failing organization would be more likely to transfer, especially if they did not have a reasonable voice in making improvements (Hirschman 1970). Therefore, the threats of accountability sanctions and community pressures, either hypothetical or actualized, should motivate schools to improve achievement to reach a certain level of performance.

Correspondingly, some researchers have found positive effects for school labels on the improvement of students' achievement after schools were graded as failing. For example, in Florida, Figlio and Rouse (2006) and Rouse, Hannaway, Goldhaber, and Figlio (2013) found that elementary schools receiving a grade of "F" improved their student performance immediately in the following year, on math and reading tests. They argue that schools actually changed in part to the increasing stigma of having received a failing grade. Other studies analyzing either school or student level data in Florida have also documented similar positive effects on student achievement in high and low stakes subjects including reading, mathematics, writing, and science for "F-rated" elementary schools (see Chakrabarti 2013a; Chiang 2009; Winters, Trivitt, and Greene 2010). Using elementary and middle school student data from New York City, which also adopted school letter grading, Winters and Cowen (2012) find that schools that received "F" labels showed positive achievement improvement in reading and math, especially for those students in the bottom quartile.

While many studies suggest that school labeling policy led to increased academic achievement for students across varied subjects, a growing body of evidence reveals that some of this positive effect may be spurious. One reason is that schools can boost test scores by deliberately and systematically manipulating the population of students taking high-stake tests (Chakrabarti 2013b; Cullen and Reback 2006; Figlio and Getzer 2006; Haney 2000; Heilig and Darling-Hammond 2008). Heilig and Darling-Hammond (2008), for instance, find that Texas schools strategically increased grade retention rates in the

9th grade and/or excluded more low-achieving students from taking the high-stakes state assessment in 10th grade. Figlio and Getzer (2006) also provide some evidence from Florida showing that schools tended to reclassify low-income and low-performing students as disabled or reassign them to special education categories that were exempt from the accountability system at the time.[13]

In summary, despite a series of studies that have detected some positive relationship between school labeling policies and student performance, there are several limitations with these prior empirical studies which can be addressed by studying the persistently lowest achieving (PLA) designation in Michigan. First, in the above studies, researchers could not determine whether understand whether it is stigma or threat of sanction behind the school labeling process that serves as a major driving force to make school change. In this study, we are able to compare the effects of two types of school labels (e.g., PLA and watch list) and determine the extent to which schools respond to different forms of labeling (i.e., consequential labeling vs. non-consequential labeling) with varying levels of accountability pressure and social stigmatization process. Second, while a majority of previous studies focus on elementary and middle schools, this analysis is targeted at a statewide sample of traditional high schools, which are generally considered to face more challenges in terms of turning around or restructuring (Louis et al. 2010). Third, prior studies tend to examine school labels created by state governments or local school districts. This study is one of the first to offer evidence assessing assignment to a PLA list based on the criteria suggested by the U.S. Department of Education through the SIG program in 2009 (U.S. Department of Education 2009).

3.3 Persistently Lowest Achieving Schools in Michigan

Starting in 2010 the Michigan Department of Education (MDE), like many other states, annually published a Top-to-Bottom (TTB) school ranking list. The state-wide TTB ranking is calculated by incorporating average achievement levels and improvement rates both in mathematics and reading over the past four years (from 2005-2006 to 2008-2009) for both elementary and secondary schools. Schools

[13] Similar trends of increasing in the incidence of grade retention and reclassification of students as disabled were also documented in New York City (Allington and McGill-Franzen 1992) and Chicago (Jacob 2005) when a high-stakes testing accountability system was mandated.

that fall in the bottom 5 percent are classified as Persistently Lowest-Achieving (PLA) schools.[14] The PLA policy was adopted by the state of Michigan in December 2009 to identify schools eligible for the federal incentive grant program School Improvement Grants (SIG) and therefore had to follow specific ranking criteria for the grants developed by the U.S. Department of Education (Hurlburt et al. 2011; The Revised School Code Act 2009).[15]

There are two major criteria for a school to receive a PLA designation, including: (1) having at least 30 students (enrolled over the entire academic year) with math and reading scores for two prior years, and (2) eligibility to receive Title I aid. Schools eligible for PLA status consisted mainly of schools receiving Title I funding and in the state of "in improvement, corrective action, or restructuring" under NCLB (i.e., "Tier 1" schools), and middle/high schools eligible for but not receiving Title I funding (i.e., "Tier 2" schools). In August 2010, 92 elementary, middle, and high schools were notified that they were placed on the 2010 PLA list in Michigan, of which we will focus on the high school sample (n=56), where comparable student test scores are available in the following year.[16]

Once placed on the PLA list, schools immediately started to face a series of accountability and community pressures. The names of the PLA schools were publicly disclosed by local media. All PLA schools were required to issue a notification letter to parents of students, explaining their PLA status. They were placed under the supervision of the State School Reform/Redesign Office (SRO) at MDE and were required to develop and implement a three-year reform plan that aimed to rapidly increase student achievement. To monitor the progress of the PLA schools, the SRO developed a school performance matrix which includes improvements in all five academic subjects (i.e., reading, writing, mathematics,

[14] The PLA schools are now termed as Priority Schools under the Elementary and Secondary Education Act (ESEA) Flexibility (U.S. Department of Education 2012).

[15] Michigan's Adequate Yearly Progress (AYP) report cards, which were required under NCLB, were continued in force during the first few years of the implementation of the PLA policy until being replaced by the Michigan School Accountability Scorecards, beginning in 2013.

[16] We chose not to include elementary and middle school sample in this study for several reasons. The first is that student achievement at different school levels (i.e., elementary, middle, and high schools) is measured by different assessment systems. Thus those test scores are not comparable across school levels. The second is that in total only 29 elementary and middle schools have outcome measures in the following year. The sample size is too small for a separate supplementary analysis.

science, and social studies) tested in annual statewide assessments.[17] Those schools not making satisfactory progress over the course of the period from 2012 to 2014 were potentially subject to be taken over by the SRO, meaning that the local community would lose governance of the community school identified as PLA (The Revised School Code Act 2009).

Along with the PLA list, the Michigan Department of Education also created a state *watch list* of schools in the bottom 6 to 20 percent, which were identified as being in danger of becoming PLA schools in the future.[18] In total, there were 60 high schools placed on the watch list in 2010. The watch list does not affect the PLA ranking; however, it provided an alert to the schools of their potential to fall into the PLA category. There were no sanctions imposed on these watch list schools (the MDE has since stopped releasing new watch lists). Without a real threat of sanctions, watch list schools may not be as responsive as those schools designated as being on the consequential PLA list. Nonetheless, those schools being identified as watch list schools still received a low-performing label. Studying solely the consequential school labeling, as in most prior studies, cannot separate out the effect of labeling and the effect of sanction threats. In this study, we can simultaneously analyze and compare the PLA and watch list in Michigan and make a theoretical contribution to school labeling literature by shedding some light on which mechanism, specifically stigma or sanction threats, is responsible for motivating low-performing schools to react.

3.4 Methodological Approach

3.4.1 Sharp Regression Discontinuity Design

To identify the causal impact of being on the PLA or watch list in Michigan we use regression discontinuity design, a commonly used quasi-experimental approach in prior studies on school labeling (e.g., Chakrabarti 2013; Chiang 2009; Winters and Cowen 2012). We estimate effects of being on a low-

[17] The performance matrix also involves instructional time, teacher performance level, student attendance, discipline incidents, course completion, dropout rate, graduation rate, and other school indicators that account for variability in each PLA school's individual plan.

[18] To the best of our knowledge, not many states created a watch list of schools in the bottom 6 to 20 percent, in addition to the PLA list of the bottom 5 percent schools. One exception is the state of Washington which initiated a program called "Struggling School Innovation Cluster" to supporting schools in the bottom 6 to10 percent in 2010. (https://www2.ed.gov/programs/racetothetop/phase2-applications/washington.pdf).

performing list by examining the difference in student achievement at the school level between schools just below and above a fixed threshold, (the fifth percentile for the PLA list, and the 20th percentile for the watch list). The assignment to either list is determined by the value of the percentile ranking on either side of a single cutoff point of perfect treatment compliance (the probability of being placed on a list jumps from 0 to 1 at the cut score). Separate regressions are estimated for the two treatments using sharp RD designs (Hahn, Todd, and van der Klaauw 2001; Trochim 1984) and following the procedures laid out by Imbens and Lemiuex (2008).

The model specification for estimating the effects of PLA treatment is:

$$Y_i^{2011} = \beta_0 + \beta_1 List_i^{2010} + f(PctRank_i^{2010}) + \beta_2 Y_i^{2009} + \Gamma' \boldsymbol{X}_i^{2010} + \varepsilon_i \qquad (1)$$

where Y_i^{2011} is the school level student achievement in a given subject for school *i* in year 2011, $List_i^{2010}$ is a dummy variable for whether school *i* was placed on the PLA or watch list in year 2010 (with β_1 the effect of being on the list), $PctRank_i^{2010}$ is the Top-to-Bottom percentile rankings rated for school *i* in year 2010 (which serves as the forcing variable in the RD designs), Y_i^{2009} is the average test score for each corresponding dependent variable in 2009 (the latest school performance measures used for the 2010 percentile ranking calculation), and $\boldsymbol{X}_i^{2010}$ is a vector of selected school characteristics obtained in year 2010.

One of the most critical aspects of the RD modeling is the functional form specification of the relationship between the forcing variable (i.e., $PctRank_i^{2010}$), and outcome variable (Schochet et al. 2010). Using an incorrect functional form in RD designs typically biases the estimate of the treatment effect. Following the strategy of selecting the most appropriate function form(s) suggested by Lee and Lemieux (2010), we test a variety of functional forms by including quadratic and cubic terms for $PctRank_i^{2010}$ as well as interaction terms between $PctRank_i^{2010}$ and treatment assignment into equation (1) to determine which best fits the data. These alternative specifications do not lead to improvements in goodness of model fit based on the results of *F*-statistics = $\frac{(R_u^2 - R_r^2)/K}{(1-R_u^2)/(n-K-1)}$, where R_u^2 and R_r^2 are *R*-squared values from the unrestricted and restricted regression, n is the total number of observations, and *K*

is the number of new added indicators, including interaction and polynomial terms. Therefore, we only present estimation results from the more parsimonious linear specification (i.e., equation [1]).

Consistent with the RD designs, only school samples within a specific bandwidth of the treatment cutoff are included in the estimation. When using an RD estimator (in our case, β_1), researchers can compare observations that are extremely close to the cutoff, thereby approximating experimental conditions. However, without a large sample around the cutoff, the RD estimators will be imprecise. On the other hand, including samples that are far away from the cutoff may produce biased or inconsistent RD estimators if the functional forms modeling the relationship between forcing variable and outcomes are not correctly specified. In this study, based on our analysis of statistical power (see Appendix C) and functional form specifications, we find that including schools within 10 percentile points of the threshold (36 PLA schools and 32 comparison schools) optimizes the bias-precision trade-off. In particular, our sample yields sufficient statistical power (80 percent power for two-sided tests at the 0.05 significance level, as recommended by Schochet [2009] and Bloom [2012]) to detect the effects of PLA list while producing consistent RD estimators and constraining the sample to be close enough to the cutoff to establish comparability between the schools above and below the cutoff. In this paper, we report the results from estimating RD models that include school samples within 9, 10, or 11 percentile ranks of the cut score.

In sharp RD designs, inferring a causal impact on an outcome relies on several fundamental assumptions including: (1) the jump at the cut score is truly discontinuous; (2) the forcing variable is observed with random measurement error; (3) the dependent measure, in this case the achievement score, is a continuous function of the forcing variable (percentile ranking) at the cutoff in the absence of treatment, and (4) the treatment units are sorted unconditionally by assignment. We conduct a series of tests regarding violations of the above assumptions, recommended by Imbens and Lemieux (2008), including unconfoundedness, no-manipulation, and no jumps at non-discontinuity points. Results from testing these identification assumptions are presented in Appendix A and generally support the use of the sharp regression discontinuity design with our data and specification described in equation (1).

To assess the robustness of our primary RD estimation results, we employ several sensitivity analyses. First, we test the sensitivity of our results to bandwidth choices, the width of the "window" of cases used for defining comparable treatment and control units, as recommended by Imbens and Lemiuex (2008). Second, following Frank, Maroulis, Duong and Kelcey (2013), we quantify the bias necessary to invalidate our RD inferences in terms of sample replacement. Given the targeted population in this study, we address concerns about the generality of estimated RD effects by calculating the proportion of the schools in our sample that would need to be replaced to invalidate our inferences.

Third, we compare the 2010 PLA list effects to estimated results of a pseudo 2009 PLA list which is constructed by using data from the school years prior to the state mandated assignment for low-performing schools to a PLA list. Similar falsification tests or placebo tests are often used in policy analysis, especially when evaluating intervention programs which reward or penalize schools based on students' average performance (e.g., Chiang 2009). By applying the same estimation models to a historical counterfactual school sample before the PLA list was implemented, we expected to verify that there would be no effect of "the list" on achievement outcomes, since none of the schools in this pseudo treatment group (bottom 5 percent in 2009) would have experienced labeling and threat of sanctions from being on a "persistently lowest achieving" school list.

3.4.2 Data and Measures

We use school-level data provided by the Michigan Department of Education to examine the effect of the 2010 PLA list on school outcomes measured in 2011.[19] The school data contain all criteria used for determining the 2010 PLA and watch lists, including information on percent proficient in mathematics and reading in statewide high school examinations (Michigan Merit Examination, MME) for the past four years (from 2006 to 2009), number of students tested in math and reading for the past four

[19] Prior empirical studies evaluating the effect of school labeling on student achievement have used both school-level (e.g., Chakrabarti 2013; Figlio and Rouse 2006) and student-level data (Chiang 2009; Rouse et al. 2013; Winters and Cowen 2012; Winters et al. 2010). Although there has been a continuing debate about the adequacies of school-level data (as opposed to individual-level data) for assessing school effects (Henig 2008), school-level data have been empirically proven to be adequate for evaluating the impact of school-based interventions (e.g., Jacob, Goddard, and Kim 2014; Stuart 2007).

years, calculated four-year improvement slopes in math and reading, graduation rates for the most recent three years, Tier 1 and Tier 2 status, TTB (Top-to-Bottom) percentile rank, Tier 1 and Tier 2 percentile rank, and whether the school was placed on the PLA or watch list in 2010.

The 2010 TTB (Top-to-Bottom) percentile ranking, which is used for determining PLA and watch lists, serves as the forcing variable in our RD estimation models. While all schools were ranked on the TTB list, schools in the Tier 1 and Tier 2 pool received an additional *within-pool* percentile ranking. The within-pool rankings were created to ensure that those schools in the bottom 5 percent on both within-pool rankings were placed on the PLA list. The bottom 5 percent cutoffs on the Tier 1 and Tier 2 ranking correspond to the percentile rank of 0.5 and 11.4 respectively on the TTB list. As the threshold of 11.4 on the TTB percentile ranking is higher than that of 0.5, thus it is used as the PLA cutoff. All PLA eligible schools both from the Tier 1 and Tier 2 pools that ranked equal or lower than 11.4 percentile on the TTB list (PLA cutoff) were identified as PLA schools, as shown in column (a) in Table 3.B1.

A similar rescaling was computed to identify the watch list. The bottom 20 percent cutoff on the Tier 2 ranking has a corresponding percentile rank of 28.1 on the TTB list. Therefore, PLA eligible schools that have a TTB rank equal or lower than 28.1 percentile (watch list cutoff) were identified as watch list schools. It is important to note that the percentile rank of 11.4 and 28.1 on the 2010 TTB ranking (forcing variable) respectively serve as the "bottom 5 percent" and "bottom 20 percent" cutoff for determining PLA and watch list schools in Michigan in 2010. The total number of high schools placed on the PLA (n=56) and watch list (n=60) in 2010, and of the final analytic sample used in this study are respectively reported in column (b) and (d) in Table 3.B1. The relationship between the TTB ranking (forcing variable) and school outcomes is assumed to be linear and smooth, hence any discontinuity of the conditional distribution of the school outcomes as a function of the percentile ranking at the cut score is considered evidence of a causal effect of being on the PLA or watch list.

The major goal of this study is to examine the differential effects of consequential labeling (i.e., PLA list) versus non-consequential labeling (i.e., watch list), not the impact of school reform/redesign plans in PLA schools. Therefore, we use the outcome variables that are drawn from student achievement

on the Michigan Merit Examination (MME) assessed at the end of the 2010-2011 school year, prior to the implementation of PLA school reform/redesign plans. Based on Michigan high school curriculum standards, the MME is administered annually in the spring to high school juniors. Five subjects are tested, including reading, writing, mathematics, science, and social studies. Student performance falls into one of four categories: advanced, proficient, partially proficient, and not proficient. Students who score either "proficient" or "advanced" are considered as having met the proficiency level in a specific subject. The school level *percent of students who met the proficiency level* in reading and mathematics are the two critical components used in computing state-wide school percentile rankings in 2010. However, starting in 2011, the *average of students' scale scores* in all five subjects was used to calculate the new state-level percentile rankings in the coming years. Therefore, to carefully examine the impact of the PLA and watch list on student outcomes, we use students' scale scores in the five tested academic subjects in 2011 as the primary dependent variables. Furthermore, recognizing uncertainty about the equating across years, we present both percent of students who exceeded proficiency levels and average of the test scores in each subject.

To address the limitation of the small sample size in our study, we include several school covariates in our RD models to mitigate the small sample size biases (Imbens and Lemieux 2008) and to improve the precision of our estimation of the effects of being on the PLA and watch lists (Schochet 2009; Wing and Cook 2013). The school level measures include percentage of free/reduced lunch students, percent of minority students, school size, and pupil-teacher ratio. These school variables are obtained from the Common Core Data (CCD) provided by the U.S. Department of Education's National Center for Education Statistics (NCES). They are collected in the 2009-2010 academic year. Additionally, we include 2009 measures for each school outcome, which are the latest school performance used in the calculation of the 2010 TTB percentile ranking, as pretreatment covariates in our models. Together with the percentile ranking as the forcing variable, these covariates are assumed and also have been empirically tested not to be influenced by the treatment (see the results of the unconfoundedness assumption testing in Appendix A). By adding these school covariates, the R-squared values of our RD models for each

outcome are relatively high, ranging from 0.70 to 0.88, which yield sufficient statistical power (80 percent power for two-sided tests at the 0.05 significance level) for detecting treatment effects in school-based RD designs, as recommended by Schochet (2009). We report results of statistical power analysis for this study in Appendix C.

Table 3.B2 reports descriptive statistics of the school outcome variables and other covariates for the PLA list, watch list, and no designation school samples. Of the five subjects, over the years and across school categories (i.e., PLA, watch, and no designation), reading has the highest percentage of students who were at least proficient, writing the second highest, followed by social studies, mathematics, and science. On average, schools across the three categories have little difference in student population and pupil teacher ratio. Lower ranked schools, however, have a higher percentage of students who are minorities and from low-income families.

3.5 Results

3.5.1 Effects of Being on the 2010 PLA List

Before estimating our RD models, we present graphical evidence on the relationship between the forcing variable of percentile ranking and school-level performance for all five subjects in 2011 (i.e., reading, writing, mathematics, science, and social studies) for both (1) the percent of students who met proficiency level (see Figure 3.A1), and (2) the average of students' scale scores (see Figure 3.A2). The regression line on the left of the cutoff in each figure represents PLA schools that fall below the threshold of 11.4 percentile rank whereas the regression line on the right of the cutoff in each figure represents watch list schools that fall above the threshold. As shown in Figures 3.A1 and 3.A2, for example, there is a jump at the cutoff in writing, implying that schools placed on the 2010 PLA list increased their student performance in writing in 2011.

To quantify the magnitude and significance of the discontinuities in school outcomes due to being on either list, we estimate parametric RD models specified in equation (1) for subject-specific outcomes.

The estimated PLA list effects are reported in Table 3.B3.[20] Panel A in Table 3.B3 displays the estimated PLA effects on the percent of students who met proficiency at the school level in the five subjects (within 9, 10, or 11 percentile rankings on either side of the PLA cutoff) whereas the lower panel presents the effects on the average of students' scale scores. The positive effect of the 2010 PLA list on writing is notably strong, about 7.9 to 9.4 points (corresponding to 0.53-0.63 standard deviations), for the 2011 school outcome as measured by the average of students' scale scores (statistically significant at the critical level of 1 percent regardless of selection of bandwidth). The estimated effects are weaker for the percent of students who met proficiency level for writing.

Additionally, our RD estimation results indicate that the PLA list has some marginal positive effects on school performance in mathematics and social studies. As reported in Panel A in Table 3.B3, being on the 2010 PLA list increases the percent of students who met proficiency level in social studies by about 5.8 to 6.2 percentage points (statistically significant at the critical level of 5 percent for the bandwidth of 9 and 10 percentile rankings). Furthermore, as shown in Panel B in Table 3.B3, schools on the 2010 PLA list raise the average of students' scale scores in social studies in 2011 by about 4.3 points (corresponding to 0.39 standard deviations; statistically significant at the critical level of 5 percent for the bandwidth of 9 percentile ranking). We also observe that the PLA list has some marginal positive effects on the average of students' scale scores in mathematics. However, the magnitudes are small and they are only statistically significant at the critical level of 10 percent for the bandwidth of 9 and 10 percentile rankings.[21] The effects on mathematics and social studies are weak and we are careful not to over-interpret these results as they are generally sensitive to the specifications based on different choices of bandwidth (checks for robustness described in the next section).

[20] Estimates on the forcing variable and other covariates are not shown in the table. They are all in the expected directions (coefficients available upon request).

[21] Given the schools were drawn from two separated tiers, we examine whether the PLA list effects in writing, mathematics, and social studies varied by tier by adding interaction terms between Tier 2 status and PLA status. The results show that there is no systematic pattern relating tier status to the PLA effects.

3.5.2 Effects of Being on the 2010 Watch List

The above RD results using first-year outcomes of student achievement suggest that schools placed on the PLA list improve student performance on writing scale scores even before a formal implementation of a reform plan took place. However, the precise interpretation of the findings proves challenging as we cannot distinguish whether the positive effects are due to the practice of stigmatized labeling, threat of sanctions, or both, faced by these schools. Nonetheless, taking advantage of the unique context in Michigan in which two forms of low-performing school lists were announced at the same time, we are able to examine the differential effects of labeling with varying amounts of accountability pressures and social stigma. Specifically, we estimate and compare the causal impact of the non-consequential watch list, which did not carry any sanctions and had less media attention, to that of the consequential PLA list.

Table 3.B4 presents the results from estimating equation (1) using the school samples around the 2010 watch list cutoff. We find no evidence of an effect of being placed on a watch list versus school neither on the watch or PLA list (within the specified bandwidths). Magnitudes of all estimates are modest (or much smaller than those PLA list estimates in Table 3.B4) and none of them reach the level of statistical significance.[22]

To determine whether the positive PLA list effects are significantly larger than watch list effects, we perform a series of Wald tests. In particular, we test the hypothesis that the two coefficients for a particular outcome, reported in Tables 3.B3 and 3.B4 respectively, are equal. We focus only on those significant RD impact estimates of the PLA list (within bandwidth of 10), specifically on the writing and social studies. For average of students' scale score in writing (PLA effect=8.706; watch list effect=-0.305), the results yield a chi-square value (χ^2) of 6.42 (degree of freedom=1) and a p-value of 0.0113. Hence, we

[22] As shown in Appendix E Table 3.E1, additional analyses using student-level data show similar findings of PLA and watch list effects. We chose to report school-level results for two major reasons. The first is that the PLA policy is designed to evaluate and monitor school-level student performance from year to year. Thus, school-level estimates have direct policy implications. The second is that student-level data with larger sample sizes tend to produce deflated standard errors for statistical significance tests, whereas school-level analysis yield a set of more conservative impact estimates.

can reject the null hypothesis at the 5 percent significance level, suggesting that the positive PLA list effect on writing is stronger than the watch list effect. The results for the percent of students who met proficiency level in social studies (χ^2=2.82, p=0.0929) show that we can only reject the null hypothesis at the 10 percent significance level.[23]

3.5.3 Interpreting the Robustness of RD Inferences

3.5.3.1 Sensitivity to Bandwidth Choice

The following analyses focus only on testing the robustness of our primary results for the 2010 PLA list effects. As a first set of robustness tests, we conduct our RD models using various bandwidths. Figure 3.A3 displays the RD causal impact estimates, along with 95 percent confidence intervals (CI), of the 2010 PLA list on the five subjects for both the percent of students who met proficiency level and average of students' scale scores in 2011 for different bandwidths around the cutoff. The range of bandwidth is from the percentile rank of 8 (*n*=49) to the maximum limit 11.4 (*n*=84), with increments of 0.2. [24] In every panel, the solid line represents the PLA causal estimate whereas the upper and lower dashed line represent the upper and lower limit of 95 percent CI respectively.

Figure 3.A3 shows that the positive PLA list effect on the average of students' scale scores in writing is especially robust to variation in bandwidths, where the values of zero fall below the lower limit of 95 percent CI. The positive PLA list impact is modestly robust for the percent of students who met proficiency level in social studies. For other schools' outcome measures, it appears that the positive PLA list effect is critically dependent on a particular bandwidth choice.

---

[23] We also perform a set of Wald tests to test the joint significance of the impact estimates across all five subjects for both PLA and watch list with different bandwidths. As reported in Appendix E Table 3.E2, the results suggest that we can reject the null hypothesis at the 5 percent significance level for the PLA list effects on the average of students' scale score for the bandwidth of 9 and 10 percentile ranking and on the percent of students who met proficiency level for the bandwidth of 9 percentile ranking. However, when testing the joint significance without writing scores, none of the Wald tests is statistically significant, suggesting that the significance of joint test is largely driven by the effect in writing. In addition, none of the joint significance tests for watch list is statistically significant.

[24] Ideally, we might use a wide range of observations on either side of a boundary to explore the sensitivity of the results to bandwidth choices. For our study sample, however, in the absence of large amounts of data, restricting analyses to bandwidth very close to the threshold results in imprecise estimates. Specifically, if we use any bandwidth smaller than 8, our analysis sample size reduces to fewer than 50 cases which will lead to a great reduction in precision and generally at the cost of statistical power. On the other hand, the maximum bandwidth that we can use on one side is limited to 11.4 as the cutoff for PLA list is the bottom 11.4 percentile on the TTB ranking.

3.5.3.2 Quantifying the Robustness of RD Inferences

To inform policy debates and theoretical interpretations of the causal effects of the PLA list, it is useful to quantify the discourse about the robustness of the inferences in this study. We quantify how much bias there must be in our RD estimates to invalidate inferences in terms of replacement data,[25] focusing only on the positive PLA list effects on the average of students' scale scores in writing and the percent of students who met proficiency level in social studies. As shown in Table 3.B5, to invalidate our causal inference of the PLA list effects on the average of students' scale scores in writing, we would need to replace about 25 to 32 percent of our PLA schools with school samples for which there is zero effect of being on the list. These 17 to 22 replacement schools could represent populations not directly in our sample, such as schools from outside of the selected bandwidth. Additionally, to invalidate the inference of an effect of assignment to the PLA list on social studies achievement we would have to replace 6 to 8.6 percent of schools with schools in which there was no effect of being on the PLA list.

This analysis helps us quantify the robustness of the inference with respect to internal validity by considering the replacement schools to come from different bandwidths. The same analysis can apply to external validity by considering the replacement schools to come from a different state or time. Thus the analysis tells us how much we would have to change our sample to change our inference. In summary, based on this quantification of possible bias to invalidate an inference, we find that our RD estimates of the PLA list on the average of students' scale scores in writing are particularly robust while they are less so for the percent of students who met proficiency level in social studies.

3.5.3.3 Falsification Test

Our primary results show that being on the 2010 PLA list increases student achievement in writing, has marginal positive effects on mathematics and social studies, and no evidence of effects on reading and science scores. These findings raise the question of whether the estimated positive effects are due to a placebo effect in which the lowest-performing schools in a specific year revert back to normal

[25] As defined by Frank et al. (2013):

The proportion of bias to make inference invalid = 100% × (estimate – (s.e. × $t_{critical,df}$)) /estimate

performance levels in student achievement the following year. As a robustness test, Table 3.B6 presents the results from estimating the effects of being placed on a pseudo-2009 PLA list which is created using data prior to 2010 and similar criteria used for 2010 PLA designation. The only even marginal effect is for math as measured as the percent of students who met proficiency level (only statistically significant at the critical level of 10 percent). Overall, we find no significant effect of the pseudo-2009 PLA list on a majority of school outcome measures, suggesting that the bottom 5 percent schools in 2009, which would have been on a PLA list if the policy was being enforced one year earlier, did not substantially improve student performance in the following year. Taking together all three set of sensitivity analyses above, we find no reasons to doubt the robustness of our primary findings regarding the positive causal effects of being placed on the 2010 PLA list in Michigan.

3.5.4 Statistical Power

When evaluating the quality of RD analysis, it is important to examine the precision of its estimates in terms of statistical power. The precision of RD impact estimates can be quantified based on the calculation of minimum detectable effect, which is the smallest treatment effect that a RD design has an acceptable chance of detecting. Following Bloom (2012) and Schochet (2009), we define an minimum detectable effect as the smallest true treatment effect that has an 80 percent chance (or 80 percent power) of providing a treatment impact estimate that is statistically significant at the critical level of 5 percent based on a two-tailed test.[26]

---

[26] Minimum Detectable Effect = $2.8 \times \sqrt{\frac{\left(\hat{\sigma}^2_{Y|T}\right)\left(1-R^2_{Y.\boldsymbol{R}.\boldsymbol{X}|T}\right)}{N\pi(1-\pi)\left(1-R^2_{T.\boldsymbol{R}.\boldsymbol{X}}\right)}}$

N = sample size used in the estimation

π = proportion of this sample assigned to the treatment

$\hat{\sigma}^2_{Y|T}$ = within-group sample variance in the outcome variable

$R^2_{T.\boldsymbol{R},\boldsymbol{X}}$ = proportion of variance in treatment status explained by the forcing variable and other covariates

$(1 - R^2_{Y.\boldsymbol{R},\boldsymbol{X}|T})$ = proportion of error variance left unexplained by the forcing variable and other covariates

We calculated the minimum detectable effects for the estimated causal effects of the 2010 PLA and watch list. The minimum detectable effect sizes for our PLA and watch list analysis range from 0.25 to 0.45, which are common for school-based RD designs that are likely to have relatively large effects (about a minimum detectable effect of 0.33 standard deviations or more; Schochet 2009).[27] The results show that our RD impact estimates of the PLA list on the average of students' scale scores in writing and mathematics exceed its minimum detectable effect, indicating we have the capacity to detect a positive treatment effect for the PLA list (see Table 3.E3). A similar conclusion can be drawn when we recomputed the minimum detectable effects with a more stringent standard of 90 percent power for two-tailed tests at 5 percent significance level (see Table 3.E4). Additionally, we find that none of the RD impact estimates of the watch list for any outcome exceed its minimum detectable effect (results not reported here).

3.5.5 Concerns about the Test-Taking Eligibility of Student Sample

One crucial concern about studying effects of school sanctions is that schools might strategically manipulate the population of test-takers, particularly: (1) disproportionately retaining low-achieving students in a low-stakes grade (in our case, 10th grade); (2) reassigning low-performance students into a special education category; or (3) excluding low-scoring students in a high-stakes grade from taking the test (Figlio and Getzer 2006; Heilig and Darling-Hammond 2008). We address these issues below.

On the surface, our student level data (see Appendix E Table 3.E5) show that compared with watch list schools, PLA schools tend to have relatively lower 10th grade promotion rates (75.1 percent vs. 55.0 percent), higher retention rates (4.3 percent vs. 7.6 percent), and higher transfer rates (20.4 percent vs. 37.2 percent). But, the differences are quite similar to the two prior years (see 2nd and 3rd panel in Appendix E Table 3.E5) and are not statistically significant when we examine it in the same RD framework used for testing the discontinuity of baseline covariates (see Appendix E Table 3.E6). More

[27] As our study is a school-level analysis, the MDEs are calculated at the school level. Typically, the variance in achievement scores at the school level is only 10 to 15 percent of the variance in student level measures. Thus we can rescale the school level MDEs into student-level by dividing the MDEs by 3.2 as recommended by Dee (2012). The calculated student level MDEs are ranging from 0.07 to 0.14.

importantly, we find that the differences in prior academic achievement (8th grade test scores in reading, writing, mathematics, and science) between regular promoted students and those who retained/transferred out are considerably smaller in PLA schools than those in non-PLA schools (see Appendix E Table 3.E7), suggesting that watch list schools lost relatively more lower-scoring students than PLA schools did. In this sense, our primary findings of positive PLA list effects are likely to be underestimated.

Second, in both PLA and watch list schools, the incidence of 11th grade students being reassigned into a special education category is very low (less than 0.4 percent) and negligible (see the last column in Appendix E Table 3.E5). Third, although PLA schools are more likely to have a higher percentage of 11th grade students not taking the MME test in 2011 (26.7 percent), as compared with watch list schools (10.1 percent), it is not an unusual rate when comparing with the two prior years (see Appendix E Table 3.E8). The difference is also not statistically significant when examining in RD models with school covariates (see Appendix Table 3.E6). Critically, the achievement gaps (prior performance in reading, writing, mathematics, and social studies) between those test takers and non-test takers in PLA versus comparable non-PLA schools did not create a positive bias in favor of a PLA effect (see Appendix E Table 3.E9). For example, on average, non-test takers in PLA schools scored 8.1 points lower than test takers did in prior writing test whereas in non-PLA schools the difference is 9.2 points. Overall, it appears that the estimated positive PLA list effects are not due to the manipulation of test-taking population.

3.6 Discussion

This study contributes to the relatively limited research on school labeling by presenting new evidence on the effects of consequential and non-consequential labeling among traditional high schools in Michigan. To summarize, our RD estimations suggest that being on the 2010 PLA list has a positive effect on student performance in writing, a marginal effect in mathematics and social studies, and no effect in reading and science. The positive effects in writing are quite robust but the effects on mathematics and social studies are less so. For writing, PLA schools gained on average about 0.53-0.63

standard deviations at the school level on statewide achievement test.[28] At the bottom end of the distribution of school performance in writing, such magnitude of achievement gain would improve the ranking of PLA schools on writing scale per se (not Top-to-Bottom percentile ranking) by about 4 to 5 percentile upward.

We find no evidence that the positive PLA effects are produced by the manipulation of student population. However, as shown in prior studies, when facing increased accountability pressure, schools may strategically reallocate instructional resources and teachers' time on subjects which are easier to improve in the short term (Smith, Roderick, and Degener 2005). Goldhaber and Hannaway (2004), and Chakrabarti (2013a), for example, find that students in failing schools are more likely to have the biggest score gain in writing, which is considered one of the subjects on which students can improve quickly. We recognize that the observed immediate strong positive effects in writing in our study may be a result of test preparation coaching at the early stage of sanction (Darling-Hammond & Wise, 1985).

While we find significant positive effects demonstrated by the PLA list schools, our analyses show that there is no remarkable difference between the 2010 watch list schools and comparable schools in student performance in 2011. This finding provides empirical evidence on how low-performing schools respond to different forms of labeling in terms of accountability pressures and social stigmatization processes. Analyzing the unique context in Michigan in which there are two types of low-performing school lists, our results suggest that schools act differently in response to labeling intensities. On one hand, the consequential PLA label, which is publicly known and accompanied by accountability requirements and potential resource based sanctions, appears to be a strong triggering factor for failing schools to

[28] To rescale school-level effects sizes into student-level ones, Lipsey et al (2012) recommend to divide the effect sizes by 2 or a larger number. In a similar analysis studying school turnarounds, Dee (2012) uses denominators ranging from 2.6 to 3.2. Using 3.2 as denominator (which should yield lower bound estimates of student level effects sizes), the corresponding student-level effect sizes range from 0.17 to 0.20 standard deviations. The effect size represents a significant improvement in writing score when comparing to the national benchmarks of annual achievement gain for 11th graders, ranging from 0.14 to 0.19 standard deviations for academic standardized tests, as computed by Lipsey et al. (2012). It is important to note that the benchmarks provided by Lipsey et al. (2012) are derived from national norming studies, whereas our study sample is only from a single state (i.e., Michigan), which tends to produce larger effect sizes even when the actual intervention effects are identical as the variance in achievement measures for narrower populations is likely smaller than broader samples.

improve. On the other hand, schools may not actively respond to the non-consequential labeling that is relatively low-intensity in the stigmatization process and without immediate threat of sanctions.

The overall findings of our study are notable for the various reasons discussed above, yet we recognize that there are several limitations of this study both in terms of analytical approach as well as data set. First, the small sample size precludes us from estimating the effects of PLA/watch list in RD designs by selecting a group of school samples that are very close to the cutoff while having enough statistical power. This small sample size issue is not uncommon in the RD literature evaluating the impact of low-performing school labels, in which only a limited number of schools would have been labeled as low-performing or "F"-rated in a given district or state each year (e.g., Rouse et al. 2013; Winters and Cowen 2012). Nevertheless, in this study, we are able to optimize the bias-precision trade-off by choosing a bandwidth that allows us to detect a positive effect of PLA list with an acceptable level of statistical power (80 percent power for two-sided tests at the 0.05 significance level, see Appendix E Tables 3.E1 and 3.E2).

Furthermore, in our RD designs, the control group is selected from the cluster of schools that ranked above the cutoff and the treatment schools ranked below the cutoff, the assumption here was that the two populations are similar and comparable at a certain degree. However, this assumption could be violated considering the difficulty of improving school performances consistently at the bottom. Particularly, when the bottom schools are characterized as enrolling consistently disadvantaged students in terms of educational resources, it might be difficult to find differences between the treatment and control schools as they both serve similar populations with similar resources. To make the two populations comparable, we needed to narrow the window which might cause attenuated regression coefficients due to the limited range of variance. Therefore, a causal inference from the regression discontinuity design inevitably relies on the assumption that we made a comparable control group using the range of a narrow window, and this assumption may be questionable. Fundamentally, this study cannot be fully free from the limitation of a regression discontinuity design. By quantifying concerns about this assumption, we show that at least 25 percent of the estimated PLA list effect on writing would

have to be due to uncontrolled bias to invalidate our inference. This level of robustness is about at the median for educational evaluation studies with observational data (Frank et al. 2013).

The limited variables in the data set also leave the following concerns. First, we lack implementation information. We do not have the proper information to investigate what actual changes are being implemented in the schools in the first and second year after the PLA list announcement. Second, the school-level variables are quite limited in content and there may be other factors such as change in school leadership or intensive professional development that could be affecting the performance in the schools. To achieve a more definitive conclusion, such variables should be considered.

Nonetheless, even with these few variables and a short-time comparison of a narrow band of similar schools, PLA label appears to have some positive effects. It seems prudent to investigate these types of labels as a part of the school sanctioning process as they are commonly applied and relatively cost efficient. Moreover, understanding how schools respond to specific rules of accountability is crucial for designing effective school reform programs. To date, all 50 states have started to identify and publish a lowest performing 5 percent school list annually, based on PLA criteria similar to those used in Michigan. Thus, our findings from studying the PLA and watch list in Michigan may have broader applications to the similar systems or school labeling policies that are becoming prevalent across the nation. Although the labeling practice varies to some extent from state to state, our study highlights that low-performing schools are likely to respond differently to varying forms of labeling, particularly with or without implied consequences.

APPENDICES

APPENDIX A

FIGURES FOR CHAPTER 3

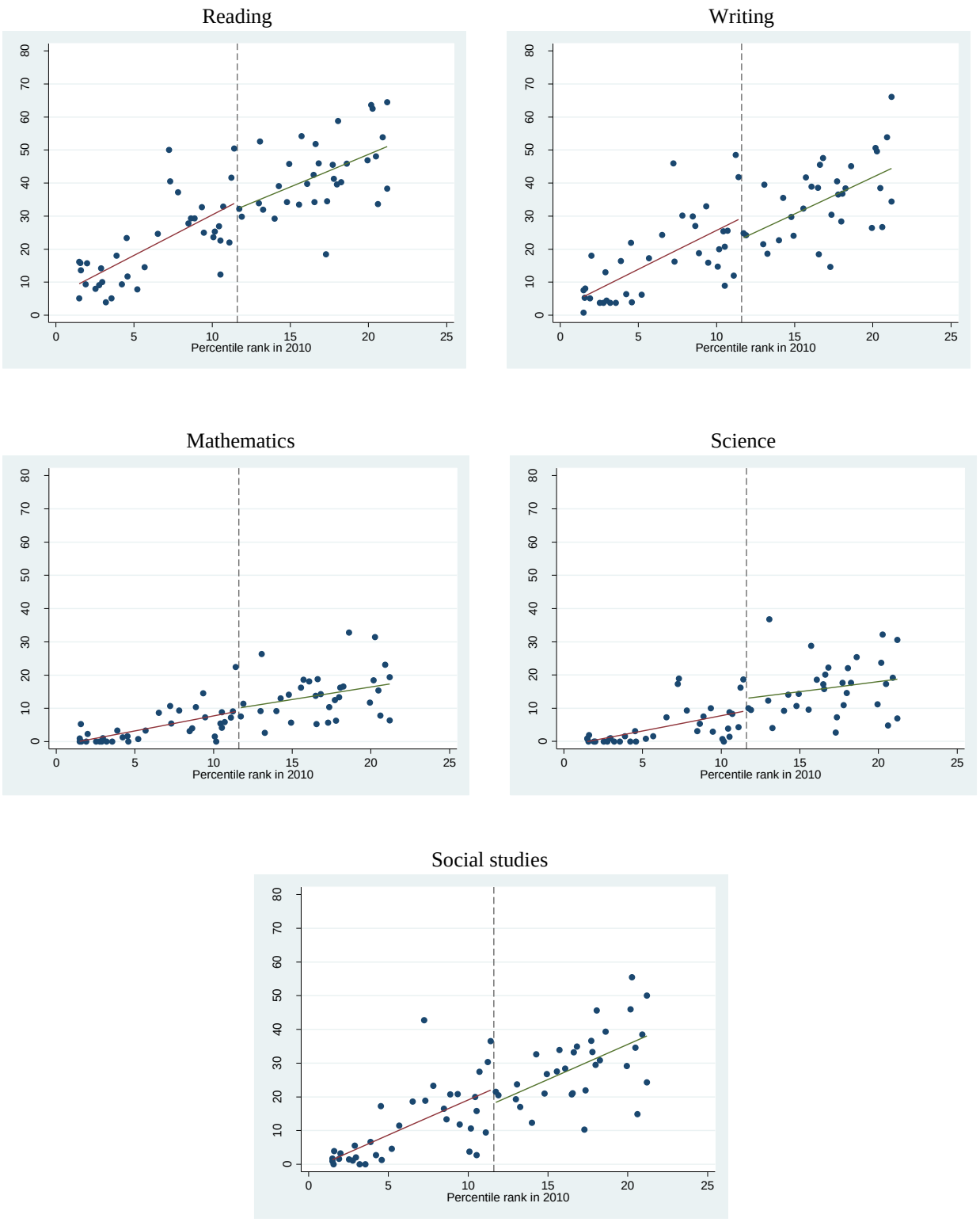


Figure 3.A1 The Relationship between Percentile Rank in 2010 and Percent of Students Met Proficiency Level in Five MME Subjects in 2011.

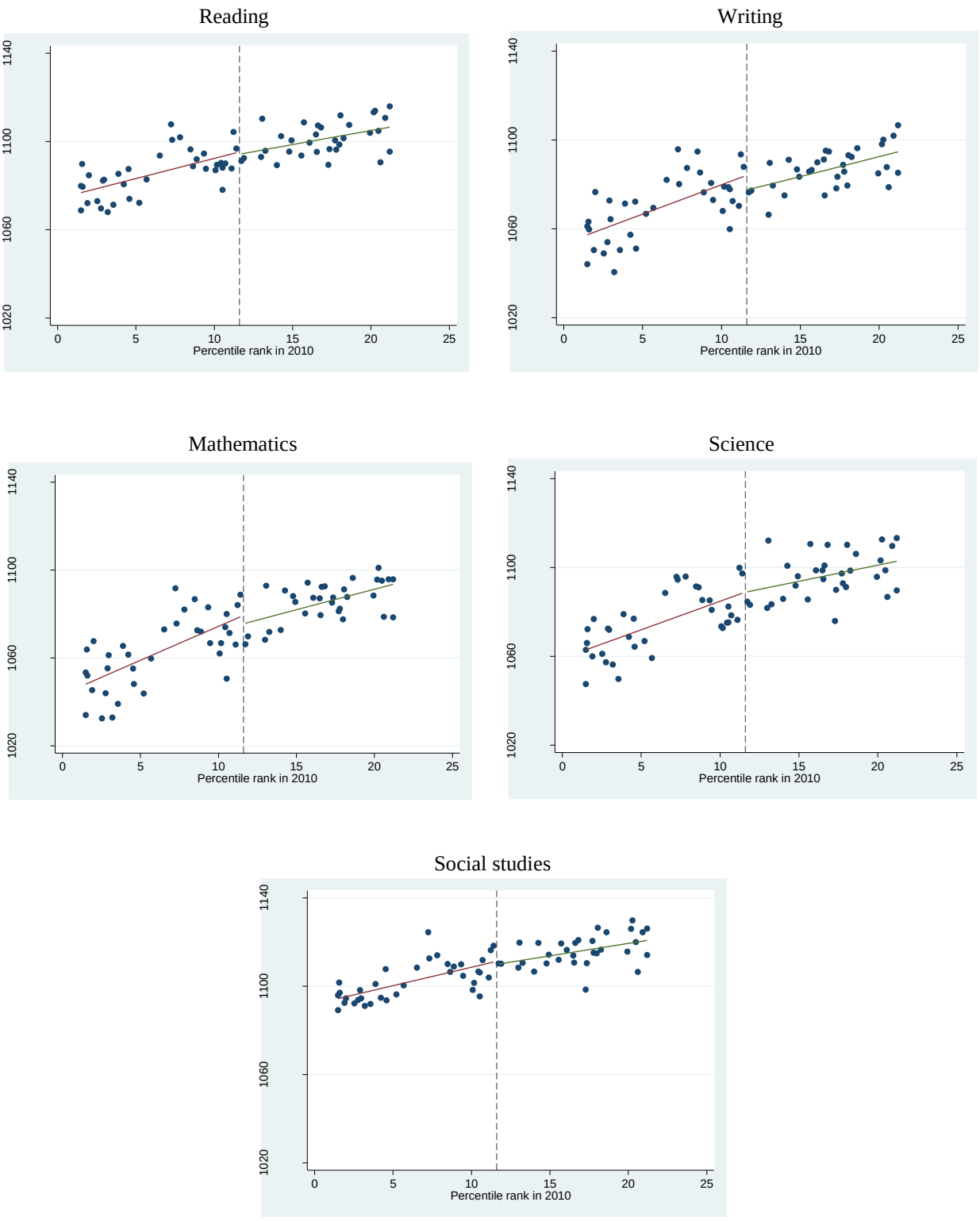
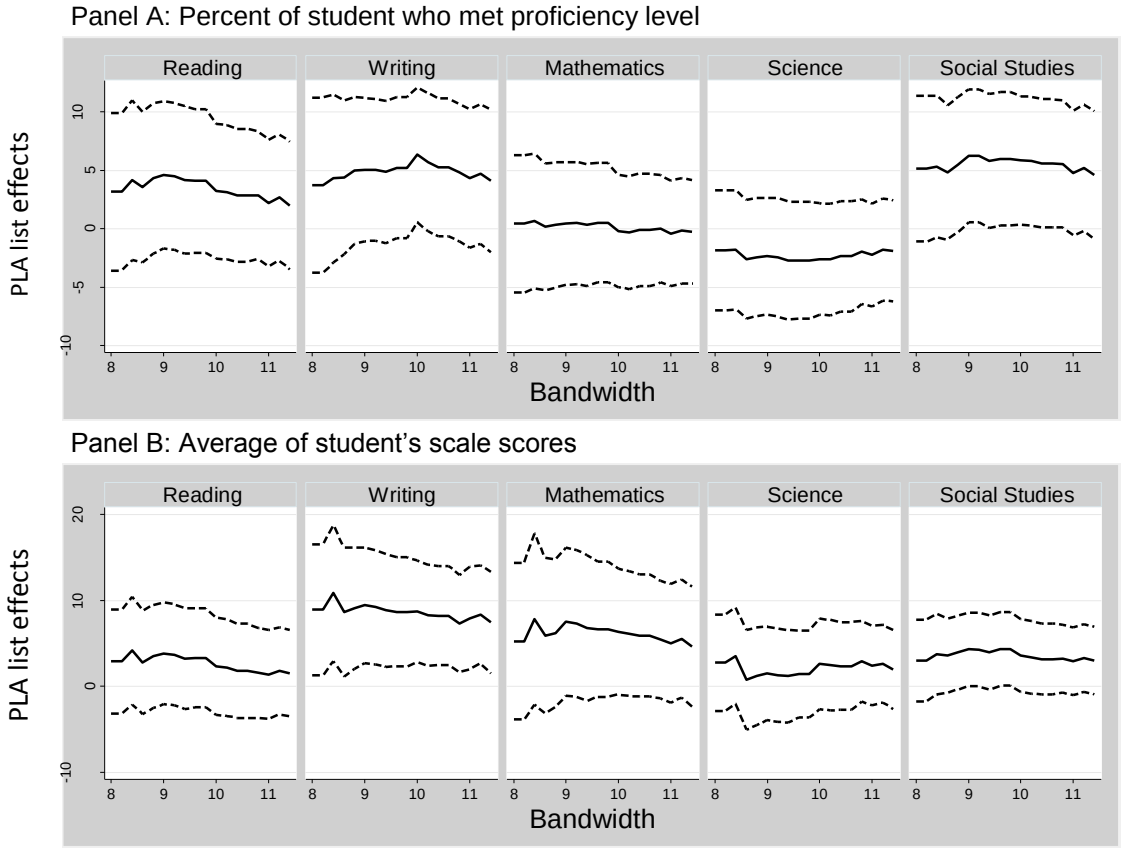


Figure 3.A2 The Relationship between Percentile Rank in 2010 and Average of Students' Scale Score in Five MME Subjects in 2011.



Note. The upper and lower dashed lines represent the upper and lower limit of 95% CI respectively. The greater fluctuations on the left sides for the smaller bandwidths are due to the smaller samples in these bandwidths which create more sampling variability.

Figure 3.A3 RD Impact Estimates of the 2010 PLA List (and 95% CI) by Selection of Bandwidth.

APPENDIX B

TABLES FOR CHAPTER 3

Table 3.B1 Low-performing School Labels by State-Wide Ranking in 2010

Label	(a) TTB percentile rank (range)	(b) PLA eligibility school	(c) Excluded cases		(d) Final school sample
			School closed	SIG school	
No designation	28.1 – 100	265	1	2	262
Watch list	11.4 – 28.1	60	0	4	56
Persistently Lowest Achieving (PLA)	0 – 11.4	56	4	9	43
Total		381	5	15	361

Note. TTB = Top-to-Bottom. Our final analysis sample excluded those schools which have been closed in the 2010-2011 school year and schools that received School Improvement Grant (SIG) funding in 2010.

Table 3.B2 Descriptive Statistics

	PLA list (*n*=43)		Watch list (*n*=56)		No designation (*n*=262)	
	TTB ≤ 11.4		11.4 < TTB ≤ 28.1		TTB > 28.1	
Variable	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Dependent variables						
% of students who met proficiency level (2011)						
Reading	19.51	12.13	44.41	10.79	56.69	10.39
Writing	15.06	12.27	35.61	10.43	50.20	12.06
Mathematics	3.83	4.76	15.42	7.21	28.92	11.99
Science	3.72	4.83	13.44	5.60	23.31	9.09
Social studies	10.03	10.65	31.43	9.96	45.14	11.15
Average of students' scale scores (2011)						
Reading	1083.91	10.30	1102.25	7.38	1110.50	6.90
Writing	1067.17	15.41	1087.56	8.41	1097.81	8.66
Mathematics	1059.92	16.67	1086.43	8.71	1098.80	8.13
Science	1072.91	13.91	1097.80	9.49	1107.95	7.82
Social studies	1101.04	8.64	1117.45	6.52	1125.48	6.78
Covariates						
% of students who met proficiency level (2009)						
Reading	19.83	10.81	40.62	8.05	52.99	10.27
Writing	14.80	10.37	33.05	8.21	46.72	11.93
Mathematics	4.08	5.02	14.72	6.11	27.83	11.47
Science	3.72	4.83	13.44	5.60	23.31	9.09
Social Studies	11.59	10.00	32.54	8.71	45.33	11.96
Average of students' scale scores (2009)						
Reading	1083.42	8.94	1098.45	11.63	1108.68	6.92
Writing	1067.57	10.89	1083.43	7.24	1093.70	9.97
Mathematics	1067.32	11.36	1087.15	6.63	1098.69	7.78
Science	1070.12	12.81	1092.79	8.07	1103.55	9.70
Social studies	1103.42	8.28	1118.38	15.26	1129.12	8.01
% of free/reduced lunch students (2010)	70.59	12.89	50.14	13.98	34.58	14.98
% of minority students (2010)	80.66	29.43	27.02	27.29	10.54	12.00
School size (2010)	852.23	504.62	752.29	487.02	794.53	549.31
Pupil teacher ratio (2010)	19.27	3.74	19.17	2.96)	19.26	2.73

Source. Michigan Merit Examination (MME), Michigan Department of Education (MDE); Common Core Data (CCD), National Center for Education Statistics, U.S. Department of Education.
Note. *n* = sample size; TTB = Top-to-Bottom percentile ranking.

Table 3.B3 RD Impact Estimates of the 2010 PLA List across Three Bandwidths

	Reading		Writing		Mathematics		Science		Social studies	
Panel A: % of students who met proficiency level (2011)										
Bandwidth = 9 (n=57)	4.596	(3.135)	5.060	(3.047)	0.463	(2.615)	-2.332	(2.481)	6.233*	(2.833)
Bandwidth = 10 (n=68)	3.215	(2.878)	6.335*	(2.884)	-0.202	(2.404)	-2.590	(2.380)	5.846*	(2.752)
Bandwidth = 11 (n=79)	2.203	(2.718)	4.311	(2.970)	-0.410	(2.260)	-2.253	(2.207)	4.775†	(2.680)
Panel B: Average of students' scale score (2011)										
Bandwidth = 9 (n=57)	3.837	(2.938)	9.409**	(3.334)	7.507†	(4.280)	1.509	(2.708)	4.298*	(2.135)
Bandwidth = 10 (n=68)	2.338	(2.825)	8.706**	(2.950)	6.355†	(3.644)	2.611	(2.637)	3.602†	(2.119)
Bandwidth = 11 (n=79)	1.380	(2.583)	7.915**	(2.982)	5.024	(3.463)	2.399	(2.318)	2.935	(1.970)

Note. *n* = sample size. Taken from a separate regression model on schools, each cell in the table shows the estimated coefficient on a dummy variable indicating the effect of being on the 2010 PLA list. All estimation models include the 2010 percentile ranking (as forcing variable), Tier 2 pool status, and 2009 pretest measure for a given subject, as well as other school characteristic covariates, including percent of free/reduced lunch students, percent of minority students, school size, and pupil teacher ratio, which are collected in the 2009-2010 academic year. Robust standard errors reported in parentheses. Statistical significance is determined using two-tailed tests.
*** p<.001; ** p<.01; * p<.05; †p < .10.

Table 3.B4 RD Impact Estimates of the 2010 Watch List across Three Bandwidths

	Reading		Writing		Mathematics		Science		Social studies	
Panel A: % of students who met proficiency level (2011)										
Bandwidth = 9 (n=64)	1.510	(3.327)	-1.415	(4.227)	0.511	(2.710)	1.721	(2.863)	-0.320	(3.378)
Bandwidth = 10 (n=73)	2.844	(2.938)	-1.731	(3.711)	-0.662	(2.642)	0.014	(2.745)	-0.968	(3.313)
Bandwidth = 11 (n=83)	3.789	(3.000)	-0.367	(3.479)	0.914	(2.569)	0.854	(2.553)	-0.063	(3.046)
Panel B: Average of students' scale score (2011)										
Bandwidth = 9 (n=64)	-0.639	(2.038)	-0.606	(2.516)	-1.377	(2.452)	0.937	(2.530)	0.825	(2.128)
Bandwidth = 10 (n=73)	0.052	(1.853)	-0.305	(2.387)	-1.396	(2.254)	0.832	(2.614)	1.139	(2.004)
Bandwidth = 11 (n=83)	1.048	(1.962)	-0.091	(2.256)	-1.085	(2.285)	1.622	(2.561)	1.675	(2.101)

Note. *n* = sample size. Taken from a separate regression model on schools, each cell in the table shows the estimated coefficient on a dummy variable indicating the effect of being on the 2010 Watch list. All estimation models include the 2010 percentile ranking (as forcing variable), Tier 2 pool status, and 2009 pretest measure for a given subject, as well as other school characteristic covariates, including percent of free/reduced lunch students, percent of minority students, school size, and pupil teacher ratio, which are collected in the 2009-2010 academic year. Robust standard errors reported in parentheses. Statistical significance is determined using two-tailed tests.
*** p<.001; ** p<.01; * p<.05; †p < .10.

Table 3.B5 Quantifying the Robustness of Inferences from RD Impact Estimates of the 2010 PLA List

				Average of students' scale scores in writing			Percent of students who met proficiency level in Social Studies		
	n	*df*	$t_{critical}$	β	*s.e.*	*% of bias*	β	*s.e.*	*% of bias*
Bandwidth = 9	57	50	2.008	9.409**	(3.334)	28.9	6.233*	(2.833)	8.6
Bandwidth = 10	68	61	1.999	8.706**	(2.950)	32.3	5.846*	(2.752)	5.9
Bandwidth = 11	79	72	1.993	7.915**	(2.982)	24.9	4.775†	(2.680)	-

Note. *n* = sample size; df = degree of freedom; $t_{critical}$ = critical value of the *t* distribution (two-tailed test); β = RD estimate; *s.e.* = standard errors.
*** p<.001; ** p<.01; * p<.05; †p < .10.

Table 3.B6 RD Impact Estimates of the 2009 Pseudo-PLA List across Three Bandwidths

	Reading		Writing		Mathematics		Science		Social studies	
Panel A: % of students who met proficiency level (2010)										
Bandwidth = 8 (n=77)	-1.246	(2.818)	-2.307	(1.994)	2.130†	(1.232)	0.054	(1.249)	-2.558	(2.723)
Bandwidth = 9 (n=85)	-1.906	(2.746)	-2.630	(1.906)	2.179†	(1.128)	0.567	(1.217)	-2.136	(2.562)
Bandwidth = 10 (n=95)	-2.410	(2.560)	-2.270	(1.778)	2.229†	(1.122)	0.893	(1.107)	-1.300	(2.394)
Panel B: Average of students' scale score (2010)										
Bandwidth = 8 (n=77)	-1.951	(1.998)	-2.392	(2.874)	1.913	(3.044)	-1.355	(3.103)	-0.242	(1.690)
Bandwidth = 9 (n=85)	-1.586	(1.953)	-1.568	(2.998)	2.693	(2.979)	-0.940	(2.984)	0.109	(1.616)
Bandwidth = 10 (n=95)	-1.519	(1.864)	-1.727	(2.794)	1.233	(2.956)	-1.597	(2.793)	0.307	(1.500)

Note. *n* = sample size. Taken from a separate regression model on schools, each cell in the table shows the estimated coefficient on a dummy variable indicating the effect of being on the 2009 pseudo-PLA list. All estimation models include the 2009 percentile ranking (as forcing variable), Tier 2 pool status, and 2008 pretest measure for a given subject, as well as other school characteristic covariates, including percent of free/reduced lunch students, percent of minority students, school size, and pupil teacher ratio, which are collected in the 2008-2009 academic year. Robust standard errors reported in parentheses. Statistical significance is determined using two-tailed tests.
*** p<.001; ** p<.01; * p<.05; †p < .10.

APPENDIX C

ADDITIONS FOR CHAPTER 3

C. RD Validity Tests

1. Testing of Sharp RD Assumptions

(a) *No-manipulation assumption*. We assess the no-manipulation assumption which suggests that no individual school managed to manipulate the value of the assignment variable in order to be on one side of the threshold rather than the other. If this happens, one might expect to observe a discontinuity in the density of the forcing variable at the cut score. Typically, researchers will test the null hypothesis of the continuity of the density of the running variable which determines the treatment assignment at the cut score, against the alternative hypothesis of a discontinuity in the density function at that cutoff point (McCrary, 2008). In our case, all ranked schools (including elementary, middle, and high schools) are evenly distributed on a 100-percentile scale. By design, there is no discontinuity in the density of the percentile ranking at any cut scores, as represented by the gray bars in Figure 3.D1. Although the density of high schools at different levels of the rating score is varied (as represented by solid line bars in Figure 3.D1), there is no systematic pattern showing that the differences in the density of the forcing variable for the high schools appear only around the cutoffs of PLA (11.4%) and Watch list (28.1%). Nonetheless, one might question that some schools acted strategically in an effort to manipulate their rank. We cannot test this directly but a test of unconfoundedness assumption (presented in the following section), which examines whether the school covariates are continuous around the cutoffs, can provide useful evidence for detecting such possible manipulation. In addition, we believe that it is unlikely for any schools to manipulate their ranking as the PLA policy was announced in December 2009 and their student test scores from 2006 to 2009 (prior to the PLA announcement) were used for the calculation of 2010 percentile ranking.

(b) *Unconfoundedness assumption*. We evaluate the unconfoundedness assumptions by testing the null hypothesis of a zero average effect on other school characteristics as pseudo outcomes known not to be impacted by the treatment (Imbens & Lemieux, 2008). We estimate regressions taking the form which is

similar to the RD basic specification in equation (1) but with a different school covariate serving as the outcome:

$$Y_i^{2009} = \beta_0 + \beta_1 List_i^{2010} + \beta_2 PctRank_i^{2010} + \mu_i \quad (2)$$

$$X_i^{2010} = \beta_0 + \beta_1 List_i^{2010} + \beta_2 PctRank_i^{2010} + \mu_i \quad (3)$$

where Y_i^{2009} is the pretest scores for each school outcome in 2009, X_i^{2010} is some pretreatment covariate measured in 2010, and the other variables are as previously defined. The results are presented in Table 3.E10. We find no significant jumps in the value of school measures at both the cutoff points of PLA and Watch list, which may invalidate the regression discontinuity design.

As suggested by Lee and Lemieux (2010), it is useful to perform a seemingly unrelated regression (SUR) analysis if there are multiple covariates available. Our SUR model consists of the specification in equation (2) and (3). Then, we perform a chi-square test for testing the hypothesis that all discontinuity terms across the 14 covariates are jointly equal to zero. For PLA list, results yield a chi-square value of 9.96, with 14 degrees of freedom, and a p-value of 0.7655; for Watch list, results yield a chi-square value of 10.01, with 14 degrees of freedom, and a p-value of 0.7612. Therefore we cannot reject the null hypothesis of no discontinuities of covariates around the fixed thresholds.

(c) *No jumps at non-discontinuity points assumption*. A third set of specification tests for a zero effect in settings where it is expected that there would be no effect (Imbens & Lemiuex, 2008). If there is an extraneous discontinuity in the dependent variable away from the fixed threshold, the assumption of smoothness in the absence of treatment will be called into question. In practice, we test if the average outcome is discontinuous at other values of the percentile ranking, particularly at the median of the subsamples on either side of the threshold, as suggested by Imbens & Lemieux (2008). We choose bottom 5.7% (median of PLA list) and 19.7% (median of watch list) as two placebo cutoff points and test the continuity of school outcomes at each using the specification in equation (1). The results are reported in Table 3.E11. Overall, we find no evidence to reject the null hypothesis of a zero jump at various values of percentile ranking away from the cutoffs.

APPENDIX D

SUPPLEMENTAL FIGURE FOR CHAPTER 3

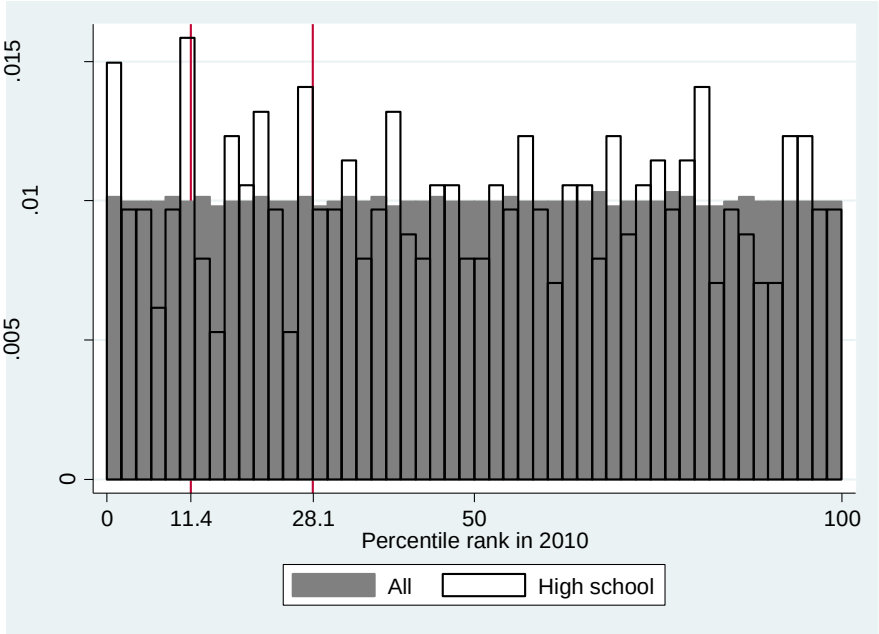


Figure 3.D1 Density of Forcing Variable (Percentile Rank in 2010).

APPENDIX E

SUPPLEMENTAL TABLES FOR CHAPTER 3

Table 3.E1 RD Impact Estimates of the 2010 PLA and Watch List across Three Bandwidths, Student Level

Student scale score in MME (2011)	Reading		Writing		Mathematics		Science		Social studies	
Panel A: PLA list effect										
Bandwidth = 9 (n_1=57; n_2=6,237)	3.137†	(1.589)	8.498***	(2.351)	7.188*	(2.782)	4.196*	(1.739)	2.839†	(1.642)
Bandwidth = 10 (n_1=68; n_2=7,113)	2.655	(1.624)	7.909***	(2.283)	6.145*	(3.070)	3.689*	(1.740)	2.531	(1.568)
Bandwidth = 11 (n_1=79; n_2=8,174)	2.425	(1.518)	6.493**	(2.414)	5.242†	(3.012)	3.875*	(1.818)	1.919	(1.539)
Panel B: Watch list effect										
Bandwidth = 9 (n_1=57; n_2=6,897)	-0.335	(1.960)	1.785	(2.579)	0.878	(1.744)	-0.626	(1.709)	-0.689	(1.604)
Bandwidth = 10 (n_1=68; n_2=8,057)	0.264	(1.923)	1.753	(2.373)	0.045	(1.785)	-0.836	(1.847)	0.069	(1.703)
Bandwidth = 11 (n_1=79; n_2=9,280)	1.751	(1.739)	1.558	(2.246)	0.224	(1.717)	0.653	(1.815)	-0.325	(1.665)

Note. n_1 = number of schools; ; n_2 = number of students. Taken from a separate regression model on students, each cell in the table shows the estimated coefficient on a dummy variable indicating the effect of being on the 2010 PLA list. All estimation models include the 2010 percentile ranking (as forcing variable), Tier 2 pool status, and other student characteristics (i.e., gender, race/ethnicity, age, free/reduced lunch status, English learner status, migrant status, and test scores in 8th grade reading, writing, math, science, and 9th grade social studies) as well as school covariates (i.e., percent of free/reduced lunch students, percent of minority students, school size, and pupil teacher ratio), which are collected in the 2009-2010 academic year. Robust standard errors are clustered by schools and reported in parentheses. Statistical significance is determined using two-tailed tests.
*** p<.001; ** p<.01; * p<.05; †p < .10.

Table 3.E2 Joint Hypotheses Tests for Impact Estimates of the 2010 PLA and Watch List across All Subjects

	PLA list				Watch list	
	All five subjects		Four Subjects Without Writing		All five subjects	
	χ^2	*p*-value	χ^2	*p*-value	χ^2	*p*-value
% of students who met proficiency level (2011)						
Bandwidth = 9	9.60	0.0874	7.74	0.1016	2.28	0.8096
Bandwidth = 10	11.42	0.0437	8.22	0.0837	4.47	0.4835
Bandwidth = 11	7.38	0.1937	6.39	0.1718	4.22	0.5177
Average of students' scale score (2011)						
Bandwidth = 9	11.84	0.0368	6.47	0.1666	3.44	0.6323
Bandwidth = 10	14.18	0.0145	6.30	0.1778	2.28	0.8096
Bandwidth = 11	11.03	0.0507	5.35	0.2534	1.87	0.8674

Note. χ^2 = chi-square.

Table 3.E3 Minimum Detectable Effects (MDE) for the Estimated Causal Effects of the 2010 PLA List, for Two-Tailed Tests at 80% Power and A 5% Significance Level

	Reading			Writing			Mathematics			Science			Social studies		
	β	*MDE*	*diff.*	*β*	*MDE*	*diff.*	*β*	*MDE*	*diff.*	*β*	*MDE*	*diff.*	*β*	*MDE*	*diff.*
	(1)	(2)	(1)-(2)	(1)	(2)	(1)-(2)	(1)	(2)	(1)-(2)	(1)	(2)	(1)-(2)	(1)	(2)	(1)-(2)
% of students who met proficiency level (2011)															
h = +/-9 (n=57)	4.596	5.721	-1.125	5.060	8.512	-3.452	0.463	8.718	-8.255	-2.332	5.456	-7.788	6.233	7.049	-0.816
h = +/-10 (n=68)	3.215	4.912	-1.697	6.335	7.019	-0.684	-0.202	7.755	-7.957	-2.590	5.254	-7.844	5.846	5.908	-0.062
h = +/-11 (n=79)	2.203	4.445	-2.242	4.311	6.509	-2.198	-0.410	6.991	-7.401	-2.253	4.816	-7.069	4.775	5.369	-0.594
Average of students' scale score (2011)															
h = +/-9 (n=57)	3.837	5.036	-1.199	9.409	7.216	2.193	7.507	6.381	1.126	1.509	4.562	-3.053	4.298	4.424	-0.126
h = +/-10 (n=68)	2.338	4.444	-2.106	8.706	5.918	2.788	6.355	5.373	0.982	2.611	4.049	-1.438	3.602	3.827	-0.225
h = +/-11 (n=79)	1.380	3.992	-2.612	7.915	5.501	2.414	5.024	4.805	0.219	2.399	3.587	-1.188	2.935	3.531	-0.596

Note. *h* = bandwidth; *n* = sample size; *β* = estimated PLA effects; *MDE* = minimum detectable effect; *diff.* = difference between *β* and *MDE*.

Table 3.E4 Minimum Detectable Effects (MDE) for the Estimated Causal Effects of the 2010 PLA List, for Two-Tailed Tests at 90% Power and A 5% Significance Level

	Reading			Writing			Mathematics			Science			Social studies		
	β	*MDE*	*diff.*	*β*	*MDE*	*diff.*	*β*	*MDE*	*diff.*	*β*	*MDE*	*diff.*	*β*	*MDE*	*diff.*
	(1)	(2)	(1)-(2)	(1)	(2)	(1)-(2)	(1)	(2)	(1)-(2)	(1)	(2)	(1)-(2)	(1)	(2)	(1)-(2)
% of students who met proficiency level (2011)															
h = +/-9 (n=57)	4.596	6.620	-2.023	5.060	9.849	-4.789	0.463	9.998	-9.625	-2.332	6.313	-8.645	6.233	8.156	-1.924
h = +/-10 (n=68)	3.215	5.684	-2.469	6.335	8.122	-1.788	-0.202	8.973	-9.175	-2.590	6.080	-8.669	5.846	6.836	-0.991
h = +/-11 (n=79)	2.203	5.143	-2.940	4.311	7.532	-3.221	-0.410	8.090	-8.500	-2.253	5.573	-7.826	4.775	6.212	-1.437
Average of students' scale score (2011)															
h = +/-9 (n=57)	3.837	5.827	-1.991	9.409	8.350	1.059	7.507	7.383	0.124	1.509	5.278	-3.770	4.298	5.119	-0.820
h = +/-10 (n=68)	2.338	5.142	-2.804	8.706	6.848	1.859	6.355	6.218	0.138	2.611	4.686	-2.074	3.602	4.429	-0.826
h = +/-11 (n=79)	1.380	4.619	-3.240	7.915	6.366	1.550	5.024	5.560	-0.536	2.399	4.151	-1.752	2.935	4.087	-1.152

Note. *h* = bandwidth; *n* = sample size; *β* = estimated PLA effects; *MDE* = minimum detectable effect; *diff.* = difference between *β* and *MDE*.

Table 3.E5 The Whereabouts of Regular 10th Graders by School PLA Status, 2008-2009 to 2010-2011

	10th graders in 2010	Promoted as regular 11th graders in 2011		Retained as 10th graders in 2011		Transferred out after 2010		Reassigned as special ed. 11th graders in 2011	
Cohort 2010	N	N	%	N	%	N	%	N	%
PLA	8,046	4,422	55.0	610	7.6	2,989	37.2	25	0.3
Non-PLA	8,936	6,714	75.1	383	4.3	1,824	20.4	15	0.2
	10th graders in 2009	Promoted as regular 11th graders in 2010		Retained as 10th graders in 2010		Transferred out after 2009		Reassigned as special ed. 11th graders in 2010	
Cohort 2009	N	N	%	N	%	N	%	N	%
PLA	8,739	4,610	52.8	937	10.7	3,177	36.4	15	0.2
Non-PLA	9,484	7,095	74.8	440	4.6	1,922	20.3	27	0.3
	10th graders in 2008	Promoted as regular 11th graders in 2009		Retained as 10th graders in 2009		Transferred out after 2008		Reassigned as special ed. 11th graders in 2009	
Cohort 2008	N	N	%	N	%	N	%	N	%
PLA	9,785	5,220	53.4	1,005	10.3	3,528	36.1	32	0.2
Non-PLA	9,527	7,263	76.2	309	3.2	1,934	20.3	21	0.2

Note. PLA = persistently lowest-achieving schools; N = number of students.

Table 3.E6 RD Impact Estimates of the PLA list on Changes in Student Populations

Dependent variable	PLA list effect (*n*=68)	
10^{th} grade promotion rate (2011)	-0.029	(0.053)
10^{th} grade retention rate (2010 to 2011)	0.044	(0.038)
Transfer rate (after 2010)	-0.017	(0.031)
% of 11^{th} graders not taking MME (2011)	0.017	(0.053)

Note. *n* = sample size; MME = Michigan Merit Examination. Taken from a separate regression model on schools, each cell in the table shows the estimated coefficient on a dummy variable indicating the effect of being on the 2010 PLA list. All estimation models include the 2010 percentile ranking (as forcing variable), Tier 2 pool status, and 2009 pretest measure for a given subject, as well as other school characteristic covariates, including percent of free/reduced lunch students, percent of minority students, school size, and pupil teacher ratio, which are collected in the 2009-2010 academic year. Robust standard errors clustered by schools and reported in parentheses. Statistical significance is determined using two-tailed tests.
*** p<.001; ** p<.01; * p<.05; †p < .10.

Table 3.E7 The Differences in Prior Achievement Scores by Student Status for the 10th Graders of Cohort 2010

	Promoted as regular 11th graders in 2011	Retained as 10th graders in 2011		Transferred out after 2010	
	Mean (1)	Mean (2)	(2)-(1)	Mean (3)	(3)-(1)
Panel A: Reading					
PLA schools	808.7	796.6	-12.1	802.8	-5.9
Non-PLA schools	818.5	805.7	-12.8	809.3	-9.2
Panel A: Writing					
PLA schools	804.7	795.4	-9.3	799.6	-5.1
Non-PLA schools	810.9	802.2	-8.7	803.8	-7.1
Panel A: Mathematics					
PLA schools	803.6	794.5	-9.1	797.6	-6.0
Non-PLA schools	814.8	801.6	-13.2	804.5	-10.3
Panel A: Science					
PLA schools	809.9	797.8	-12.1	802.5	-7.4
Non-PLA schools	822.6	807.4	-15.2	811.0	-11.6

Note. PLA = persistently lowest-achieving. The prior achievement scores are based on the statewide Michigan Educational Assessment Program (MEAP) test when the students were in 8th grade.

Table 3.E8 The Distribution of MME Test-Takers and Non-Test Takers in 11th Grade, 2008-2009 to 2010-2011

	11th graders	MME test takers		Non-test takers	
	N	N	%	N	%
Panel A: Year 2011					
PLA schools	6,653	4,877	73.3	1,776	26.7
Non-PLA schools	7,833	7,043	89.9	790	10.1
Panel B: Year 2010					
PLA schools	6,942	4,736	68.2	2,206	31.8
Non-PLA schools	8,289	7,459	90.0	830	10.0
Panel C: Year 2009					
PLA schools	7,227	5,181	71.83	2,036	28.2
Non-PLA schools	8,552	7,595	88.8	957	11.2

Note. PLA = persistently lowest-achieving; N = number of students.

Table 3.E9 The Differences in Prior Achievement Scores between MME Test-Takers and Non-Test Takers in 2011

	Test-takers	Non-test takers	Difference
	Mean (1)	Mean (2)	(2)-(1)
Panel A: Reading			
PLA schools	808.5	797.3	-11.2
Non-PLA schools	818.1	807.5	-10.6
Panel A: Writing			
PLA schools	804.4	796.3	-8.1
Non-PLA schools	810.8	801.6	-9.2
Panel A: Mathematics			
PLA schools	803.2	793.2	-10.0
Non-PLA schools	814.4	800.9	-13.5
Panel A: Science			
PLA schools	809.3	797.6	-11.7
Non-PLA schools	822.2	807.5	-14.7

Note. PLA = persistently lowest-achieving. The prior achievement scores are based on the statewide Michigan Educational Assessment Program (MEAP) test when the students were in 8th grade.

Table 3.E10 Testing the Unconfoundedness Assumption

	PLA list		Watch list	
Dependent variable (School covariate)	(*c*=11.4%, *n*=68)		(*c*=28.1%, *n*=73)	
Percent of students met proficiency level (2009)				
Reading	3.839	(4.023)	-3.067	(3.775)
Writing	5.403	(3.012)	-3.285	(4.211)
Mathematics	0.445	(2.296)	-5.502	(3.678)
Science	0.027	(2.304)	-4.251	(3.536)
Social studies	0.562	(3.662)	-3.437	(4.932)
Average of students' scale score (2009)				
Reading	3.173	(3.405)	-1.412	(2.658)
Writing	4.586	(2.808)	-1.044	(4.348)
Mathematics	1.591	(3.482)	-3.108	(3.170)
Science	2.773	(4.315)	-8.338	(5.337)
Social studies	3.510	(3.863)	-3.354	(3.385)
% of free/reduced lunch students (2010)	-0.024	(0.061)	0.108	(0.059)
% of minority students (2010)	-0.029	(0.146)	0.069	(0.078)
School size (2010)	-0.307	(0.277)	0.019	(0.326)
Pupil teacher ratio (2010)	0.989	(1.317)	0.368	(1.317)

Note. *c* = cutoff; *n* = sample size. Taken from a separate regression model on schools, each cell in the table shows the estimated coefficient on a dummy variable indicating the effect of being on the 2010 PLA list. All estimation models include the 2010 percentile ranking (as forcing variable) and Tier 2 pool status. Robust standard errors reported in parentheses. Statistical significance is determined using two-tailed tests.
*** p<.001; ** p<.01; * p<.05; †p < .10.

Table 3.E11 Estimated Effects at the Median of the Two Subsamples on Either Side of the Cutoff

	PLA list subsamples		Watch list subsamples	
	(*c*=5.7%, *n*=43)		(*c*=19.7%, *n*=56)	
Percent of students met proficiency level in 2011				
Reading	-4.925	(5.552)	-3.967	(3.179)
Writing	3.119	(6.611)	-1.894	(4.054)
Mathematics	-1.190	(2.509)	-1.023	(3.299)
Science	0.047	(1.687)	-0.465	(2.627)
Social studies	0.883	(5.978)	-1.662	(3.333)
Average of students' scale score in 2011				
Reading	-3.739	(4.245)	-1.125	(2.430)
Writing	-3.207	(8.360)	0.892	(3.107)
Mathematics	-8.142	(7.605)	1.102	(3.122)
Science	-5.921	(5.981)	1.084	(2.669)
Social studies	0.003	(3.984)	-0.456	(2.375)

Note. *c* = cutoff; *n* = sample size. Taken from a separate regression model on schools, each cell in the table shows the estimated coefficient on a dummy variable indicating the effect of being on the 2010 PLA list. All estimation models include the 2010 percentile ranking (as forcing variable), Tier 2 pool status, and 2009 pretest measure for a given subject, as well as other school characteristic covariates, including percent of free/reduced lunch students, percent of minority students, school size, and pupil teacher ratio, which are collected in the 2009-2010 academic year. Robust standard errors reported in parentheses. Statistical significance is determined using two-tailed tests.
*** p<.001; ** p<.01; * p<.05; †p < .10.

REFERENCES

REFERENCES

Allington, R., & McGill-Franzen, A. (1992). Unintended effects of educational reform in New York State. *Educational Policy, 6*, 396-413.

Bloom, H. S. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness, 5*(1), 43-82.

Chakrabarti, R. (2013a). Vouchers, public school response and the role of incentives: Evidence from Florida. *Economic Inquiry, 51*(1), 500-526.

Chakrabarti, R. (2013b). Accountability with voucher threats, responses and the test-taking population: Regression discontinuity evidence from Florida. *Education Finance and Policy, 8*(2), 121-267.

Chakrabarti, R. (2013c). Incentives and responses under No Child Left Behind: Credible threats and the role of competition. *Journal of Public Economics, 110*(1), 124-146.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics, 93*(9-10), 1045-1057.

Cullen, J., & Reback, R. (2006). Tinkering towards accolades: School gaming under a performance accountability system. In T. J. Gronberg & D. W. Jansen (Eds.), *Improving school accountability: Check-ups or choice* (Advances in Applied Microeconomics, Volume 14) (pp. 1-35). Amsterdam: Elsevier Science.

Darling-Hammond, L., & Wise, L. (1985). Beyond standardization: State standards and school improvement. *The Elementary School Journal, 85*(3), 315-336.

Dee, T. (2012). *School turnarounds: Evidence from the 2009 Stimulus*. (NBER Working Paper No. 17990). Cambridge, MA: National Bureau of Economic Research.

Frank, K. A., Maroulis, S., Duong, M., & Kelcey, B. (2013). What would it take to change an inference? Using Rubin's Causal Model to interpret the robustness of causal inferences. *Education Evaluation and Policy Analysis, 35*(4), 437-460.

Figlio, D. N., & Getzler, L. (2006). Accountability, ability, and disability: Gaming the system? In T. J. Gronberg & D. W. Jansen (Eds.), *Improving school accountability: Check-ups or choice* (Advances in Applied Microeconomics, Volume 14) (pp. 35-49). Amsterdam: Elsevier Science.

Figlio, D. N., & Rouse, C. E. (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics, 90*(1-2), 239-255.

Florida Department of Education. (2015). *Florida school grades: School accountability report.* Retrieved January 16, 2015, from http://schoolgrades.fldoe.org/reports

Goldhaber, D., & Hannaway, J. (2004). Accountability with a kicker: Observations on the Florida A+ Accountability Plan. *Phi Delta Kappan, 85*(8), 598-605.

Hahn, J., Todd, P., & van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica, 69*(1), 201-209.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives, 8* (41). Retrieved August 24, 2014, from http://epaa.asu.edu/ojs/article/view/432/828

Heilig, J. V., & Darling-Hammond, L. (2008). Accountability Texas-style: The progress and learning of urban minority students in a high-stakes testing context. *Educational Evaluation and Policy Analysis, 30*(2), 75-110.

Hurlburt, S., Le Floch, K. C., Therriault, S. B., & Cole, S. (2011). *Baseline analyses of SIG applications and SIG-eligible and SIG-awarded schools* (NCEE 2011-4019). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Hirschman, A. O. (1970). *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states*. Cambridge, Mass.: Harvard University Press.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics, 142*(2), 615-635.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics, 89*(5-6), 761-796.

Jacob, R. T., Goddard, R. D., & Kim, E. S. (2014). Assessing the use of aggregate data in the evaluation of school-based intervention: Implications for evaluation research and state policy regarding public-use data. *Educational Evaluation and Policy Analysis, 36*(1), 44-66.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature, 48*(2), 281-355.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. (NCSER 2013-3000). Washington, DC: U.S. Department of Education.

Louis, K., Leithwood, K., Wahlstrom, K., & Anderson, S. (2010). *Learning from leadership: Investigating the links to improved student learning.* New York: Wallace Foundation.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics, 142*(2), 698-714.

Michigan Department of Education. (2012). *School rankings business rules*. Retrieved February 3, 2014, from http://www.michigan.gov/documents/mde/2011-12_School_Rankings_Business_Rules_393915_7.pdf

Mintrop, H. (2004). *Schools on probation: How accountability works (and doesn't work)*. New York: Teachers College.

National Center for Education Statistics. (2016a). *State education reforms: Table 1.3. Rewards and sanctions for schools, by state (2011-2012).* Retrieved February 5, 2014, from http://nces.ed.gov/programs/statereform/tab1_3.asp

National Center for Education Statistics. (2016b). *State education reforms: Table 1.4. Types of school sanctions, by state (2011-2012).* Retrieved February 5, 2014, from http://nces.ed.gov/programs/statereform/tab1_4.asp

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. (2013). Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. *American Economic Journal-Economic Policy, 5*(2), 251-281.

Schochet, P. Z. (2009). Statistical power for regression discontinuity designs in education evaluations. *Journal of Educational and Behavioral Statistics, 34*(2), 238-266.

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J. R., Porter, J., & Smith, J. (2010). *Standards for regression discontinuity designs.* Retrieved November 23, 2014, from http://ies.ed.gov/ncee/wwc/pdf/wwc_rd.pdf

Smith, B. A., Roderick, M., & Degener, S. C. (2005). Extended learning time and student accountability: Assessing outcomes and options for elementary and middle grades. *Educational Administration Quarterly, 41*(2), 195-236.

Stuart, E. A. (2007). Estimating causal effects using school-level data sets. *Educational Researcher, 36*(4), 187-198.

Texas Education Agency. (2015). *Accountability rating system for Texas public schools and districts*. Retrieved October 29, 2015, from http://ritter.tea.state.tx.us/perfreport/account

The Revised School Code Act, 451 of 1976, §§ 380-1280c, Laws of Michigan, 2009.

Trochim, W. M. K. (1984). *Research design for program evaluation: The regression discontinuity approach*. Beverly Hills, CA: Sage.

U.S. Department of Education. (2009). *Guidance on School Improvement Grants Under Section 1003(g) of the Elementary and Secondary Education Act of 1965*. Washington, DC: Author.

U.S. Department of Education. (2012). *ESEA flexibility*. Washington, DC: Author. Retrieved May 21, 2015, from http://www.ed.gov/esea/flexibility

U.S. Department of Education. (2015). *School Improvement Grants*. Washington, DC: Author. Retrieved February 23, 2016, from http://www.ed.gov/category/program/school-improvement-grants

Wing, C., & Cook, T. D. (2013). Strengthening the regression discontinuity design using additional design elements: A within-study comparison. *Journal of Policy Analysis and Management, 32*(4), 853-877.

Winters, M. A., Trivitt, J. R., & Greene, J. P. (2010). The impact of high-stakes testing on student proficiency in low-stakes subjects: Evidence from Florida's elementary science exam. *Economics of Education Review*, *29*(1), 138-46.

Winters, M. A., & Cowen, J. M. (2012). Grading New York: Accountability and student proficiency in America's largest school district. *Educational Evaluation and Policy Analysis*, *34*(3), 313-327.