UNIVERSITY OF CALIFORNIA,
IRVINE

Failing Grades: Examining The Long-Term Effects of Failure in Education

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Sociology

by

Tanya Natasha Sanabria

Dissertation Committee:
Professor Andrew Penner, Chair
Associate Professor Thurston Domina
Professor Evan Schofer
Assistant Teaching Professor Jacob Avery

2019

DEDICATION

To

my family and friends

in recognition of their worth

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor and committee chair, Professor Andrew Penner, whose selfless time and care inspired me to finish. With the attitude and grace of a genius, he continually encouraged lines of research yet unexplored, and easily spreads that excitement to everyone he mentors. Without his guidance, persistence, and understanding, this dissertation would not have been possible.

Along with Professor Penner, Professors Thurston Domina and Jacob Avery have been amazing mentors from the beginning of my graduate program, providing much needed feedback, encouragement, and inspiration to keep me going during the times where I didn't believe in myself. And I am also extremely grateful to Professor Evan Schofer, for his incredible generosity and words of encouragement.

Outside of my committee, I received much in the way of mentorship from a variety of places, and would like to especially thank Professor Ann Hironaka for her unending patience with my work as I struggled to tell a coherent story, from start to finish. I would also like to thank Professor Judith Treas for her support and wisdom as I developed my dissertation. Outside of UC Irvine, I would like to thank the gracious mentorship and guidance received from Professors Yasmiyn Irizarry, Chandra Muller, Kelly Raley, and Catherine Riegle-Crumb at UT Austin. I am also indebted to my undergraduate mentor Professor Mary Yu Danico, whose enthusiastic support and never-ending guidance put me on a path I never would have thought possible.

My friends and classmates have been a continual source of support and encouragement. If it were not for writing sessions and happy hours, I don't think I would have survived graduate school without my sanity intact. I cannot imagine a better support group in graduate school than friends like Bonnie Bui, Edelina Burciaga, Miles Davison, Alma Garza, Ben Gibson, Martin Jacinto, Jess Lee, Joseph King, Jessica Kizer, Setareh Mahmoudi, Hector Martinez, John McCollum, Anna Penner, Stephanie Pulles, and Emma Smith. Outside of graduate school, I could not imagine the journey through graduate school without my best friends, who are family to me: Shabnam Azari, Danielle Madrigal-Upchurch, Jonathan Serrano, and Nazira Paulette Taylor. I cannot begin to express my gratitude for your unending support.

Finally, thank you to my family. Words are not enough to express my gratitude for the constant source of support and inspiration to keep me going. Sarah, you have taught me to always remember what's important in life from the day I held you in my arms. I'm fortunate to witness the young intelligent woman that you've become. I couldn't be more proud of you. To Deguie (Azalie), Anthony, Leonel and my parents, your love illuminates my life everyday.

Financial support was provided by the University of California, Irvine, 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 award number K01HD073319, 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.

CURRICULUM VITAE

Tanya Natasha Sanabria

2012	B.A. in Sociology, California State Polytechnic University, Pomona
2012-15	Teaching Assistant, School of Social Sciences, University of California, Irvine
2014-15	Pedagogical Fellow, Division of Teaching Excellence and Innovation, University of California, Irvine
2014-15	Research Assistant, ADVANCE Program, University of California, Irvine
2016-17	Research Assistant, Center for Administrative Records Research and Applications, U.S. Census Bureau
2017-18	Research Assistant, Center for Administrative Data Analysis, University of California, Irvine
2019	Ph.D. in Sociology, University of California, Irvine

FIELD OF STUDY

Sociology of Education

PUBLICATIONS

"Is free and reduced-price lunch a valid measure of educational disadvantage?" *Educational Researcher* 47(9): 539-555, 2018.

"Weeded Out? Gendered responses to failing calculus." *Social Sciences* 6(2): 47-61, 2017.

"Marital Status and Union Formation across the Life Course". Pp. 247-269 in *Gerontology Changes, Challenges, and Solutions*, edited by Madonna Meyer and Elizabeth Daniele. Santa Barbara, CA: ABC-CLIO Publications, 2016.

"Reactions to Failure." Pp. 664-665 in *The SAGE Encyclopedia of Economics and Society*, edited by Frederick F. Wherry: Thousand Oaks, CA: SAGE Publications, 2015.

ABSTRACT OF THE DISSERTATION

Failing Grades: Examining the Long-Term Effects of Failure in Education

By

Tanya Natasha Sanabria

Doctor of Philosophy in Sociology

University of California, Irvine, 2019

Professor Andrew Penner, Chair

Academic course failure is perhaps surprisingly common in the US education system. Studies in educational development suggest that course failure has universally negative effects on students' educational outcomes. However, it is unclear the degree to which these negative results may be driven by differences between students who do and do not fail, and whether all students are equally impacted. My dissertation examines how course failure and academic labels impact young adults' academic and labor market outcomes, focusing on differences across student characteristics, and using quasi-experimental research designs to account for differences in prior academic or demographic characteristics between students who do and do not fail.

Drawing on research in life course theory, I argue that failure more broadly should not be understood as aberrant but as an important part of the developmental process. Using three large, longitudinal data sources with academic transcript information, and employing quasi-experimental methods where possible, I demonstrate that course failure can compound advantages or disadvantages to have substantial impacts on future outcomes that vary by gender and race. Chapter 2 uses students' college transcripts from the National Education Longitudinal Study (NELS 88) to show that failing a so-called "weed out" course discourages women from pursuing STEM majors, but that men who fail major in STEM fields at similar rates to those who

pass. In Chapter 3, I use transcript data from the National Longitudinal Survey of Youth (NLSY 97) to show that while college remedial coursework benefits some students, the substantial number of students who fail remediation are considerably worse off (e.g., they are less likely to graduate, take longer to graduate, and earn less) than peers who were not placed in remediation. Finally, Chapter 4 uses statewide administrative data to show that students with a more negative performance label have lower test scores and worse behavioral outcomes over time. My dissertation thus underscores how students who failed in school will likely access vastly different opportunities in education, the labor market, and other institutions.

CHAPTER 1: INTRODUCTION

Course failure is common in the contemporary US educational system. In California, for example, nearly half of 8th graders fail Algebra (Liang, Heckman, and Abedi 2012). Likewise, over half of college students have failed a course at some point during their academic career (NELS PETS: 2000). While research suggests that students who have failed a course experience negative effects, such as future lower achievement and behavioral issues (Andrew 2014; Jimerson 2001), it is unclear the degree to which these differences may be driven by differences between students who do and do not fail, and whether these differences occur across different contexts (Allen et. al 2009; Reschly and Christenson 2013). Little work has thoroughly examined how course failure can have meaningful long-term differences for students' educational experiences.

Drawing on research in life course theory, human development, and social psychology, I argue that failure more broadly should not be understood as aberrant, but rather as an important part of the developmental process. For example, when young toddlers fall, we do not view this failure at walking as problematic, but rather as a part of the growth process. Yet academic course failure is defined as problematic, and as a result plays an important role in widening the gap between marginalized and privileged students. This is because while failure is part of the learning process for all students, in an educational system that defines failure as problematic, students from advantaged backgrounds can marshal resources that allow them to do their failing (and the attendant learning) outside of school, hiding much of their failure from this institutional context. Students from disadvantaged backgrounds, by contrast, do not have this luxury, and as a result more of their learning (and failure) happens in the context of formal schooling. When educators place a limit on the amount of failure within schools that is tolerated before students

are categorized as failing, and social and academic consequences are attached to this, we observe differences in both who is more likely to fail in school and whose failures are more costly.

My dissertation thus examines how course failure shapes young adults' educational and labor market trajectories across demographic and institutional contexts, moving beyond one-size-fits-all understandings of the effects of failure. Using three separate educational datasets, and employing quasi-experimental methods where possible, I show that course failure does not uniformly affect students, where those who were at risk before failure are hurt the most for future outcomes. Ultimately, my research shows why and how the effects of failure vary across contexts, generating new insights about how inequality operates in education throughout the life course.

THEORETICAL FRAMEWORK

Drawing on research in life course theory, development psychology and social psychology, I argue that failure should be understood as aberrant, but rather as an important part of the developmental process. Then, I argue that modern public discourse around failure, particularly the concept of "grit" neglects to consider the structural barriers that students face in schools. Instead, within the context of schooling as a sorting process, students in privileged positions have the opportunity to fail outside of school, thus hiding much of their failure within an environment that assigns academic and social consequences to failure, while students in disadvantaged positions are able to do so. Lastly, I use the concept of "turning points" and "scarring" from life course theory to argue that course failure has the potential implications for impacting students beyond the classroom and into their transitions into adulthood.

Categorical Inequality of Failure

In this section, I provide an overview of how failure is perceived (and encouraged) in human development, while failure in education is interpreted as problematic. Recently, educators and researchers believe the answer lies in encouraging the student to develop perseverance in the face of failure. However, I argue that this belief ignores the structures that not only facilitate the categorization of students in "successful" and "failed" categories, but also reproduces these inequalities by sorting students and correspondingly allocating institutional resources based on these categories.

Developmental scholars recognize that children first must fail before they can succeed in a task as part of their growth. For example, a longstanding puzzle in early childhood development literature is that, when an infant can crawl excellently, why would an infant take a

risk to adopt a new strategy that is as unstable and unknown as walking? While there is no unifying theory as to what motivates a child to make these changes, Adolph et al. (2012) sought to uncover why children learn to walk, considering that novice walkers fell more per hour than expert crawlers. Their findings suggest that walking covers more distance in travel than crawling. However, toddlers and parents do not view falling while trying to walk as a bad thing, but rather a normal progression from crawling to walking. This is a familiar pattern with many developments through infancy and childhood, where children will adopt new strategies for executing a task that is initially more difficult than their current strategy (Centers for Disease Control and Prevention 2018). Learning a new strategy, such as walking, is accompanied by more frequent falling (i.e., failing at walking) but through this process, a child will fall less over time to eventually walk successfully. Thus, failure is not atypical when a child is practicing a new strategy for a task, but rather a necessary component to gain the confidence to execute a new and risky strategy.

Recently, educators and parents believe that students should become "gritty", or continually persistent, in the face of failure. Duckworth, Peterson, Matthews, and Kelly define grit as "passion and perseverance for very long-term goals" (2007, p. 1087). Testing the concept of grit among new cadets at the United States Military Academy, Duckworth and her colleagues developed a grit scale based on a series of twelve statements (e.g. "I finish what I start) to successfully predict which cadets would finish the training and who would drop out. Duckworth (2013) suggests that, where cognitive ability falls short in predicting success, grit can help explain why some students perform better than others, and the success of highly accomplished individuals can be explained by their possession of grit.

This interpretation of grit parallels troubling aspects of cultural deficit beliefs in the public discourse about education, which refers to the notion that students (particularly those from low income or racial/ethnic minority backgrounds) fail in school because these students and their families have internal deficits that hinder the learning processes (e.g. "lack of motivation"). The deficit model suggests that academic failure is rooted in the students' cognitive and motivational deficits, rather than examining how structures within institutions thwart student learning and block future opportunities for success (see Bereiter and Engelman 1966; Valencia 1997). Likewise, Duckworth's grit theory has been criticized for failing to consider inequalities faced by low-income students, students of color, and students with disabilities (Denby 2016; Kohn 2014; Nathan 2017; Rose and Ogas 2018). Duckworth's grit theory implies that, for students who face barriers and fail, they simply do not have the "grit" that their successful (and more likely advantaged) peers possess. On the contrary, students of color, with disabilities, or those living in poverty do not lack grit, but rather display grit to navigate barriers in their communities, homes, and schools that goes unrecognized and unrewarded in the contemporary educational system.

In sociology of education, schools are considered structural sites that create important social categories, sort students based on these categories, and assign social and academic consequences based on this categorization (Domina, Penner, and Penner 2017). The definition of failure (and success) in an academic subject is at the discretion of the institution, not universally defined. Moreover, educators place a limit on the amount of failure that is tolerated within schools before students are categorized as failing. When schools place students in the "fail" category, students then face tangible and long-lasting consequences, facilitating the production (and reproduction) of social inequity within, between, and outside of schools. This is because placement in the social and academic category of failure shapes the educational resources and

incentives in which students are exposed, and students who failed a course will likely access vastly different opportunities in education, the labor market, and other institutions.

Privileged students have opportunities to fail (and learn) outside of formal schooling (e.g. private supplementary tutoring known as "shadow education") (Entrich 2017). Students from disadvantaged backgrounds, by contrast, do not have this privilege, and as a result more of their learning (and failing) happens in the formal school setting. For example, we can imagine two fourth graders (Student A and B) struggling to grasp the concept of fractional equivalencies. Student A has access to tutoring services and parents who attained college degrees; thus, Student A is able to continually fail and learn from their failure in a low-stakes environment (such as with their tutor or parents). In this way, Student A is afforded the opportunity to fail "back-stage," gaining the advantage to succeed in the "front-stage" (i.e. formal schooling). Their advantages are then rewarded and compounded throughout education and beyond. On the other hand, Student B does not have access to similar resources as Student A; therefore, Student B fails more in their school. From a teacher's perspective, Student B appears to be struggling more than Student A, as more of their failure takes place in the context of formal schooling where it is visible to the teacher. If Student B continues to fail and does not master a skill in the time allotted in formal schooling, their failure can compound in two ways. First, to the degree that skill accumulation compounds, with early learning providing the basis for later learning (Siegler et al. 2012), Student B will be disadvantaged by their incomplete foundation when they are attempting to learn new material that builds on areas that they have yet to master. For example, Student B may not only be penalized for not displaying proficiency in fractional knowledge, but the initial disadvantage of struggling with this mathematical concept means that Student B will have a harder time understanding how to add or subtract fractions, falling behind their peers.

Without intervention, it is likely Student B will remain behind their peers academically. Second, Student B's teacher may formally or informally label them as a student who fails more than their peers, resulting in course failure, grade retention, or academic dismissal.

While in many contexts in human developmental failure facilitates the adoption of new strategies and the development of confidence to use these new strategies, failure is often characterized as problematic in student learning. A recent development in the educational public discourse suggests that students do not possess enough "grit" when faced with failure. However, I argue that grit theory in the current public discourse is reminiscent of earlier arguments posited by cultural deficit models (e.g. "racial/ethnic minorities lack the motivation to succeed in school"). Focusing on grit ignores the issues posed by social inequality in education - the biases, barriers, and lack of privilege faced by students of color, those living in poverty, or with disabilities. Rather, I argue that schools explicitly define the categories of "success" and "failure" (which are not universally agreed upon) at their discretion, and sort these students accordingly. Thus, schools give meaning to these categories through placing a limit on how many students can succeed (and thus creating a structure that arguably necessitates the placement of students into a "failure" group). That is, once failure is defined as a category, this category takes on a life of its own independent of a particular student population's range of achievement, and the utility and logic of this category allow the educational system to see failure in a population and utilize this category, even if sparingly so (c.f., McDermott 2001). Resources are then allocated based on this category, shaping the educational resources and incentives to which students are exposed.

Failure as a Turning Point in the Life Course

In this section, I draw on life course theory to provide an overview of the concepts of timing (the social and developmental implications of life events) and turning points (a significant life event that changes the direction of an individual's trajectory) within the life course. Despite its theoretical and empirical importance, the concept of turning points has rarely been explicitly utilized in the educational context. Drawing from life course theory and the concept of "scarring" in developmental psychology, I conceptualize course failure as a harmful turning point within an individual's educational career, which have long-lasting consequences after the student has left the classroom.

The life course paradigm serves as an ideal framework for examining education an understanding the nexus of social pathways, developmental trajectories, and social change over time. For the purpose of my dissertation, I draw heavily on the principle of timing – the social and developmental implications of events in the life course, which can vary based on when these events occur in a person's life and the developmental stage in which this event occurs (Elder, Johnson and Crosnoe 2007). I also refer to the concepts of trajectories, transitions, and turning points in the life course model. Trajectories refer to the socially organized pathways that individuals and groups move through institutions in the life cycle, such as education, work, and family. These trajectories are shaped by historical forces and often structured by social institutions, although these pathways can be altered from the impact of broader contexts (e.g. war) and from a demographic shift of populations entering or exiting these pathways (e.g. increased funding for higher education). Trajectories are comprised of transitions, or changes in social status or identity (e.g. entering the labor market, or becoming a parent). A turning point is a particular event or experience that results in changes in the direction of a pathway or transition

in the life course (Elder 1998). A significant turning point in an individual's life can potentially have lifelong implications through the accumulation of advantages or disadvantages – a cascade of positive or negative events and influences over time, which depends on both the social structure in which the turning point occurred and how the individual responds to the circumstances.

Despite its theoretical importance, turning points have rarely been studied in the context of the educational career. This is puzzling, given that schooling can be viewed as a sequence of developmental environments through which students move (Attewell and Domina 2008). Moreover, the U.S. contemporary educational system has adopted a high-stakes approach by attaching test scores or other singular indicators of academic ability to significant promotions, such as advancing to the next grade or entering a particular type of postsecondary school (Alon and Tienda 2007; Hanushek and Raymond 2005). While we can consider school contexts as directly influencing the acquisition of human capital through the development of skills, life course theory emphasizes how these school contexts also shape later life outcomes by altering students' pathways into various future educational contexts and trajectories (Elder, Johnson and Crosnoe 2007). Thus, conceptualizing failure in school as a turning point in the life course can provide insight into the complicated underlying processes of how individuals move through education into other social institutions, such as the labor market. Examining failure as a turning point in the educational career may also reveal why, for instance, the same life event (e.g. failing a remedial course in college) can have harmful consequences for some, but not for others.

Findings from prior research suggest that turning points in the educational career (e.g. primary school grade retention) can have lasting impacts on future educational attainment. While not referred to as a turning point, Andrew (2014) argues that "scarring" is an important form of

cumulative advantage in the life course, brought about by a triggering event such as grade retention in primary grade school. This triggering event shifts an individual's status in a given hierarchy (e.g. student) to a new status (e.g. retained student) that impacts subsequent outcomes, even after the individual has moved on from this new (and usually temporary) status (DiPrete and Eirich 2006). For example, being retained at an early age reduces the odds of high school completion, even if the student is able to recover academically in high school (Andrew 2014). In this way, the initial differences in being retained at an early age magnifies over time, making it difficult for a student who is academically "behind" to catch up with their peers. However, it is important to note that harmful turning points do not affect students in the same way, nor does it always result in long-lasting scars (Jacob and Lefgren 2009).

Since the schooling process is organized as a sequence of formative environments (Attewell and Domina 2008), and since turning points have important long-lasting implications (Elder, Johnson and Crosnoe 2007), I expect that a harmful turning point, such as failing a course, will negatively affect transition to adulthood outcomes (Andrew 2014).

PREVIOUS RESEARCH

Although prior research has extensively studied what leads to academic success and its effects thereafter, relatively little is known about the effects of course failure, particularly during important transitions in the life course (e.g. eighth grade or first year in college). In the following section, I review prior literature on the effects of course failure, highlighting two central findings: (1) failure is not always associated with negative outcomes, as is often assumed, and (2) there is evidence suggesting that the effect of failure is heterogeneous across demographic groups, course topic, institutional settings, and time points within an individual's educational trajectory.

Contrary to popular belief that failure can have devastating effects, research in the fields of psychology and education do not provide a clear answer as to whether course failure produces universally harmful outcomes. For example, failure in an academic setting has been demonstrated to have both detrimental and positive effects upon subsequent performance. One study finds that failure can have both facilitative and harmful effects on future performance dependent on the perceived importance of the given task (Roth and Kubal 1975). Moreover, researchers hypothesize a curvilinear relationship between failure and future performance as opposed to a decreasing linear relationship (Brehm and Brehm 1981; Clifford 1979; Wortman and Brehm 1975). In other words, failure in moderation can actually optimize future performance depending on the context in which the academic failure occurred.

Given that the effects of failure depends on the perceived importance of the task (from the student and/or the educational institution, in what contexts is course failure deleterious or facilitative? To answer this question, I review prior research on grade retention in K-12 education, or the practice of having a student repeat a year of schooling due to course failure. In

doing so, I argue that the mixed findings present in the grade retention literature point to the greater puzzle of the effects of course failure.

Much attention in educational research has been dedicated to evaluating the efficacy of retention. However, empirical studies on the effects of grade retention have yielded inconsistent conclusions. The literature commonly characterizes grade retention as harmful and negative (Dennenbaum and Kulberg 1994; Jimerson 2001a; Frey, 2005). The most frequently cited meta-analyses conducted by Holmes (1989) and Jimerson (2001b) conclude retention has almost "universal negative effects", pointing to lower achievement levels and/or disciplinary problems among retained students than continuously promoted students (Reschly and Christenson 2013). Furthermore, repeating a grade in elementary school has been identified as a significant predictor for future poor academic adjustment in high school and dropping out (Stearns, Moller, Blau, and Potochnick 2007; Gottfried 2013). The empirical evidence demonstrates negative consequences of failing a course; however, it is unclear that retention is "universally negative" as claimed.

Recent research has challenged the findings that retention is "universally negative," referring primarily to the methodological limitations of these studies (Alexander, Entwisle, and Horsey 2002; Allen, Chen, Willson, and Hughes 2009; Lorence and Dworkin 2006). The most difficult obstacle in examining retention is to determine its causal effects in the absence of a randomized study. Moreover, factors, such as family socioeconomic background, associated with the treatment (i.e. retention) are also associated with the measured outcomes (Reschly and Christenson 2013). Other studies examine the effect of retention only after the retention had occurred without accounting for pre-retention characteristics. For example, students displaying lower levels of effort after retention may have very well had lower levels of effort before

retention. Due to these potential selection biases, scholars cannot adequately ascertain the effects of course failure.

Given that randomly assigning students in grade retention is neither ethical nor feasible, previous studies try to control for pre-retention by selecting a group of similarly low achieving students who were promoted in the subsequent year. However, this approach ignores differences not captured in the pre-retention measures related to academic performance or grade retention placement (e.g. behavioral issues, special education sorting, ability group tracking). Alexander, Entwisle, and Dauber (2003) point out that if only retained children are followed, the retainees' performance could reflect a general decline in performance with age that would be observed for students, regardless of grade retention. The authors also point out that grade retention is examined separately from administrative sorting into special education classes, ability groups, and curricular placements when these interventions occur simultaneously in a student's experience. Thus, it is vital for researchers to understand that grade retention is correlated with characteristics associated with, but does not directly cause, negative academic and socio-emotional outcomes.

Studies employing strong methodological designs (based on the quality of the comparison group and the statistical control) find no significant difference for retention on achievement (Allen et. al 2009). Moreover, studies that compared same-age peers found that achievement levels declined less steeply than studies that compared same-grade peers. These results indicate that retained students seem to gain a boost in achievement compared to their younger same-grade peers but lose this advantage over time. However, researchers should exercise caution to claim that repeating a year can help than harm students. Rather, retention appears to have varying effects dependent on what group of students is examined. For example,

Alexander, Entwisle, and Dauber (2002) find that retention appeared to help first-graders who were not too far behind *before* retention. However, retention appeared to least help students who were already substantially struggling academically before retention, as well as those who would eventually be placed in special education or a second retention. Therefore, future research on failure must consider who is hurt by failure more, and at which time points failure has its most pronounced effects.

For example, would course failure matter differently among students who do not have to repeat a year of schooling and who are not closely monitored by an educational institution? In the case of higher education, students who fail courses are not required by the university to repeat the failed course. Additionally, a failing grade in a university setting represents a summary of the students' performance in a course or test. Given the absence of an all-encompassing intervention program like grade retention in higher education, researchers tended to examine the consequences of failure at the individual level. Thus, the literature on the effects of failure in college is sparse, briefly speculate its effects, or limit the scope to short-term outcomes.

In higher education, studies have found gender differences in causal attribution for success and failure. Women are more likely to attribute failure to lack of ability as opposed to task difficulty (Beyer 1998; Nelson and Cooper 1997; Ryckman and Peckham 1998; Sweeney, Moreland, & Gruber 1982; Wortman and Brehmn 1975) and experience relatively more stress following failure on a task (Wortman and Brehmn 1975). For example, women were also more likely to "feel like a failure" after receiving an imaginary F on an exam (Beyer 1998). These findings have implications beyond individual stress levels. Correll (2004) found that gender status beliefs contribute to a gender-differentiated double standard for attributing to performance to ability or task difficulty, accounting for actual ability in the participant. In her study, men and

women differently assessed their own competence to tasks that are relevant for future careers. These assessments then shape how students view themselves and pursue specific career paths or activities, based on their assessment of competence in these tasks. While researchers studying motivation and attribution of causality acknowledge that there are differences in response to failure, scholars need to continually improve our understanding on the effects of failure through considering differences across gender, race, family background, and institutional characteristics.

Prior research has demonstrated that the risk of course failure in high school varies by students' race, ethnicity, and gender as well as prior achievement (Mickelson 1989; Roderick and Camburn 1999; Riegle-Crumb 2006). Moreover, institutional and school climate factors, such as school size and instructional policies, can shape the distribution of student achievement and educational attainment independent of individual student characteristics (Rumberger and Palardy 2005; Rutter 1985; Werblow and Duesbery 2009). Following this, we should assume that the effects of failure would also vary as a function of demographic and institutional characteristics. For example, Riegle-Crumb (2006) finds that among female students, failure had less of a negative effect for African American females compared with white females. This could be interpreted as greater resiliency among African American female youth on their performance or that African American female youth may be more immune to any type of institutional feedback, positive or negative. Women were less likely to continue as an economic major after poor academic performance than men (Rask and Teifenthaler 2008; Owen 2010) and may be drawn to higher grades in social science or humanities courses (Ost 2010). These findings suggest that the consequences of failure are not uniform across all student groups.

Overall, prior research on the effects of course failure is sparse and disparate across fields of study. Previous studies had to contend with major methodological limitations – finding a

comparable group to students who failed a course in an educational setting. Moreover, a nontrivial number of studies have not adequately addressed these major methodological limitations in studying course failure, concluding that failure is deleterious across student groups and across institutions. These answers do not sufficiently take into account variation across demographic groups, schooling contexts and transitions, and institutional settings. Additionally, it is imperative that sociological research provides a unifying framework to understand the effects of course failure and provide insight on important educational processes. Thus, my dissertation seeks to answer three main questions:

(1) What is the relationship between failure in a course and future student outcomes?

(2) How do the effects of failure in school vary by demographic and institutional characteristics?

(3) In what contexts does failure matter more for future outcomes?

The dissertation is organized as three separate but interrelated studies that share a core focus on the consequences of course failure and its varying effects. Each chapter uses a separate, large-scale longitudinal data source for the dissertation. In Chapter 2, I use the National Education Longitudinal Dataset (NELS) to examine the relationship between course failure in college and degree completion, whether these relationships vary by race and gender, and whether failing different types of courses matter differently for these outcomes. I find that women who fail a course have a lower likelihood of completing a bachelor's degree compared to men, but find no racial or course subject differences. Chapter 3 uses data from the National Longitudinal Survey of Youth (NLSY) to examine whether the relationship between college remediation and degree and labor market outcomes is moderated by whether students fail remedial coursework. I find that students who failed their remedial coursework had substantially lower odds of degree

completion and earned less than those who passed their remedial coursework. Finally, Chapter 4 uses statewide school data from Oregon to estimate the impact of receiving a negative performance label on reading test scores in third grade on future academic and behavioral outcomes from 2004 through 2015. I find that students who were assigned a negative performance label have lower test scores and worse behavioral outcomes over time. [1]

I would like to note that Chapter 4 does not examine course failure in the same way that Chapters 2 and 3 examine course failure. However, we can think about Chapter 4 as failing to meet the standard for that particular year among third graders, and that this analysis enables me to specifically examine the effects of these kinds of labeling practices without other consequences, such as repeating a course or grade, academic probation, or other signals that mark course failure.

CHAPTER 2: Weeded Out? Gendered Responses to Failing Calculus

Abstract

Although women graduate from college at higher rates than men, they remain underrepresented in science, technology, engineering, and mathematics (STEM) fields. This study examines whether women react to failing a STEM weed-out course by switching to a non-STEM major and graduating with a bachelor's degree in a non-STEM field. While competitive courses designed to weed out potential STEM majors are often invoked in discussions around why students exit the STEM pipeline, relatively little is known about how women and men react to failing these courses. We use detailed individual-level data from the National Educational Longitudinal Study (NELS) Postsecondary Transcript Study (PETS): 1988–2000 to show that women who failed an introductory calculus course are substantially less likely to earn a bachelor's degree in STEM. In doing so, we provide evidence that weed-out course failure might help us to better understand why women are less likely to earn degrees.

Keywords: higher education; gender; STEM; inverse probability weighting

Introduction

A longstanding body of research on gender differences in education suggests that women are underrepresented in many science, technology, engineering, and mathematics (STEM) fields—particularly in the physical sciences and engineering (Xie and Shauman 2007). Research seeking to understand gender differences in who majors in a STEM field has identified a plethora of factors, ranging from discrimination, cultural stereotypes around gender and science, confidence, peer networks, and a preference for flexible curricula not offered in STEM departments (Correll 2001; Charles and Bradley 2009; Cech et al. 2011; Riegle-Crumb 2006; Mann and Diprete 2013). Underlying much of this research is the notion that STEM undergraduate training occurs in an environment that ranges from disengaging to competitive to chilly, and that this climate leads students to opt for other fields (Seymour and Hewitt 1997; Niederle and Versterlund 2007). While the factors that contribute to this climate are likewise numerous, competitive weed-out courses at the introductory level are a source of considerable dissatisfaction among undergraduates (Seymour and Hewitt 1997). These courses serve a gatekeeping function, as they are required for many STEM majors, and are often failed by a substantial number of students, promoting a competitive "sink or swim" environment (Seymour and Hewitt 1997; Kokkelenberg and Sinha 2010; Olson and Riordan 2012).

Importantly, both women and men see this as problematic. The women interviewed by Seymour and Hewitt express their thoughts like "I knew I could have done it if I wanted to. But I just said 'Do you really want to do this? Is it really worth killing yourself for?'" or "It's been unadulterated hell. Major overloads, no rest, stress—and it's getting worse. That's why I'm looking elsewhere" (Seymour and Hewitt 1997, pp. 202–3). Men's assessments are largely similar: "I mean, why stay [in science]? You know, there's no reason. And the rewards are—

there's no rewards. I mean, I can see no logical reason why you'd stay." and "You go through hell in the sciences without any guarantee that you will be able to work. Why do it? Why not be an English major?" This sentiment is summarized by Meg Whitman, who noted in an interview that "I took calculus, chemistry, and physics my first year. I survived. But I didn't enjoy it…After that, I had to find something else to do. I began selling advertising for a magazine that was published by Princeton undergrads. It was more fun than physics" (Fishman 2001).

However, despite the fact these weed-out courses are often invoked by students as a significant source of disengagement, surprisingly little is known about how undergraduates respond to failing these courses. While not examining weed-out course failure per se, research on grade inflation suggests that failing a weed-out class could play an important role in shaping students' future majors. One study, for example, found that students were "pulled away" by their higher grades in the humanities, arts, and social sciences courses and "pushed out" of STEM because of lower grades (Ost 2010). Grade inflation in introductory classes may be particularly important, as the grades that students receive in introductory courses strongly predict whether students choose to enroll in more courses in the discipline (Ost 2010). Introductory courses in STEM departments tend to be among the lowest graded courses (Rask 2010). Simulations suggest that if the grading distribution in introductory science courses resembled the college average, there would be 2–4 percent increase in advanced science course taking in later semesters (Rask 2010).

We build on this research by examining whether there are gender differences in the rates at which men and women fail introductory calculus (which we henceforth refer to simply as calculus), and how they respond to failure. Calculus often serves a gatekeeping function across STEM disciplines, limiting the rate at which students can take advanced coursework in their

major. Introductory math courses, such as calculus, were found to be important factors for students' decisions to stay or switch out of STEM (Chen 2013). Although several studies have indicated that performance in introductory courses has been linked to STEM persistence, little attention has been given to failing weed-out courses like calculus. A key limitation in previous research is that these studies pool grades across STEM courses, using GPA as an indicator of poor performance. While important, these studies cannot ultimately address the role of weed-out course failure. Given the important signal that failing a weed-out course provides to students (Crisp et al. 2009), we argue that examining the gendered responses to calculus failure can provide researchers a better understanding of the critical junctures that shape a student's academic trajectory.

Gender might play an important role in shaping how students respond to failing calculus given societal stereotypes about math competence. Correll (2004) shows that beliefs about gender differences in a domain can shape self-assessments of competence and interest in pursuing a career using these skills. Specifically, when women are exposed to the belief that men are superior in a particular domain, women rate their performance worse than men, even when men and women receive identical feedback about their actual performance in the domain. Given widespread stereotypes about gender differences in mathematics, Correll's findings suggest that women who fail a calculus course might perceive their math skills to be worse than men who fail, and might have less interest in pursuing math-dependent careers. Gender differences in self-assessments driven by these stereotypes may explain why women tend to express doubts in their mathematical skills (Charles and Bradley 2009; Noel-Levitz 2014) and are more likely to switch to a female-typed major when receiving lower grades in coursework (Rask and Tiefenthaler

2008). As Charles and Bradley (Charles and Bradley 2009, p. 926) note, "Beliefs about gender difference can thus spawn powerful self-fulfilling prophesies".

While previous research suggests that women are more likely to re-evaluate and change their career pathways in response to negative feedback, we know of no study that has examined the implications of calculus failure and gender differences on whether students major in STEM. This study uses a doubly robust inverse probability weighting approach to compare the degree outcomes of students who had taken and failed calculus to a comparison group who passed calculus. We thus provide the first examination of the potentially gendered ways in which students responded to failing weed-out coursework.

Research Questions

Our key research question examines whether there are gender differences in the response to failing calculus, focusing on students' likelihood of completing a bachelor's degree, and in particular, on degree completion in a STEM field. To motivate the analyses for our central research question, we first ask (1) who takes and who fails calculus? Then, we ask, (2) what are the schooling outcomes associated with failing calculus? Finally, we address our key question, (3) are there gender differences in the schooling outcomes associated with failing calculus? To understand how failing a weed-out class may affect students in the STEM pipeline (i.e., those who may be considered at risk of majoring a STEM field), we narrow our sample size for questions (2) and (3) to students who planned to major in STEM as high school seniors.

Data

Data are from the National Education Longitudinal Study (NELS:88) and the NELS Postsecondary Education Transcript Study (PETS:2000) (NCES 1988; NCES 2000). The NELS:88 is a longitudinal study that followed a representative sample of 25,000 eighth-grade students over twelve years starting in 1988. The Educational Testing Service created pencil-and-paper tests to assess each eighth-grader's skills in reading and mathematics for the NELS:88. These tests were repeated in tenth, and twelfth grades. We use the student's percentile rank in the pencil-and-paper test in twelfth grade to measure students' pre-college academic skills in reading and math.

During each follow-up survey, additional data and interviews were collected from parents, teachers, and students participating in the study. As a longitudinal panel study, NELS:88 experienced sample attrition and non-response bias. To adjust for the sampling frame, the NELS:88 replenished the sample with additional respondents. All analyses thus use weights to adjust for these differences and students in the analyses were non-missing in key outcome, predictor, and control variables.

The fourth and last follow-up study of NELS:88/2000 for the sample of the eighth-grade class of 1988 occurred in 2000. The study collected postsecondary education transcripts for the sample members who responded to the final follow-up and reported attendance at a postsecondary educational institution in the third (1994) or fourth (2000) follow-up. Approximately 16,020 postsecondary transcripts were collected for 15,240 sample members, a subsample from the third follow-up. Transcripts contained detailed information on students' coursework, credits, grades, and degree obtained. To examine postsecondary education outcomes, we restricted our sample to the base-year through fourth follow-up studies, limiting the number of valid cases with a postsecondary transcript record to 7050 individuals.

Measures

Our key independent variable is failing an introductory calculus course, a key gate-keeping course that often serves as a requirement for STEM majors. Calculus courses were identified using the 2010 College Course Map (CCM) taxonomy system to code information on the course subject and title from college transcripts. Students were coded as having failed a class if they both (1) received a grade of "0" or "F" for the course and (2) reported zero earned credits for the course. We ran additional analyses where we define failure to include grades of "D", "D-", and "F". Findings were consistent with results from analyses reported here.

The two main outcome variables in this study are whether a student completed a bachelor's degree and whether they graduated with a bachelor's degree in a STEM field. STEM majors include engineering, mathematics, physics, chemistry, and biology; a complete list of majors included as STEM fields is available in Appendix A (Table A1). The degree type and major is reported on the student's transcript at collection.

We also control for a wide range of variables. Student-level controls include race/ethnicity, gender, socio-economic status, high school GPA (standardized), twelfth grade test score percentile ranks in both reading and math, whether students planned to major in STEM as high school students, and the highest math course taken while in high school. During students' senior year of high school, students were asked if they expected to attend college and in which field they expected to major; we collapsed anticipated majors into an indicator for whether students planned to major in a STEM field. While we would ideally use a measure of intended major from the fall when students entered university, we prefer our measure from the senior year of high school to information collected in the third follow-up of NELS:88 in 1994, when most students were in their second year of college.

We also control for whether the student's primary institution was a public two-year, private not-for-profit four-year, and a public four-year institution. Because some students move from one college to another, we coded for the first college that a student entered after high school. Accounting for observable differences on these dimensions helps ensure that the associations we observe between failing calculus and degree receipt are not being driven by these factors.

Sample

The first column of Table 1.1 provides a summary of the controls and outcome measures, as well as the number of students who took calculus and the number of students who failed (*n = 3650)*. The study sample has slightly more women (52.6 percent) than men (47.4 percent). The sample consisted of primarily Non-Hispanic White (74.5 percent), with 7.5 percent identifying as Non-Hispanic Black, 11.5 percent identifying as Hispanic, and 6.6 percent as Asian. The average age that students entered college was 18.4, with ages ranging from 17 to 24.

To measure socioeconomic status, we use the socioeconomic status composite measure created by NELS, which combines information from the father's education level, mother's education level, father's occupation, mother's occupation and family income from the parent questionnaire data in NELS:88. In our sample, the average socioeconomic status (SES) composite is 0.08, meaning that the college-going students in our sample are relatively advantaged compared to the unweighted national average of −0.08 in NELS:88. For pre-college academic skills, we use the score percentile rank from the NELS pencil and paper test in reading and math that students took in twelfth grade in high school. On average, students in our sample of college-going students scored in the 60th percentile, meaning that students in our sample scored on average at the 60th percentile of the national distribution of high school seniors. The average high school grade point average (GPA) for our sample is 2.89. In our full study sample, about a

quarter of students (24.9 percent) planned to major in STEM. We also take into account the highest level of mathematics course taken in high school, creating a series of indicators for whether students' highest math course was Algebra I or similar (10 percent), geometry (13 percent), Algebra II (34 percent), Trigonometry (15 percent), pre-Calculus (16 percent), or Calculus (12 percent).

Looking at institution-level characteristics, we see that approximately 38 percent of the students in our sample entered a public two-year institution as their primary institution, while 18 percent entered a private not-for-profit four-year institution, and around 45 percent entered a public four-year institution. Approximately 15 percent of the entire sample had taken calculus and 1.6 percent of the entire sample (10.7 percent of calculus takers) had failed calculus. Regarding key outcomes, about less than half of the sample (41 percent) had earned a bachelor's degree in any field as of 2000, while 46 percent did not. About 13 percent of the sample received a bachelor's degree in a STEM field.

The second and third sets of columns of Table 1.1 provide the summary of covariates, outcome measures and independent variables among students who planned to major in STEM (n = 910) and those who did not plan to major in STEM (n = 2740), respectively. The group of students who planned to major in STEM is more evenly split by gender (49 percent men and 51 percent women) compared with the group of students who did not plan to major in STEM (47 percent men and 53 percent women). There are fewer White students (70 percent as compared with 76 percent), more Black students (10 percent as compared with 7 percent), fewer Hispanic students (11 percent compared with 12 percent), and more Asian students (9 percent as compared with 6 percent) in the group of students who planned to major in STEM. Students who planned to major in STEM demonstrate slightly higher levels of pre-college academic skills (scoring on

average at the 63rd percentile compared with the 60th percentile) and achievement (2.98 GPA compared with 2.86) than those who did not plan to major in STEM fields. A significantly larger proportion of students who planned to major in STEM had taken Calculus as their highest math course in high school (20 percent) compared to those who did not plan to major in STEM (10 percent) while a higher proportion of students who did not plan to major in STEM fields had taken up to Algebra II (36 percent compared with 29 percent).

The percentages of students who entered a public two-year, a private not-for-profit four-year, or a public four-year institution as their primary institution in each group were fairly similar to the full sample. Approximately 28 percent of students who planned to major in STEM took calculus in college compared to 11 percent of students who did not plan to major in STEM. Four percent of students who planned to major in STEM as high school seniors had failed calculus, while one percent of students who did not plan to major in a STEM field failed calculus. The percentages of students who earned a bachelor's degree in each group were fairly similar to the full sample. Approximately 28 percent of students who planned to major in STEM earned a bachelor's degree in a STEM field, while 8 percent of students who did not plan to major in STEM earned a STEM bachelor's degree.

Methods

Estimation Strategy

We use doubly robust inverse probability weighting (IPW) to examine the relationship between failing calculus and degree outcomes among calculus takers. In our observational data, we cannot randomly assign our treatment (e.g., calculus failure). As such, students who fail calculus are likely to be different from those who did not fail calculus (our "control" condition)

in both observable and unobservable ways. Table 1.2 provides descriptive results on students who take and fail calculus in the study sample. We see in Table 1.2 that there are both demographic and institutional differences between students who pass and students who fail calculus. Given these differences, we cannot estimate the effect of calculus failure on degree completion by simply comparing the estimates of degree completion likelihood among those who failed or students who passed calculus. To address this issue, we use IPW estimates to account for differences in the observable characteristics of students who pass and fail calculus.

IPW estimators use a two-step approach. First, the predicted probability of receiving the treatment is estimated for each student. Then, weights for each student are created. To balance the groups on observable characteristics, the IPW scheme up-weights students who received a given treatment but were unlikely to receive the treatment based on observable characteristics (e.g., students who were likely to fail but passed, or who were likely to pass but failed). Conversely, the scheme down-weights students who were highly likely to receive the treatment they received.

One limitation of IPW is that it assumes that the model used to predict the treatment (and therefore the weight) is correctly specified. If this model is not correctly specified, then the weighting will not account for the differences in these observable characteristics. We can relax the model specification assumption by using doubly robust IPW estimators and include controls in our weighted models predicting our outcomes. In these models, if either the weighting model or the final model is correctly specified, we will account for potential imbalance in our observable characteristics. It is important to clarify, however, that doubly robust models do not account for differences in unobserved characteristics of respondents. For a step-by-step process of how we created the doubly robust IPW estimators, see Appendix B.

Results

Predicting Calculus Taking and Performance

Table 1.2 presents the results of linear probability models in order to provide descriptive information on the characteristics of students who (a) take calculus compared to the entire study sample (n = 3490) and (b) fail calculus compared to students who had passed calculus (n = 540).

Model 1 shows that women are 11 percentage points less likely to take calculus than men, and that Asian students are nine percentage points more likely to take calculus than white students. A one-unit increase in SES composite is associated with a two percentage-point increase in taking calculus. One percent increases in students' reading and math scores and high school GPA are associated with four and 11 percentage-point increases in the likelihood of taking calculus, respectively. Compared to students who had algebra I or a similar course as their highest math class in high school, students who took algebra II are, if anything, slightly less likely to take calculus, while students who took calculus in high school were 31 percentage points more likely to take calculus in college. Students who planned to major in STEM as high school seniors were 13 percentage points more likely to take calculus. Finally, students entering a four-year private or public college (compared to entering a two-year college) were six and three percentage points more likely to take calculus, respectively.

Model 2 examines how the same set of factors from Model 1 are associated with failing calculus among students who took it. Importantly for our purposes, we see no gender differences in the likelihood of failing among calculus takers. We do find that high SES students, as well as students with higher GPAs in high school are less likely to fail. We also find that students who directly enter a four-year college are more likely to fail than students who first entered a two-year college. All other variables in the model yielded statistically non-significant findings.

General and STEM Bachelor Degree Attainment

Our results examining the relationship between failing calculus and degree attainment are presented in Table 1.3. As noted earlier, to focus on students who might plausibly be in the STEM pipeline, we restrict our analyses here to students who (a) planned to major in STEM in their senior year of high school and (b) had taken calculus in college. Students in this sample were weighted based on their probability of being assigned to treatment received. To address concerns around misspecification in the weighting model, we estimate doubly robust models that include all covariates in the models predicting our outcomes. In the first two models, we first examine whether students completed a bachelor's degree in any field. Models 3 and 4 examine whether students attained a bachelor's degree specifically in a STEM field.

In Model 1, we examine the relationship between failing calculus and completing a bachelor's degree. After accounting for demographic characteristics, prior achievement, academic skill, highest math course taken in high school, and institution-level covariates, we find that failing calculus is associated with a 12 percentage-point decrease in degree completion. In Model 2, we interact failing calculus and gender to see whether the relationship between failing calculus and bachelor degree completion varies by gender. To facilitate interpretation, we present predicted probabilities from Model 2 (holding covariates constant so that covariates are averages for the study sample) in Figure 1.1. While we find only small differences in the likelihood of receiving a bachelor's degree between men who passed and failed calculus (0.80 versus 0.76), we see that women who did not fail calculus are 32 percentage points more likely to receive a bachelor's degree than women who failed calculus (0.92 versus 0.60; $p = 0.019$). Men's likelihood of receiving a bachelor's degree is thus not strongly tied to whether they pass calculus,

while for women it is. Women who pass calculus are more likely to get a bachelor's degree than men, while women who fail calculus are less likely to do so.

Model 3 in Table 1.3 examines the relationship between failing calculus and STEM bachelor's degree completion. Here we find that, overall, failing calculus was not statistically significant ($p = 0.165$), though the point estimate is similar in magnitude and direction as in Model 1, suggesting that students who fail are less likely to obtain a STEM degree. Model 4 follows Model 2, examining the relationship between failing calculus and receiving a STEM bachelor's degree by gender. Predicted probabilities from Model 4 are reported in Figure 1.2. As above, we find no statistically significant differences among men (0.74 versus 0.86), but we do find that there is a statistically significant difference between women who do and do not fail (0.07 versus 0.78, $p < 0.001$). As is readily visible in Figure 1.2, failing calculus does not appear to weed out men, but does appear to weed women out.

Despite widespread interest in the role of weed-out classes in the STEM training pipeline, little is known about how failing a weed-out class might shape both men and women's STEM decisions to major in a STEM field. Using nationally representative data and a wide range of controls, we find that women who intended to major in STEM and fail calculus in college are significantly less likely to obtain a bachelor's degree in a STEM field. For men who intend to major in a STEM field, on the other hand, we find no evidence that failing calculus lowers their likelihood of obtaining a STEM degree. To the degree that calculus functions as a weed-out class, our findings suggest that it does so in a profoundly gendered way, weeding out women but not men.

Our results have important consequences for policies aimed at increasing the representation of women in STEM fields. Given that calculus often serves as a gatekeeper for advanced courses

in STEM, students who fail calculus face additional barriers that make it difficult to continue with their college studies in many STEM fields (Seymour and Hewitt 1997; Chen 2013). Our findings suggest that these barriers do little to dampen men's STEM degree completion, but may play a substantial role in shaping women's STEM degree completion. Policies aimed at increasing the representation of women obtaining STEM degrees may want to focus on women at this crucial stage, and efforts to assist students who have failed calculus may want to focus particularly on women. More broadly, given the lack of an effect on men's majors, these findings suggest that STEM educators may want to rethink the role of weed-out classes in STEM education. That is, it is difficult to argue that weed-out classes are doing their job and keeping unprepared individuals from pursuing these majors, when men who fail calculus are just as likely to graduate with a STEM degree as men who pass.

This lack of a difference for men is perhaps puzzling and raises additional questions. For example, it is unclear at what rate we would want men and women who failed calculus to continue pursuing STEM degrees (Penner and Willer 2015). Women are generally more responsive to grades than men (Charles and Bradley 2009), and while research on STEM persistence typically operates under the assumption that STEM persistence should be encouraged for all individuals, it seems plausible that after failing a weed-out class, pursuing a different major is potentially more adaptive than continuing to major in STEM. That is, while qualities like grit (Duckworth et al. 2007) and resilience (Masten 1994) are rightfully celebrated, adaptive goal disengagement (Heckhausen and Schulz 1995) is also an important adaptive strategy. To use a non-educational example, somebody who has repeatedly asked a romantic interest to go on a date and been turned down should potentially disengage from the goal of being in a romantic relationship with this individual, rather than continue to persist. While we are unable to

adjudicate whether the women who fail weed-out classes are best served by persisting in STEM fields, we argue that understanding the outcomes associated with weed-out class failure provides insight into the larger structural changes needed to alter students' persistence decisions.

In line with arguments around adaptive goal disengagement, our findings could in part also reflect the fact the women who fail calculus have better non-STEM options than men (Penner 2015; Wang et al. 2013). If this was the case, weed-out classes could plausibly explain both why women are less likely to major in STEM fields (they switch their majors after failing) and why men are less likely to graduate from college, net of enrollment rates (if they drop out after failing a weed-out class). As we only find evidence for the first of these processes, this suggests a gendered dimension in how calculus weeds women out of STEM fields. It also seems unlikely that these differences could produce differences of the magnitude we observe here. However, this perspective does highlight that we should not view women dropping out of the STEM pipeline as failures, but instead focus on questions around how STEM fields are structured.

In addition to questions about the larger structure of STEM education, larger societal stereotypes about gender and STEM are potentially relevant. One explanation for our findings is that the weed out culture for introductory-level coursework combines with gendered stereotypes about STEM fields to result in different self-assessments after calculus failure (Correll 2004). That is, much like the women in Correll's study who expressed less interest in pursuing fields that were said to be male advantaged, larger gender stereotypes might shape how women who fail calculus incorporate this information into their self-assessments and interests differently than men.

In supplemental analyses, we considered whether failure in any course deters women from earning a STEM degree. Taking a sample of students in the humanities "pipeline," we estimated

whether failing introductory writing composition is more likely to deter women than men from graduating with a humanities degree using the same IPW estimation strategy described above. While failing introductory writing is negatively associated with completing a bachelor's degree and a humanities bachelor's degree, we find no gender differences in humanities degree attainment rates among those who failed this course. We also examined other potential STEM weed-out courses (e.g., introductory chemistry), and do not find similar patterns in these courses as for calculus. This is perhaps surprising, and may speak to the unique space that calculus occupies.

Limitations

While we provide important evidence regarding the different ways in which women and men respond to failing weed-out courses, our study has several limitations. The first is the possibility that students who have failed calculus are different from students who did not in unobservable ways, limiting causal attributions. While we account for a wide range of observable characteristics by estimating doubly robust IPW, our approach cannot account for unobserved differences between the students who did and did not fail calculus.

Another limitation of our study is our lack of information about students' intended majors before and after taking calculus. We use information about whether students planned to major in STEM as high school seniors to indicate whether students could be in the STEM pipeline at this point, but cannot isolate failing calculus as being the factor that led students to pursue a different major. For example, we lack information on other important factors associated with college and STEM persistence, such as quality of faculty-student contact in the STEM department, peer interactions, experiences or perceptions of diversity on the college campus, student satisfaction, and participation in extracurricular activities while enrolled in college (Seymour and Hewitt

1997). Of particular note, we lack data on perceptions of failure, motivation, and self-efficacy in the NELS:88 (Tinto 1987). However, to the degree that many of these considerations could be mediators that helped explain why failing mattered, it is unclear that they should be introduced as control variables. Additionally, while we acknowledge that calculus takers across STEM majors may differ, the limited sample size in our study does not allow separating out analyses by specific major (e.g., physical versus biological sciences).

Finally, although we use a large, nationally representative dataset to examine these questions, the number of individuals who intended to major in a STEM field and took (and failed) calculus is relatively small, necessitating caution in interpreting the results. As such, these results would benefit from future replication studies. Furthermore, as noted above, in our supplemental analyses, we find evidence suggesting that calculus may be unique, as we do not find similar patterns for other introductory STEM courses. However, given the relatively small samples for these classes, future work on this question would be particularly useful in understanding if other attributes to its position in the course sequence, course content, pedagogy or other factors play a role in weeding out women but not men. In particular, while we focus on calculus, given its prominent position and relative prevalence, future work might fruitfully examine whether other weed-out classes function in similar gendered ways.

Conclusion

Gender disparities in postsecondary STEM education continue to be an enduring issue in higher education. Our study examined how men and women react differently to failing a weed-out course among potential STEM majors, which might shape their educational pathways. Using detailed individual-level data from NELS PETS:1988–2000, we find that women who planned to major in STEM and failed calculus in college were substantially less likely to obtain a bachelor's

degree in STEM. On the other hand, failing calculus did not appear to lower the likelihood of STEM degree receipt among men. Thus, we demonstrate evidence of the gendered ways these weed-out courses function—weeding out women but not men in the STEM degree pipeline.

CHAPTER 3: Failing at Remediation? College Remedial Course-taking, Failure, and Long-term Student Outcomes

Abstract

Colleges offer remedial coursework to help students enrolling in post-secondary education who are not adequately prepared to succeed in college-level courses. Despite the prevalence of remediation, previous research presents contradictory findings regarding its short- and long-term effects. We use a doubly robust inverse probability weighting strategy to examine whether the degree completion and wage outcomes associated with remedial education vary by passing or failing remedial coursework. Using the NLSY Postsecondary Transcript-1997 data, we find that almost 30 percent of remedial takers fail a remedial course. Students who took and passed their remedial coursework at both two-year and four-year colleges were more likely to graduate from college than similar students who did not take remediation. For both two-year and four-year college entrants, students who failed remedial coursework were less likely to obtain a bachelor's degree and, among degree receivers, took longer to graduate. Students who entered two-year or four-year colleges and failed remedial coursework earned lower wages over time compared to similar students who never took remediation. Among four-year college entrants, these wage differences seem to be explained completely by degree completion. However, wage differences for two-year college entrants still remain after accounting for degree receipt. Our findings suggest that while many students may benefit from remedial education, a substantial number of students struggle with remedial coursework and fail to realize the intended benefits.

Keywords: Remediation, higher education, degree completion, wages, inverse probability weighting

Introduction

Approximately two-thirds of all students entering two-year colleges and 40 percent of students entering four-year colleges enroll in some form of remedial coursework (Chen 2016). Remediation rates tend to be higher for delayed entrants and older returning students, as well as for Black and Hispanic students; remedial coursetaking is also generally higher at two-year, open-access institutions – institutions where many students begin higher education (Merisotis & Phipps 2000; Kurlaender & Howell 2012). While remediation rates are particularly high at community colleges and non-selective colleges and universities, even at selective four-year colleges and universities 30 percent of students take at least one remedial course (Chen 2016). Course failure is a common experience among students in remedial courses. In fact, less than half of community college students who enroll in remedial courses ultimately pass them all; among four-year college students in remedial courses that rate is just 59 percent (Chen 2016).

Considerable attention has focused on the enormous cost of remediation to public colleges and universities as well as to students themselves. The annual cost of remediation is estimated to be nearly $7 billion nationally to colleges (Scott-Clayton, Crosta, & Belfield 2014), with many arguing that taxpayers are "double billed" for colleges teaching academic skills that students should have learned in high school (Saxon & Boylan 2001). A recent report from the Center for American Progress indicates that students and their families paid an annual $1.3 billion in out-of-pocket costs for remediation across the nation (Jimenez, Sargard, Morales, & Thompson 2016). Moreover, although research on the effectiveness of remedial education is mixed, over the past 30 years remedial course takers earned degrees at lower rates than their non-remediated peers (Adelman 1999; Adelman 2004; Chen 2016). However, simply eliminating remedial education is problematic, as evinced by Florida's 2014 legislation that made remedial

coursework optional for students (Hu et al. 2016). In the fall semester following this policy change, fewer students enrolled in remedial coursework and the passing rates for gatekeeper courses declined, as students who would have been placed in remediation were likely underprepared for those courses. This suggests that there may always be a need for remediation in college, particularly for students who are the most unprepared for college-level coursework (Bailey, Hughes, & Jaggars 2012). With growing demands for a skilled workforce in the United States (Bailey, Jeong, & Cho 2010), it is crucial for researchers and policymakers to better understand the efficacy of remediation.

Previous research has suggested that students receiving remediation rarely have better academic and labor force outcomes than similar students who were not placed in remediation. However, when analysts focus on students who successfully complete remediation, they find much more positive outcomes (Attewell, et al. 2006; Bahr 2008; Bettinger & Long 2009; Chen 2016; Hodara and Xu 2016). In this paper, we explicitly examine how the outcomes associated with remediation vary based on whether students pass or fail their remedial coursework. Drawing upon new postsecondary transcript data linked to the National Longitudinal Survey of Youth (1997), we are able to overcome a key limitation in previous research by distinguishing between students who fail a course versus those who do not complete credits and withdraw or take an incomplete. Failure is an important signal in a student's academic trajectory (Crisp, Nora, & Taggart 2009). We hypothesize that failing remedial coursework may be a particularly large setback for incoming college students, both since it has consequences for student aid eligibility (Staying Eligible 2017) and since many students may interpret remedial course failure as an indication that they do not belong in higher education. As such, we argue that examining remedial course failure is crucial to understanding remediation, college retention, and long-term

student success.

Part of the difficulty in assessing the impact of remediation on student outcomes is that students who require remediation differ from those who do not, making it challenging to isolate the effect of remediation on college outcomes from the effects of remedial students' relatively weak academic preparation and other student characteristics associated with remediation. Some researchers have attempted to address these selection issues by employing randomized controlled trials to evaluate interventions intended to supplement or ameliorate issues with remediation policies. Examples of these evaluations include Barnett et al.'s (2012) study, which examined the outcomes of developmental summer bridge programs offered in Texas, and Logue, Watanabe, and Douglas' (2016) research, which evaluated the outcomes of students taking remedial mathematics simultaneously with introductory statistics. Another study reported findings from random assignment to "learning community" models designed to improve chances of college success (Visher, Weiss, Weissman, Rudd, & Wathington 2012). Overall, these studies find positive effects for the completion of college-level coursework across two-year community colleges, but they do not find a significant impact on persistence. These studies highlight the promise of remediation, as well as the challenges it faces in creating lasting change. In this paper we use a novel nationally representative dataset to complement existing research on the effects of specific remedial interventions, and underscore the important role of remedial course failure in whether students realize the intended benefits of remediation.

Literature Review

The Effects of Remediation on Educational Outcomes

A number of rigorous studies have focused on students who score near the cut-offs for remedial placement tests at two-year community colleges. Using a regression discontinuity (RD) design to compare students just above and below remedial placement cutoffs, Calcagno and Long (2008) find that remedial coursework promoted early persistence (into the second year), but did not affect eventual four-year degree completion in Florida, while Scott-Clayton and Rodriguez (2015) find no impact of remedial placement on enrollment, persistence, or eventual degree attainment in a large, urban community college system. Other researchers focus on remediated students in two-year colleges more broadly (and not just at the assignment threshold), finding that remediation lowers their odds of earning an associate's degree (Clotfelter, Ladd, Muschkin, & Vigdor, 2015) and their likelihood of successfully transferring to a four-year college (Crisp and Delgado, 2014). These negative effects are congruent with the Deil-Amen and Rosenbaum's (2002) argument that remedial placement "cools out", or lowers, students' educational aspirations, potentially in part by diverting remedial students away from earning college-level credits towards a degree (Scott-Clayton and Rodriguez, 2015; Clotfelter et al., 2015).

Importantly, however, there is some evidence of heterogeneity in the effects of remediation at two-year colleges. Melguizo et al. (2016), for example, find evidence of school-level heterogeneity using an RD design to examine the effects of remedial placement in six community colleges in a large community college system, showing a range of negative, positive, and null findings across the different schools. Other studies find evidence of student-level heterogeneity: Using an RD design to examine remediation in a single large community college, Ngo (2018) finds that while students who are placed in remediation due to a lack of fraction

knowledge subsequently fare worse than non-remediated students, remedial placement has no effects for those placed into remediation for other reasons.

While the effects of remediation at two-year community colleges have been studied extensively, there are fewer studies that examine remediation across two-year and four-year colleges. The findings from studies that do examine remediation in this context are largely mixed. Using an RD design, Bettinger and Long (2009) found that students placed in remedial math courses at nonselective four-year colleges were more likely to drop out or transfer to a two-year college. However, remediation did not lower the likelihood of obtaining a bachelor's degree. Moreover, students who had completed their remedial math courses were more likely to obtain a bachelor's degree, albeit taking more time to complete their degree, than students who attempted but did not complete their remedial math coursework. Boatman and Long (2018) also use an RD design to show that Tennessee students with lower levels of academic preparation actually benefit from taking remedial coursework, while those who only need a single remedial course do worse when placed in remediation. While Boatman and Long's (2018) findings are primarily driven by two-year college students, they find a similar pattern among students at four-years schools. From these findings, we might expect that students who started at two-year colleges and take remedial courses are less likely to complete a bachelor's degree than those who began at a four-year college, and remediation may add time to degree for two-year college entrants (see also Reynolds 2012; Scott-Clayton & Rodriguez 2015; Xu, Jaggars, Fletcher, & Fink 2018). On the other hand, there may be no differences across postsecondary institutions. Martorell and McFarlin (2011) directly examined the impact of remediation between two-year and four-year colleges, finding that remediation did not decrease the probability of receiving a bachelor's degree or increase time to degree at either two-year or four-year colleges.

Previous Research on Wage Outcomes

To date, little research has examined how remedial coursework affects wages, even though educational attainment more broadly is closely linked to social and economic advantages. Hout (2012) for example, found that annual earnings increased roughly 20 percent for each year of educational attainment. Moreover, the least educated workers were almost four times more likely than college graduates to be unemployed during the recession and stayed unemployed longer than college graduates. To the degree that schooling provides students with important skills, we might expect remedial education to provide students human capital (such as improved literacy) that is rewarded in the labor market (Johnson 2007). We might also expect that remedial coursework helps students succeed in the labor market by helping them succeed in college (though there is a dearth of evidence suggesting strong positive effects of remediation on college completion).

In one of the few studies to directly examine how remediation impacts labor market earnings, Martorell and McFarlin (2011) used a regression discontinuity approach for the Texas Academic Skills Program (TASP) test. Martorell and McFarlin do not find that students' post-college earnings benefit from remediation in general, but rather find negative, albeit small, effects on 6-year earnings in four-year schools with low-remediation rates. On the other hand, Hodara and Xu (2016) find that students attending community college in North Carolina and Virginia have increased earnings after earning remedial English credits, which appears to be entirely driven by an increased probability of employment. However, they also find that students assigned to the lowest level of remedial math earned less over time (seven years after college entry). This study suggests that the time to take multiple remedial math courses may mean forgone earnings and that perhaps skills acquired in remedial math courses are not valued as

much in the labor market as skills acquired in remedial English. More generally, recent research suggests that students who entered a two-year college and transferred to a four-year college may have lower earnings over time (Xu et al. 2018), even though two-year transfer students were just as likely as four-year college entrants to obtain a bachelor's degree.

The Current Study

Our study contributes to this literature in three important ways. First and foremost, we attend to the important role of course failure, separately comparing students who pass and fail remediation to their non-remediated peers. Second, while regression discontinuity designs provide rigorous quasi-experimental evidence regarding the effects for students who were near the cut-off, they cannot speak to the effects for students who were not near the cut-off score, or for students in institutions with other assignment mechanisms for remedial coursework (e.g., where advising is an integral aspect of remedial course assignment). Our study includes students who took any remedial course and thus provides descriptive evidence regarding the schooling and labor market outcomes associated with remediation across a wider range of students. Third, our study examines remediation separately for two-year and four-year colleges across the country and thus offers insights on how the outcomes associated with remediation might vary by institution type (two-year or four-year colleges).

Our key research questions examine whether the degree completion and wage outcomes associated with remedial education vary by whether students pass or fail remedial courses. To motivate the analyses for our central research questions, we first ask (1) Who takes and who fails remedial coursework? Then, we address our key research questions: (2) What are the schooling outcomes associated with remedial coursetaking and how do they vary by whether students pass

or fail their remedial coursework? and (3) What are the wage outcomes associated with remedial coursework and how do they vary by whether students pass or fail their remedial coursework ? To understand how remedial courses may affect students across postsecondary educational institutions, we examine students who first entered a two-year college separately from those who first entered a four-year college. Given the important differences between two- and four-year colleges, as well as the potential relationships between the academic and wage outcomes we are interested in, we first examine the academic and wage outcomes for two-year college entrants, and then do the same for four-year college entrants.

Data

We use data from the National Longitudinal Survey of Youth 1997 (NLSY97), a nationally representative sample of approximately 9,000 youth born between the years 1980 and 1984. In 2011, when NSLY97 respondents were between the ages of 31 and 35, the study undertook a retrospective effort to collect complete postsecondary transcripts from all respondents who reported attending a postsecondary undergraduate degree program during any of the NLSY97 interviews (rounds 1 through 15). At least one transcript was received for 3,818 of the 4,399 youth for whom one or more transcripts were requested. Our sample consists of the participants who have attended a post-secondary institution from 1997-2011, submitted valid post-secondary education transcripts, and who had work history information (n = 3,646). The postsecondary transcript data provides important chronological information about students' enrollment patterns across two-year and four-year colleges, courses taken, and academic performance in these courses, including their bachelor's degree attainment and time to degree completion.

In addition to data obtained from the NLSY Postsecondary Transcript Study 2011, the NLSY97 survey provides detailed information on employment history, such as work experiences, income, assets, and other economic characteristics. The NLSY97 work history data provide an annual work record of each respondent from January 1994 through 2011, and contain information on the respondent's labor force status each year, the usual hours worked per week at all jobs, and earnings for all jobs. These data enable us to link detailed information about two- and four-year college students' coursetaking patterns with their post-college wages.

Variables

We use NLSY97 transcript data to chart students' remedial coursetaking experience. These data flag courses as remedial based on the 2010 College Course Map (CCM) taxonomy system.[1] We consider any student who took one or more remedial courses in college as a student who was exposed to remedial instruction. We use the NLSY97 transcript course grade to identify students who failed remedial courses, coding students who received a "0" or "F" for the particular course and reported zero earned credits as failing the course.

After describing the correlates of remedial experiences, we explore the relation between exposure to remediation, as well as failure in remedial courses, and two sets of student outcomes. First, we examine whether a student graduated with a bachelor's degree, and for those who did, we also examine whether they finished in six years or after six years. These data are available for all NSLY97 respondents for whom transcript data are available. Second, we explore student labor force outcomes. These analyses use post-college wages reported in the years 2007 through

[1] While prior studies have examined remedial course by subject matter (e.g. English versus math), our paper focuses on the two-year and four-year contrast. Given issues of statistical power, this precludes separating our sample by course subject.

2011, the latest five years in the dataset.[2] Wages were averaged over the number of years of wage data that are available. If the student did not report their wages in any year between 2007 through 2011, the wages were considered missing and were excluded from the estimation of a student's average wages over time. The average hourly wages were then converted to logged wages for the analyses.

All multivariate analyses control for a wide array of student demographic, socioeconomic, and academic background characteristics. Gender was a dummy variable, coded 1 for male and 0 for female. The racial categories we use are White, Black, Hispanic, Asian, and Other (which comprised approximately 3 percent of the entire sample). Birth cohort is an indicator for the year that the student was born, which in this sample ranges from 1980 to 1984. Age at college entry is measured in years by subtracting birth date from the student's first term year. To account for family background, we also include biological mother's age at first birth (in years), years of education of the respondent's most educated parent, household income and logged per capita household income from when students were in high school, which is the earliest reported year for household income in the survey. We used the student's test score percentile rank in the Armed Services and Vocational Aptitude Battery (ASVAB) test and high school grade point average (GPA) to measure students' pre-college academic background. ASVAB measures the respondent's knowledge and skills in topics such as mathematics knowledge, paragraph comprehension, and general science; most NLSY97 Round 1 respondents

[2] We use averages from five years of post-college wage data where possible, but in cases where this is not possible, we use averages of three or four years of data. For example, if a student's last term was in 2006 or earlier, we use the five years from 2007 through 2011 to calculate their average hourly wages. If a student's last term reported year was in 2007, we use the four years from 2008 through 2011. For students whose last term was in 2008, we use the three years from 2009 through 2011. For these analyses, we do not use respondents with fewer than three years of earnings, which means that our wage analyses exclude students whose last reported college term was in 2009 or later. Supplemental analyses that include zero wages show similar findings.

took the ASVAB test between 1997 and 1998. As an overwhelming majority of students (92 percent) worked at some point while attending college, we include a logged continuous measure of average hours worked per week. Finally, given the difference between two- and four-year colleges, we conduct separate analyses for students who enrolled in a two- or four-year college as their primary institution after completing high school.

Missing Data

The percentage of missing values ranged from zero for some demographic variables, such as race, gender and birth year, to as high as 25 percent for household income. Only 45 percent (1,638) of the 3,646 students in the sample would have been available for analysis using listwise deletion. Data are primarily missing due to one of three reasons: the respondent did not participate at all in the survey year; the respondent did not provide a valid answer to the question; or was a valid skip (e.g. a question only applied to respondents in a certain age range). We address the issue of missing data using multiple imputation (Rubin 1987). Because of the complex design employed in the data collection, the primary sampling unit, strata and weights were included in the imputation process (Allison 2003). Multiple imputation produces consistent estimates if the data are missing at random conditional on the variables in the model, and the imputation model is correctly specified.

Analytic strategy

Our initial analyses use these data to understand who takes remedial coursework, and how students who pass and fail their remedial coursework differ from their non-remedial course taking peers. We begin by comparing the demographic characteristics, socioeconomic

background, and academic histories of students who take remediation with students who do not. We then separate students who passed and failed remedial coursework, first comparing students who passed their remedial coursework with their peers who did not take remedial coursework, and then comparing students who failed remedial coursework with their non-remediated peers. We estimate a series of models of the following general form:

$$Y_i = \alpha_i + \beta_k \boldsymbol{X}_{ki} + \varepsilon_i \quad (1)$$

where Y_i represents respectively whether student i took remedial coursework, whether a student passed their remedial coursework (omitting students who failed remedial coursework), and whether a student failed a remedial course (omitting students who passed their remedial courses). We predict these outcomes as a function of a range of student characteristics, including gender, race, birth cohort, age at college entry, family background, and pre-collegiate academic background, represented as $\boldsymbol{X}_{ki}$.

We then draw upon the results of these analyses to examine the relationship between remedial coursetaking and degree and wage outcomes using a doubly robust inverse probability weighting (IPW) strategy. In our observational data, we cannot randomly assign our treatment (i.e., remedial coursetaking and remedial course failure) to ensure that the treatment is independent of the outcome (e.g., degree completion). Thus, students who take remediation (and those who fail remedial coursework) are likely to be different from those who did not take any remediation (our "control" condition) during their college career in both observable and unobservable ways. Given these differences, we cannot estimate the effect of remedial coursetaking and failure on degree completion by simply comparing the point estimates for degree completion rates among those who did and did not take (or fail) remedial coursework. To

create a more plausible counterfactual, we use a doubly robust IPW approach to account for differences in the observable characteristics of students who pass and fail remediation.

To execute this strategy, we use the covariates included in Model (1) to estimate the predicted probability that each student takes remedial coursework, and their predicted probability of passing or failing their remedial coursework. We then weight each student by the inverse of the predicted probability relevant for the comparison being made (e.g., those who took and failed remedial coursework versus those who did not take remedial coursework). To balance the groups on observable characteristics, the IPW scheme up-weights students who received a treatment condition they were unlikely to receive based on observable characteristics (e.g., students who were likely to take and fail remedial coursework but did not take remedial coursework, or who were likely to be non-remediated but took and failed remedial coursework). Conversely, the approach down-weights students who were highly likely to receive the treatment they received.

A limitation of IPW is that it assumes that the model used to predict the treatment (and therefore the weights) is correctly specified. If this model is not correctly specified, then the weighting will not account for the differences in these observable characteristics. We can relax the model specification assumption by using doubly robust IPW estimators, and including controls in our final weighted models predicting our outcomes. In these models, if either the weighting model or the final model is correctly specified, we will account for potential imbalance in our observable characteristics. It is important to clarify, however, that doubly robust models do not account for differences in the unobserved characteristics of respondents. For more details on how we calculate the doubly robust IPW estimators, see Appendix C.

Results

The first and fifth columns of Table 2.1 provide information about our controls and outcome measures separately for students who first entered a two-year college (n = 1,677) and those who entered a four-year college (n = 1,884), respectively.[3] Students who entered two-year colleges are more likely to be male (47 percent compared with 45 percent of students who entered a four-year college); they are also less likely to be White (48 percent compared with 61 percent of students entering four-year colleges), more likely to be Black (25 percent compared with 21 percent), more likely to be Hispanic (22 percent compared with 14 percent), and less likely to be Asian (2 percent compared with 3 percent). On average, students entered a two-year college a year older (19.8 years) than four-year college students (18.7 years).

To measure family socioeconomic status, we use the mother's age at first birth, years of parental education (from the most educated parent), household income, and income per capita in the household. We see that the average age of mothers at first birth is 22.9 years among two-year college entrants (compared with 24.5 years among four-year college students). On average, a two-year student's most educated parent had completed 13.2 years of education while a four-year student's most educated parent completed 14.7 years of education. The average household income reported in 1997 is $44,650 for two-year students and $62,964 for four-year students. In our analyses, we use per capita household income, which had an average of $10,670 for two-year students and $15,582 for four-year students.

We use the student's ASVAB test score percentile and high school GPA to measure pre-college academic background. Students who entered a four-year college scored in the 65th

[3] Note that the subgroups for students who enter a two-year or four-year college does not add up to the total sample of students in the study (*n* = 3,646). Table 2.1 provides descriptive statistics for the study sample prior to multiple imputation, and students who were missing on information for primary institution type were not included in the descriptive statistics but are included in the analytic models after multiple imputation.

percentile on the ASVAB test, meaning that these students scored on average at the 65th percentile of the national distribution of young adult test takers. By contrast, students who entered a two-year college scored at the 45th percentile on the ASVAB test. Four-year college entrants had a higher GPA than those who entered a two-year college (3.18 GPA compared with 2.76). Both two- and four-year college students were overwhelmingly employed while in school. Employed students worked an average of 27.7 hours per week at a two-year college and 24.6 hours per week at a four-year college. Remediation rates were slightly higher in two-year colleges: Approximately 65 percent of students who entered a two-year college took a remedial course compared to 59 percent of students who entered a four-year college. About 22 percent of students who entered a two-year college failed a remedial course, while 14 percent of students who entered a four-year college failed a remedial course.

Regarding our key outcomes, 75 percent of two-year college entrants had not earned a degree, while 12 percent earned an associate's degree as their highest degree, and 13 percent earned a bachelor's degree. Among students who entered a four-year college, only 35 percent received no degree, while 6 percent earned an associate's degree as their highest degree and over half of students received a bachelor's degree (59 percent). Among those who earned a bachelor's degree, 84 percent of students who entered a four-year college earned their degree within six years while 63 percent of students who entered a two-year college did so. Students who entered a four-year college had higher wages than students who started at two-year colleges ($18.47 versus $15.41), even among those with bachelor's degrees ($19.75 compared with $18.76).

[Insert Table 2.1]

The second, third, and fourth sets of columns in Table 2.1 provide descriptive statistics separately for three groups of students who first entered a two-year college: those who did not

take remedial coursework (column 2; n = 580), those who took and passed remedial coursework (column 3; n = 732), and those who took and failed remedial coursework (column 4; n = 365). We see that female and White students are overrepresented among students who passed remediation, and that Black students are overrepresented among students who failed remediation. Students who passed remedial coursework demonstrate higher levels of pre-college academic skills (scoring on average at the 50th percentile on the ASVAB test) and achievement (2.88 GPA) than students who did not take remedial coursework, while students who failed remediation on average had lower levels of pre-college academic skills.

By contrast, among students who began at four-year colleges (columns 6-8), we find that students who did not take remedial coursework had the strongest academic backgrounds, followed by those who took and passed their remedial coursework. Socioeconomic advantage follows a similar pattern of advantage, where students who did not take remedial coursework are the most advantaged, and those who fail remediation are the least advantaged. Like students who began at two-year colleges, Black students are overrepresented among students who failed remedial coursework at four-year colleges, while White students are underrepresented among students who took remedial coursework, and particularly among students who failed remedial coursework.

Who Takes, Passes, and Fails Remedial Courses

Table 2.2 presents descriptive information about the characteristics of students who entered a two-year college and (a) took a remedial course while enrolled in college (compared to students who never took a remedial course), (b) passed their remedial coursework (compared to students who never took a remedial course) and (c) failed a remedial course (compared to students who never took a remedial course). Model 1 shows that there are relatively few

significant predictors of taking remediation. Asian students are 17 percentage points more likely to take remediation than White students, and the likelihood of taking a remedial course increases by one percentage point for each additional year of education obtained by the respondent's parent. Interestingly, there were no differences in pre-college academic ability and achievement between students who did and did not take remediation.

[Insert Table 2.2 here]

Model 2 compares those who passed remedial coursework to the sample of non-remedial takers at a two-year college (n = 1,354), and Model 3 compares those who failed remedial coursework to those who had never taken remedial classes (n = 981). We see that women are 6 percentage points more likely to take and pass remedial coursework (Model 2), while Black, Hispanic, and Asian students are more likely to fail a remedial course than White students (Model 3). Parental education similarly predicts the likelihood of passing and failing remedial coursework (though for failing this coefficient is only marginally significant). By contrast, Model 2 indicates that students who pass their remedial coursework have a similar academic background as those who take no remedial coursework, while Model 3 shows that students who fail their remedial coursework had lower GPAs in high school and ASVAB test scores than those who did not take remedial coursework.

Table 2.3 presents results from analyses that parallel those in Table 2.2, but for students who entered a four-year college. We see that Black students are 10 percentage points more likely to take a remedial course compared with White students (Model 1), and that this difference is particularly pronounced when we compare non-remedial students to those who failed remediation. We also find that students who have higher GPAs and ASVAB scores are less likely to take remedial coursework. Although ASVAB scores predict both passing and failing

remedial coursework similarly, the difference in GPA between those who pass their remedial coursework and those who did not take remedial coursework is not statistically significant (Model 2), indicating that the difference observed in Model 1 is driven by differences between students who did not take remedial coursework and those who took and failed remedial coursework (Model 3). Our measures of socio-economic background do not significantly predict remedial coursetaking at four-year schools.

[insert Table 2.3 here]

Degree and Wage Outcomes among Two-Year College Students

Our results examining how remedial coursetaking and failure are related to degree attainment and wage outcomes among students who started at a two-year college are presented in Table 2.4. The specific outcomes we examine are bachelor's degree receipt, receiving a bachelor's degree after six years (among degree recipients), and post-college wages. For all three outcomes, we compare: those who had taken remedial coursework with those who had never taken a remedial course (Model 1; Remediation vs no remediation); those who had passed a remedial course with those who had never taken a remedial course (Model 2; Passed remediation vs no remediation); and finally, those who failed a remedial course with those who had never taken a remedial course (Model 3; Failed remediation vs no remediation). [4]

[Insert Table 2.4 here]

In Panel A, we examine the likelihood of receiving a bachelor's degree for students who entered a two-year college. We find that taking remediation is associated with a nearly 9

[4] Appendix Table C2 provides information about a fourth comparison that is potentially of interest: those who pass and fail remediation. In addition to the other covariates in Tables 2.4 and 2.5, Table C2 also controls for the number of remedial courses taken to ensure that these differences are not driven by differences in the number of remedial courses taken.

percentage-point increase in bachelor's degree completion for two-year college students after accounting for demographic, familial, and academic background characteristics. In Models 2 and 3, we see that both students who passed their remedial coursework and those who failed are more likely to graduate than observationally similar students who did not take remediation, although this difference is more pronounced when looking at students who passed remediation. Our results thus suggest that taking and passing remedial coursework increases the likelihood of completing a bachelor's degree among students who started at two-year colleges. Moreover, while students who fail remediation do not benefit from remedial coursework to the same degree as students who pass, we find that among students who started at a two-year college, even those who fail remedial coursework receive bachelor's degrees at higher rates than observationally similar students who do not take remedial coursework.[5]

In Panel B we examine whether students who initially entered a two-year college and completed a bachelor's degree did so within six years, which is the national average time to complete a bachelor's degree (Kena et. al. 2015). We find no evidence that students who passed their remedial coursework took more time to complete their degree (among bachelor's degree completers). These results imply that remediation has little effect on increasing the time needed to complete a degree for those starting at a two-year college if students pass their remedial coursework. One explanation is that entering a two-year college already increases a student's time to degree if the student transfers up to a four-year school (Deil-Amen and Rosenbaum 2002), so that any additional time needed to complete remedial coursework and then earn a degree is not significant for students who are already on a pathway that takes longer to do so.

[5] Supplemental analyses examined remedial coursetaking and failure to predict whether two-year college entrants earned an associate's degree to ensure that the patterns are similar to those we report for bachelor's degree completion. While the magnitude of the differences is somewhat smaller, the overall pattern is the same.

However, Model 3 suggests that students who fail a remedial course and eventually attain a bachelor's degree are substantially more likely to take over six years to do so, although this difference is only marginally significant.

Panels C and D report results from our wage analyses. Panel C shows that students who do and do not take remedial courses have similar post-college wages (Model 1). Importantly, however, we see a different pattern of results among students who pass and fail their remedial coursework. Students who pass their remedial coursework earn similar, and if anything, slightly higher wages than their non-remedial coursetaking peers (Model 2), while students who fail remediation earn substantially less than students who do not take remedial courses (Model 3). Panel D examines the degree to which these differences in wages remain once we account for differences in associate's and bachelor's degree receipt.[6] We find a largely similar pattern of results as in Panel C, suggesting that even after accounting for associate's and bachelor's degree receipt, students who take and fail remediation at two-year colleges earn less than observationally similar students who do not take remedial courses.

Degree and Wage Outcomes Among Four-Year College Students

Table 2.5 parallels Table 2.4, but presents results from analyses examining the degree and wage outcomes associated with remediation among students who initially entered four-year colleges. In Panel A, we examine the likelihood of receiving a bachelor's degree among the three comparison groups for students who entered a four-year college. We find that students who take remedial coursework on average do not differ from observationally similar students in their probability of completing a bachelor's degree (Model 1). However, this average null effect

[6] In addition to the covariates for the Panels A-C, all three models in Panel D also control for both associate's and bachelor's degree completion.

masks very different outcomes for students who pass and fail remedial coursework. Model 2 shows that students passing remediation are more likely (8 percentage points) to obtain a bachelor's degree than those who did not take remediation. By contrast, Model 3 shows that four-year college students who fail remedial coursework are substantially less likely (25 percentage points) to receive a bachelor's degree than those who did not take remediation. Panel B examines whether students who take remedial coursework at four-year colleges and receive a bachelor's degree take longer to receive their degree, showing that only students who fail remedial coursework experience an increase in the likelihood that they will take over six years to finish their degree.

[Insert Table 2.5 here]

In Panels C and D, we examine the relationship between remedial coursetaking and average hourly post-college wages among respondents who attended four-year colleges. Similar to two-year college entrants (who do worse if they fail), among four-year college entrants, we find that those who fail remedial coursework earn substantially less (approximately 9 percent) than those who did not take remedial coursework, although this finding is only marginally significant. Panel D of Table 2.5 shows that the wage differences that we observe in Panel C appear to be largely driven by changes in the probability of receiving a degree, as once degree receipt is introduced as a control, we see that the wage differences are reduced and are no longer statistically significant. The results in Table 2.5 thus suggest that remedial coursework among students who start at four-year schools may help them complete their bachelor's degree if they pass, but that if they fail they are much less likely to complete their degree and their average earnings are lower as a result. However, unlike students who start at two-year schools, where the earnings loss associated with failing a remedial course remains after taking degree receipt into

account, for students who start at four year schools the earnings difference we observe appears to be a function of degree receipt.

Discussion

Despite extensive research on the impact of remediation in college, little is known about whether the outcomes associated with remediation differ by whether students pass or fail their remedial coursework. Our first set of analyses show that different characteristics predict whether students take and fail remedial coursework at two-year and four-year colleges. We find that although some factors predict remedial coursetaking and failure at both two-year and four-year colleges, others do not. Gender, for example, predicts taking and failing remedial coursework at four-year but not two-year schools, while Black students are more likely than White students to take and fail remedial coursework at both two- and four-year colleges. The characteristics of students who take remedial coursework are largely consistent with prior research using nationally representative data (Attewell et. al. 2006; Chen 2016), however we find that academic background measured through high school GPA and ASVAB scores predicted failing remedial coursework at both two- and four-year schools.

Like prior nationally representative research on remediation, this study cannot account for potential differences in the unobserved characteristics of students who do and do not take (and fail) remedial coursework. As such, we view our results as descriptively instructive, rather than causal. We also note that our R-squared values in Tables 2.2 and 2.3 are relatively low, suggesting that there is still much to be learned about who takes and fails remedial courses. As our doubly robust IPW results from Tables 2.4 and 2.5 highlight the substantively important and statistically significant differences in outcomes based on who takes and fails remedial

coursework, we believe that understanding the factors predicting remedial course failure will be an important undertaking for future research. In particular, we believe that it will be fruitful to examine information about whether students are placed into remediation through a college-entry test or were advised to take remediation, as well as other factors found to be associated with college completion, such as participation in precollege encouragement programs, quality of faculty-student contact, peer interactions, experiences or perceptions of diversity on the college campus, student satisfaction, perceptions of failure, motivation or self-efficacy, and participation in extracurricular activities while enrolled in college (Tinto 1993).

In spite of its limitations, this study provides important insights for understanding the outcomes associated with remediation. Prior research suggests that remediation has little to no benefits for students who begin their studies at two-year colleges (Calcagno and Long, 2008; Clotfelter et al., 2015; Crisp and Delgado, 2014; Deil-Amen & Rosenbaum, 2006; Scott-Clayton & Rodriguez, 2015). However, using nationally representative data and a wide range of controls, we find that taking remediation is positively associated with bachelor degree completion among two-year college entrants, regardless of whether they pass or fail. One possible explanation is that remedial course enrollment in our study may be capturing two-year college student's persistence in remedial course sequences and intention to transfer to a four-year school, given that only one third of students assigned to remediation enroll in a remedial course at two-year colleges (Bailey, Jeong, & Cho 2010).

Our study calls attention to the important role of remedial course failure in understanding the outcomes associated with remediation. We find that students who pass remedial coursework generally do no worse than similar students who did not take remedial coursework, while students who fail remediation take longer to graduate and earn lower wages than similar non-

remediated students. We also find notable differences in the outcomes associated with remedial coursetaking and failure among students who start at two- and four-year colleges, as among four-year college entrants failing remediation is also associated with a lower likelihood of degree receipt. Interestingly, students entering two-year colleges who fail remedial coursework appear to still be more likely to earn a bachelor's degree than their counterparts who do not take remedial coursework. Thus, in contrast to Martorell and McFarlin's (2011) findings, our results show disparate patterns across students who pass and fail remediation at two- and four-year colleges.

While we cannot speak to the mechanisms as to why failing remediation has different consequences at two-year and four-year schools, our findings might indicate that two-year colleges are better at remediation for degree completion than four-year colleges. We might also consider remediation (and remedial course failure) as a potential site for sorting students within higher education, as posited by Scott-Clayton and Rodriguez (2012) and Bettinger and Long (2009). From this perspective, failing a remedial course at a four-year college might signal to students that they do not belong in higher education. In this way, remedial course failure at a four-year college might divert students away from the college pathway altogether. By contrast, as remediation is more common among students entering two-year colleges, students could understand remedial coursework differently in this context. Given that Xu et al. (2018) find that students who begin at a two-year school have higher rates of bachelor degree attainment than students who enter a four-year school, we might speculate that remedial coursetaking among students entering two-year schools may indicate students' willingness to make longer-term investments in their education, and their intention to transfer to a four-year school. While students who enter two-year colleges and fail remedial coursework do not receive degrees at the

same rate as those who pass remedial courses, they nonetheless appear to receive some benefit from their remedial experience. However, these differences in how remediation works for students starting at two- and four-year schools may also be due to selection and differences with the non-remediated student body.

We also find differences in the relationship between post-college wages and remedial course failure at two- and four-year colleges. At both two- and four-year colleges, students who fail remedial coursework do worse compared to those who were not remediated. For students who entered a two-year college, the wage differences for failing remediation remain significant even after accounting for associate's and bachelor's degree completion, while the wage differences for those who fail remedial coursework among four-year college students appear to be completely explained by differences in associate's or bachelor's degree receipt. Although we are unable to speak to the precise mechanisms driving these results, it would appear that the negative repercussions of remedial course failure for the wages of four-year college students are largely a function of degree completion. Among two-year college students who failed a remedial course, however, remedial failure appears to matter for wages after accounting for degree receipt. This might indicate that remedial course failure has larger impacts on students' confidence that remain even after we account for degree completion, though it is not immediately clear why this would be the case particularly at two-year schools, especially given our results for degree completion.

One explanation could be that students starting at two-year colleges who failed remediation have a delayed entry into the labor market and are not able to catch up in earnings over time. Similar to findings from Jaggars and Xu (2016) and Hodara and Xu (2016), these students are not able to receive the same positive increases in earnings growth over time

compared to those who pass remediation, even after accounting for bachelor's degree receipt. Failing remedial coursework may be a particularly large setback in delaying students' entry into the workforce, not only in the loss of credits in the transfer process, but also in additional time needed for students to navigate the remedial coursetaking process after their transfer. Regardless of the precise mechanisms, these post-college wage differences are significant, as they are likely to grow over time as cumulative advantage processes widen these disparities in earnings.

The results of this study have implications for potentially identifying and meeting the different needs of students at two-year and four-year colleges. Policymakers should first be aware that two-year and four-year colleges structure remediation differently and that a one-size-fits-all approach would not be appropriate in these contexts. Four-year colleges would appear to benefit from emulating the support structures for remediation (e.g. advising, tutoring, and curriculum) at two-year colleges that help students eventually obtain a bachelor's degree, given that students who pass their remedial coursework appear to benefit from a higher likelihood of obtaining a degree than their non-remediated counterparts. However, two-year schools also need to ensure that students at these institutions take and pass remediation, given the lasting wage penalties for failing remedial coursework among students entering two-year colleges. To the degree that the issues faced by students entering two-year colleges are related to delays, co-requisite programs are a potentially promising solution, as they combine the enrollment of a transfer-level course in English and mathematics course with a support course for students underprepared for college-level coursework. While these programs are relatively new, early research indicates promising results, with higher course completion rates and improved academic performance (Logue, Watanabe, & Douglas, 2016). Other efforts to help struggling remedial students at two-year colleges should accelerate the degree attainment process so that their wages

do not suffer relative to their non-remediated peers and those who passed remediation should also be given consideration.

In sum, this study makes several contributions to the existing literature on the effectiveness of remediation. We highlight that remedial course failure is relatively common at both two- and four-year colleges and that those who fail remedial coursework do not appear to benefit to the degree that their counterparts who pass do, and have lower post-college wages than their non-remediated peers. In doing so, we not only highlight the important differences in the outcomes associated with remediation at two- and four-year schools, but also call attention to the important role that failure plays in shaping students' experiences. Future studies seeking to understand the effectiveness of remediation should attend to the important role that failure might play in determining the effectiveness of remediation, and how the effects of failing remedial coursework might be mitigated**.**

CHAPTER 4: What's in a Label? The Long-Term Effects of Student Labels

Abstract

Students receive labels on their performance on statewide tests, but there is a dearth of research on how this information impacts future academic achievement and behavioral outcomes. Employing a regression discontinuity approach, we find that students with a more negative label have lower test scores over time and increased school absences five years later. We also find that the likelihood of suspension increases in the following year, but not over time. By using a regression discontinuity design to address concerns around selection, our analyses estimate plausibly causal effects, and suggest that a negative performance label has long-term negative effects.

Introduction

Over the past few decades, the amount of information that students receive, particularly from official state standardized tests, has grown tremendously. Students, teachers, and schools are under increased pressure to improve results on these tests to meet state and federal mandated standards. The performance information embedded in these tests is ubiquitous and, perhaps, particularly salient. Standardized testing will likely continue in an era of accountability-based reform, given that the passage of Every Student Succeeds Act (2015) still maintains an accountability system across states. While some test scores are attached to advancing to the next grade, receiving a high school diploma, or entering particular types of colleges (Hanushek and Raymond 2005; Alon and Tienda 2007), other test scores did not have officially defined consequences for students. Given that students receive an indicator of their academic ability at primary and secondary school levels, it is important to understand the implications of labels, particularly when these labels that should not matter for students do have an impact.

One central feature of test-based accountability in the United States is that students receive both their test score and a label based on their test performance. Oregon, the state we examine in this paper, assigned the following labels to students: *Very Low, Low, Nearly Meet Standard, Meet the Standard,* or *Exceed the Standard*. The Oregon Department of Education created performance levels in their statewide assessment for each grade, subject, and year to evaluate students' progress toward master of the academic content. These labels have been divided into categories dependent on a cut-point. In this study, we use a regression-discontinuity design to examine the impact of test performance labels on students' future academic and behavioral outcomes. These labels do not provide any additional information beyond the student's test score, but rather serves as a way to more easily interpret the score. Educators do

not use the student's performance in the statewide assessment in their decisions for student placement or assigning a final grade. Thus, these tests do not have official state-defined consequences for the students. In this paper, we focus on the feedback that students received about their test performance. We show that, even without state-defined official consequences for students, receiving a less positive performance label based on the student's test score negatively impact students' academic and behavioral short and long-term outcomes.

Background

The No Child Left Behind Act of 2001 required each state to establish its own academic standards in the core content subjects of English language arts (ELA), mathematics, and science. NCLB also mandated that each state needed to develop an annual testing program to assess student progress toward mastery of these subjects and to define what proficient mastery of those standards meant. This legislation adopted the goal that all American public school students should be at "proficient" by grade level on state tests by the 2013-2014 school year. Under NCLB, states must demonstrate that schools are making "adequate yearly progress" (AYP) in math and reading towards the 2014 deadline, including students from racial minority backgrounds, with special education needs, or limited English proficiency. Schools that did not reach their AYP target were subjected to severe sanctions, such as the loss of federal funding (No Child Left Behind Act 2001).

The law required all states to submit a plan to the U.S. Department of Education, detailing how each state would determine adequate yearly progress, what assessments and reporting would be used, and how the state would meet other additional requirements. In Oregon, the statewide assessment system was the Oregon Assessment of Knowledge and Skills (OAKS),

which consisted of measuring student performance in mathematics, reading/literature, science, and social science through multiple-choice tests aligned with grade-level content standards. The tests were administered through a computer-adaptive testing system, although students were also offered to take tests in Spanish and Russian for English Language Learners, and Extended Assessments for students with special education needs. The assessments in math and reading were used for accountability to NCLB standards. The purpose of OAKS was solely to provide information to teachers and administrators about individual student progress toward meeting the high school "certification of mastery of the knowledge and skills" content standards (p. 4, Oregon Department of Education 2010). Students did not face official consequences for not meeting the threshold defining the AYP of the particular school, but merely provided information on individual student progress.

Students received information about their test performance several months after taking the OAKS. The report includes the student's scaled score, the standard error of measurement (which describes the precision of the score), and the performance level associated with the student score. The performance level provides a way for parents and students to interpret what the scaled score means in relation to the target score for the subject, grade, and academic year. Thus, these reports give students a substantial amount of information about not only their scaled score, but also on how the student performed relative to the achievement standards. While the label does not provide additional information, the label allows for an intuitive interpretation of the student's score.

Prior Research on the Effects of Labels in School

While performance labels with school or state-defined sanctions have been studied greatly, there is a dearth of research on whether the impact of labels associated with students' performance remains despite the absence of sanctions or consequences for the student. We review literature on studies examining official consequences for statewide assessments (e.g. grade retention) and then research that examines labels without state-defined sanctions.

We can consider research on grade retention, the practice of a student repeating another academic year with the intention that this student will catch up with their peers academically and socially. Grade retention is usually an official consequence of students receiving less positive performance labels on statewide assessments. However, the consensus is mixed as to whether retention has benefits for students. Earlier studies found few beneficial effects for academic achievement (Jimerson 2001a; Jimerson and Ferguson 2007; Wu, West, and Hughes 2008) and harmful effects for high school completion (Alexander, Entwisle, Dauber, and Kabbani 2004; Jimerson, Anderson, and Whipple 2002; Rumberger and Larson 1998). Other researchers have found positive effects in subsequent achievement and little to no impact for future outcomes (Allen et. al 2009; Greene and Winters 2007; Jacob and Lefgren 2009; Mariano and Martorell 2013). Findings from Andrew (2014) suggest that performance labels with official consequences like grade retention can have a long-lasting "scarring" impact over time, as students retained in primary grade school were 60 percent less likely to complete high school in the future. Overall, these findings paint a conflicted picture of the impact of grade retention, a specific consequence for negative performance labels on statewide tests.

There are fewer studies that have examined performance labels in the context of no consequences for students. Papay, Murnane, and Willet (2016) find evidence that performance

labels on Massachusetts mathematics test impact students' college-going decisions among urban low-income students, even in a setting where there are no official consequences for students. In this way, they examined performance labels in a "low-stakes" setting – that is, these tests in eighth grade hold schools and districts accountable, but not students or individual teachers. Students who received a positive label were more likely to decide to enroll in college, and this effect was strongest among students who previously stated they did not intend to enroll in college by tenth grade. However, it is unclear whether these labels matter if students are in primary grade school, and whether these labels can have lasting consequences over time.

Third grade reading statewide assessments may be a particularly important to examine the impact of performance labels for students' future outcomes. Given that students begin taking statewide assessments in most states in third grade (No Child Left Behind Act of 2001; Every Student Succeeds Act of 2015), it is important to examine performance labels associated with assessments that students begin taking at the state-level. Students, teachers, and schools are likely under increased pressure to improve third grade statewide assessments, since it represents the trajectory of the students' academic growth over time. Third grade is also an important transition for students, as they switch from a "Learning to Read" approach to a "Reading to Learn" approach (Hernandez 2011; Zakariya 2015). Appendix D shows an example of an OAKS question, which illustrates that students are not only tested on the definition of words, but also need to demonstrate proficiency in reading comprehension.

While there are few studies that examine reading in third grade specifically, some studies suggest that third grade reading literacy may have long-term consequences than other subjects or grades. Denning, Murphy, and Weinhardt (2018), for example, find that students with a higher third grade academic rank have better educational outcomes 19 years later. Moreover, third grade

reading skills significantly predict future academic and behavioral outcomes (Hernandez 2011). Students who are at risk of dropping out of school can be identified retrospectively as early as third grade on the basis of attendance patterns, academic performance, and behavior (Lehr, Sinclair, and Christenson 2004). Additionally, students' future educational decisions depend on their academic performance (Jacob and Linkow 2011; Papay, Murnane, & Willett 2016). Thus, we would expect that, given that a performance label is an intuitive way for students to interpret their reading skills, students would assess their reading skills with their peers and for future academic performance and behavior.

Current Study

In this paper, we focus on performance labels that do not have officially state-defined consequences for students. We examine how students respond to feedback based on their test performance – a label that they receive for the first time as third graders on the OAKS. Using a regression-discontinuity design, we examine the impact of labeling by comparing the educational achievement and behavioral outcomes of third graders who were assigned exogenously to different labels because they scored just below or just above the state-mandated labeling cut-points. We focus on testing in reading based on previous research that finds that third grade reading skills in particular have significantly predicted future academic and behavioral outcomes (Schwerdt, West, and Winters 2017). We use this approach to examine test performance labels on academic and behavioral outcomes when students are in fourth grade and when students are in eighth grade. Thus, my central research questions ask: (1) does the performance label that third graders receive from their OAKS reading score affect their future academic achievement? And, (2) does the performance label that third graders receive from their OAKS reading score affect their future attendance and probability of suspension?

Data

We use longitudinal data from the Oregon Department of Education covering the 2004-05 through 2014-15 school years. For our primary analyses, we use data from six cohorts of students enrolled in third grade in 2005 through 2010, took the reading assessments for the OAKS, and had non-missing values for characteristics and outcome measures. In total, the analytic sample contains 281,973 third-graders, with varying samples for students near each cut-point.

We examine the effect of test performance labels assigned from the Oregon Department of Education for the third grade reading OAKS assessment on four outcomes – reading test score percentile rank, math test score percentile rank, school absences, and suspension. For percentile ranks of test scores, we rank the student's test scores among their peers from their OAKS math and reading test score for that academic year. For school absences, we use the total number of absent days in the academic year. We collapse in-school and out-of-school suspensions into a binary outcome, excluding expelled students. Note, however, suspension data is only available from 2007-2008 through 2014-2015 (see Table 3.1 for data structure).

Sample

Table 3.2 provides descriptive statistics of characteristics and outcomes in the entire sample of third graders in Oregon from the 2004-2005 through 2014-2015 academic years (n = 281,973). About 68 percent of the sample is White and 18% of students are Latinx. Only 3 percent of students are Black, while students identified as Other comprise 11 percent of the sample. The sample is fairly evenly split by gender (49 percent girls and 51 percent boys). We also include indicators of whether students participated in programs while enrolled in school.

Half of the sample is identified as participating in the free or reduced lunch (FRL) program in third grade, while 10 percent of students are classified as English Language Learners (ELL). Almost 14 percent of students received special education services. The average third grade OAKS reading score is near 214, which is about 13 points higher than the 201 *Meet the Standard* threshold for years 2005 through 2006 and 10 points higher than the 204 *Meet the Standard* threshold for years 2007 through 2011. In fourth grade, we see the average reading and math scores are near the 50^{th} percentiles, and that students miss about 8 days of school on average. Less than 3 percent of students were ever suspended. In eighth grade, the students' math and reading scores still remain near the 50^{th} percentiles, students missed about 10 days of school, and 14% of students were suspended.

Empirical Strategy

To examine the causal effect of performance labels on student outcomes, we estimate four separate models using a regression discontinuity (RD) design. By examining students near the cut-point for each performance label (the forcing variable), we are able to compare outcomes for two groups of students – those who scored at the cut-off and received a more positive label and those who scored below the cut-off yet received a less positive label. If the cut score is determined exogenously, then students on either side of the threshold are similar on observable characteristics. The estimated difference between these two groups of students provides an unbiased estimate of the causal impact of the performance label (Lee and Lemieux 2010). Because the labels are enforced rigidly in that students who score just below the threshold are assigned one label and students who score just above the threshold are assigned a different and more positive label, the RD is sharp.

We estimate all models using ordinary least squares (OLS); although the probability of suspension is a dichotomous outcome, estimates from OLS linear probability models are more easily interpretable. For the analytic strategy, we use the general form below to estimate each of the four models:

$$Y_{ict} = \beta_0 + \beta_1 Read_i + \beta_2 Above_i + \beta_3 (Read_i \times Above_i) + \beta_4 \times X_i + \varepsilon_i \quad (1)$$

where Y_{ict} represents an outcome for student *i* in cohort *c* in year *t*. The variable $Read$ represents the third grade reading test score in OAKS, centered on the placement threshold, and the variable $Above$ indicates whether a student scored at or above the specific cut-point for each performance label assigned to third graders for reading. The interaction term allows the slopes for the reading score and whether this score is above the particular threshold, and estimate the difference between students just below and above the cutoff. X_i represents the covariates in the analytic models; we include dummy variables for student's racial or ethnic category, free or reduced lunch (FRL) status, and English Language Learner (ELL) status. We also include the fixed effect of school and academic year to account for average differences in the outcomes across schools and years. In the model above, β_2 indicates the parameter of interest, which represents the average effect of receiving a performance label on the outcome for a student with a score right at the margin of the threshold. Thus, the estimate value for β_2 being statistically significant and positive would indicate that a detrimental label, as opposed to the more positive label, causes the student's future test scores to increase discontinuously, on average, in the population.

In using RD, several assumptions must be met in order to maintain the internal validity of the RD analyses and to make unbiased causal inferences. First, the cut-score must be determined exogenously and students cannot manipulate their position on the forcing variable relative to the cut-score. The scaling procedures used to determine the cutoffs are complicated and would have

been implausible for a student to manipulate their scores beforehand (Oregon Department of Education 2011). We also present evidence further in the paper that demonstrates that this assumption holds for the analyses.

Another key assumption for an RD design is that students just above and below the threshold are conditionally random (Schochet et al. 2010). We examine covariate balance, or whether relevant student characteristics jump at the threshold that defined the ITT, to check this assumption. In Table 3.3, we present the key results from RD regressions where each observed student characteristic is the dependent variable. While the estimated jump for gender is not significant and small across performance labels, the other point estimates for other traits are statistically significant and large, suggests that there is an imbalance of observed student covariates for gender, race, FRL, ELL, and special education. We address this imbalance by including the student characteristics as covariates in the RD models and exclude students classified as special education because they were most likely assessed differently (Oregon Department of Education 2010).

Main Results

Table 3.4 reports the main results for the effects of performance labels on fourth grade math and reading percentile ranks, number of absent days, and probability suspension. For each of the four outcomes, we estimate models that include terms for the variable of interest (i.e. binary indicator of whether students scored above cut-point), the running variable (third grade OAKS reading scores), and a linear spline that allows this assignment variable to have a distinct relationship above and below the threshold. Each model also includes dummy variables controlling for gender, race, FRL status, and ELL status. (Supplemental models without controls

yield largely similar results, although the magnitude of the point estimates of models without controls is slightly larger). All models condition on school and year fixed effects.

[Insert Table 3.4 here]

Overall, we find that earning a more negative label causes students to perform worse academically, at least at certain performance levels. In Panels A and B of Table 3.4, we present the causal effects of earning a more negative label on fourth grade reading OAKS scores at each of the cut scores of third grade reading in OAKS. Being classified as *Nearly Meet the Standard* as opposed to *Meet the Standard* in third grade decreases fourth grade reading percentile ranks by 3.6 percentage points (p =0.000) and math percentile ranks by 3.2 percentage points (p = 0.000). We can translate these effect sizes to months of learning based on Hanushek, Woessmann, and Peterson (2012), which assumes a scaling factor of 0.25 standard deviations per year for all grades and subjects (see Appendix E for calculation of effect sizes into months of learning). Thus, a back-of-the envelope calculation would give an estimate of 1.5 months less of instruction for students in reading and 1.3 months of less instruction in math in fourth grade. Interestingly, while this is the cutoff that is used to define Adequate Yearly Progress under No Child Left Behind, the schools face sanctions for not meeting this cutoff, not students. While not as large of a percentage point difference, we also find that being classified as *Meet the Standard* instead of *Exceed the Standard* decreases the reading percentile rank by 1.1 percentage points (p = 0.000) and decreases the math percentile rank by 1.4 percentage points (p = 0.000).

On the other hand, we find no effect of earning *Low* instead of *Nearly Meet the Standard* on any academic outcomes. While we would expect that the effect would be negative, we surprisingly find that receiving a *Very Low* as opposed to *Low* label increased fourth grade reading percentile ranks by 1.8 percentage points (p = 0.000), or three-quarters of one month of

additional learning. One possible explanation is that teachers or schools may pay special attention students identified as needing significant improvement in reading comprehension and skills. Thus, they focus on increasing the test scores of students in the *Very Low* category, and for these students, receiving the least positive label may benefit students academically. Alternatively, students classified with labels in the middle of the distribution, like *Nearly Meet* and *Meet the Standard*, may not receive the same resources that students on either extreme receive to improve test scores (*Very Low* or *Exceed the Standard*).

Panels C and D of Table 3.4 present the behavioral fourth grade outcomes of receiving a more negative label. While we find significant effects for fourth grade academic percentile ranks, we find no effects of these labels for attendance (Panel C), measured as the number of absent days in the school district, in fourth grade. In Panel D of Table 3.4, we find evidence that being classified as *Nearly Meet the Standard* increases the likelihood of suspension in fourth grade by nearly 1 percentage point ($p = 0.074$). While small, this effect is substantively important given that only less than 3 percent of fourth graders were ever suspended in the entire sample. Translated into percentage of likelihood, the 1 percentage point roughly translates to an increase in the probability of suspension by 33 percent in fourth grade.

Table 3.5 parallels Table 3.4, but presents the main results from analyses examining the effect of performance labels received in third grade on academic and behavioral outcomes in eighth grade. Overall, we find that earning a more negative performance label causes students to have lower percentile ranks in their reading and math OAKS scores in five years at certain performance levels. The effects observed in Table 3.4 for reading and mathematics hold considerably, while we also find evidence of the impact of negative performance labels on attendance but not suspension. In Panels A and B of Table 3.5, we present the estimated causal

effects of earning a more negative performance label on reading and mathematics OAKS score percentile ranks in eighth grade at each of the cut scores. Being classified as *Nearly Meet the Standar*d as opposed to *Meet the Standard* in third grade decreases the reading percentile rank by 3.7 percentage points (p = 0.000) and the math percentile rank by 3.1 percentage points (p = 0.000), respectively. Back-of-the-envelope calculations provide an estimate of 1.3 months less of instructional time in reading and 1.2 months less of instructional time in math in eighth grade. We also see a negative effect on math and reading percentile ranks if a student is labeled *Meet the Standard* as opposed to *Exceed the Standard* (1.4 percentage points and 1.6 percentage points, respectively), although these effects are not as large as for receiving *Nearly Meet the Standard*.

[Insert Table 3.5]

The results in Table 3.5 show no effect of earning the label *Low* instead of *Nearly Meet the Standard* on any of the outcomes. Earning a *Very Low* as opposed to *Low* label increases the reading percentile rank by 2.2 percentage points and the math percentile rank by 2.8 percentage points. Back-of-the-envelope calculations suggest that students in the *Very Low* category benefit from almost an additional full month of learning in reading and a little over a month of learning in math by the end of eighth grade.

For Panel C in Table 3.5, we find that being classified as *Meet the Standard* instead of *Exceed the Standard* increases the number of absent days by almost half a day (p = 0.002). While missing a half of a school day may seem harmless for students behaviorally, given that the average number of absent days in eighth grade is around 10 days, additional absent days may increase the risk of drop out in high school (Parr and Bonitz 2015). In Panel D of Table 3.5, the marginally significant coefficient found in Table 3.4 for suspension disappears, suggesting that

receiving a negative performance label does not increase the likelihood of suspension in eighth grade.

Discussion

In this study, we used a regression discontinuity design to estimate the effects of receiving a more negative label for the population of students near performance label cut-offs on future achievement, school absence, and likelihood of suspension. Our analyses show evidence of substantial negative effects of a less positive label for students for certain performance labels, while for students at the lowest category for reading appeared to benefit from this classification. We also find evidence that these negative effects do persist over time, as we note declines in achievement, and an increased number of school absences in five years. While regression discontinuity design can overcome selection issues by estimating the effect for students near a threshold, the estimates cannot speak to outcomes among students who are further away from this threshold. Nonetheless, our analyses show plausibly causal effects, suggesting that the negative effects of performance labels can be long-term.

Importantly, one key limitation of our methodology using an RD approach is that it estimates the effect for students near a particular threshold. However, the strength of this study is the ability to examine the impact of labeling at different cut-points and to estimate not only the immediate effects in their fourth grade, but also to see whether these effects last over time into students' eighth grade outcomes. Another limitation is that, while the RD approach is able to provide causal inferences, our study cannot speak to the precise mechanisms to explain how labeling impacts student outcomes.

Implications of our study suggest that performance labels may also contribute to schooling processes such as tracking/sorting students into classrooms and ability-based grouping (e.g. gifted or special education) based on an arbitrary label, rather than solely on academic performance (Domina, Penner, and Penner 2017). This sorting thus exposes students to different incentives and educational opportunities, and thus advantages or disadvantages accumulate over the course of a student's trajectory through school and into other institutions. Our research also contributes to the prior research that examines officially defined consequences for labels, and we add to this literature that performance labels still have an impact, regardless of whether students face sanctions for a negative performance label.

Our findings also suggest that, while performance labels should not matter for students, they do matter for their future achievement and behavior. We demonstrate that there are still the unintended consequences of performance labels that contribute to categorical inequality that is produced and maintained within schools. Thus, we might ask, what purpose do these performance labels serve for students, teachers, and parents if there are no officially defined consequences? While it may be beneficial for educators to identify students at the extreme ends for placement into gifted or "at-risk" groups, there is simply no purpose for students to find out that they did not meet the standard for reading if the school does not use this information. One policy suggestion is that schools can explore alternatives for performance labels to identify students who are academically struggling in third grade, but does not impact students who are near the threshold of meeting the standard for that particular year. Given that a substantial number of students are near this threshold, it is important for schools to consider how best to improve test scores and reading literacy among students near this threshold.

CHAPTER 5: CONCLUSIONS

My dissertation began by highlighting the difficulty of assessing whether negative results from course failure may be driven by differences between students who do and do not fail, and whether all students are equally impacted. I proceeded to approach this question drawing from human development, categorical inequality, and life course perspectives. My dissertation seeks to ensure that the differences I observe are not being driven by differences in prior achievement by employing strong methodological design and longitudinal data with transcript information. In the conclusion, I summarize the central findings from my dissertation, discuss the implications of these findings for theories on categorical inequality and the life course approach, and then discuss some of the more promising areas for future research raised by my findings. I end the dissertation with concluding remarks.

Summary of Findings

To summarize the findings of this dissertation, the results can be organized into two themes: differences across demographic and institutional contexts and life course trajectories. However, before detailing the main findings in each of these areas, I highlight the three central findings of my dissertation. First, my findings show that course failure has both short-term and long-term outcomes among students, such as degree attainment, wages over time, academic performance, and student behavior. Second, I find statistically significant differences in both the rates and outcomes of course failure by gender, race, and school type, suggesting that social factors operating within education are important in producing these differences. Third, students from disadvantaged backgrounds experience more harmful consequences of failure than their privileged counterparts, which I argue is evidence that failure is a social category through which resources and opportunities are allocated.

Differences across demographic and institution type

1. Among college students interested in pursuing a STEM degree, there are no statistically significant differences between men who fail introductory calculus, but there is a statistically significant difference between women who fail introductory calculus.
2. Gender predicts taking and failing remediation among those who began at four-year colleges, but not at two-year colleges.
3. Black students were more likely to take and fail remedial coursework at both two-year and four-year colleges.
4. Remedial course performance, the likelihood of bachelor degree attainment, and time to bachelor's degree completion varied across four-year and two-year colleges.
5. Academic performance and student behavioral outcomes varied by different performance categories even though students were observationally similar to each other on the third grade reading assessment.

Life course trajectories

1. Women who were interested in pursuing a STEM degree and failed an introductory calculus course were less likely to attain a STEM degree than men, obtaining degrees in humanities or social sciences.
2. Failing remediation lowers the likelihood of attaining a bachelor's degree if a student entered a four-year college, but not a two-year college. Rather, students were more likely to receive a bachelor's degree if they failed a remedial course and started at a two-year college.
3. Students who fail their remedial coursework earn less wages over time if they started at a two-year college, even after accounting for highest degree attained.

4. The scores of students who received a less positive label in third grade reading (except for the lowest possible label) declined in fourth and eighth grade, and students miss more days of school over time.

Implications for Categorical Inequality and Life Course Theories

In this section, I discuss the implications of these findings for theories on categorical inequality in education and the life course perspective on timing, turning points, and the future of students' trajectories.

Categorical Inequality

The analyses from this dissertation demonstrate not only the importance of considering categories based on failure to understand processes of inequality, but also explore how different characteristics can compound the advantages or disadvantages from course failure.

These variations found in this dissertation have implications for grit theory about student differences in achievement and access to educational opportunities. For example, the finding that observationally similar third-graders just above and below a threshold for a specific performance label have different future academic performance and student behavior is difficult for the grit theory to accommodate. It is of course possible that grit theory would explain differences in motivation across students that is not captured in the Oregon administrative data, however, if this were the case, students would need to know beforehand how test scores for reading are calculated and what the score cut-offs are for each performance label in order to manipulate their placement in the distribution of scores. Moreover, we would expect that students who receive a positive performance label would usually experience positive outcomes; however, findings show that students receiving the lowest possible performance label actually have better outcomes

compared to those whose score was in the performance label above. These analyses suggest that the grit theory that emphasizes motivational differences needs to account for structural contexts before it is useful in explaining differences in outcomes in cases where students are observationally similar in prior ability, demographic characteristics, and academic performance.

The differences across gender presented in this dissertation also have implications for theories on categorical inequality in education. The finding that failing calculus weeds out women but not men from earning a bachelor's degree in a STEM field speaks to the argument on how course failure serves as potential sites for sorting students within education, where students in privileged positions are not harmed by failure to the extent that students in disadvantaged positions experience. There may be unobserved differences between the students who did and did not fail calculus, however, we would expect to see no gender differences in responses to failing calculus – similar to what we observed for other similar weed-out courses and the introductory writing course. Rather, because calculus is a predominantly male-typed subject, men perceive themselves as having more advantage in the course and continue their pursuit of a STEM degree regardless of failure, while women are at a disadvantage in failing a subject that is stacked against them. Moreover, women encounter academic and social consequences of failing a "gatekeeping" course for a bachelor's degree in STEM that their male counterparts do not (to the same extent).

The findings on the impact of remedial course failure across two-year and four-year colleges presented in this dissertation suggest that the effect of failure also varies by the resources and opportunities available to students at the institutional level. Results from the second empirical chapter show disparate patterns for the effects of remedial course performance among students entered a two-year or four-year college. Students placed in remediation in

college face additional barriers that their non-remedial counterparts do not, in that remedial coursework usually does not count towards the degree. Additionally, findings show that racial minority students are the most likely to be taking and failing remediation than other groups. Thus, racial minority students being placed in remedial coursework compounds the disadvantages associated with these statuses. However, the harm of these disadvantages are mitigated by the institution type. For example, two-year college entrants appear to benefit from remedial coursework even if they fail a remedial course, while four-year college entrants are less likely to obtain a bachelor's degree after failing a remedial course. This is because taking (and failing) remediation is the juncture in which students are sorted, and produce different signals to similar groups of students because of the institutional context.

In sum, I find that differences across demographic and institution types explain one of the ways categorical inequality operates within education, and how resources are distributed and allocated based on the definition of these categories. My findings also indicate that grit theory, or identifying motivation and persistence among students, falls short in explaining the differences across demographic and institutional characteristics. While it is difficult to isolate motivation and persistence, I show that categories of failure play an important role in how students are sorted in education, shaping the opportunities and incentives to which these students are exposed.

Life Course: Timing, Turning Points, and Trajectories

The findings from this dissertation show that it is important to consider failure as a turning point in the educational career, with advantages and disadvantages compounding over a series of developmental sequences in education. Examining failure as a turning point also allows

for researchers to understand the extent to which these categories can have long-lasting implications in institutions outside of education, such as the labor market.

We can draw implications from each empirical chapter's findings for life course theories for failure as a turning point that shapes students' future trajectories and transitions into adulthood. Chapter 2 examined gendered responses to failing introductory to calculus, which often serves as a gatekeeping purpose across STEM discipline by limiting the rate at which students progress to more advanced coursework in their major. However, my findings indicate that calculus also serves as a gatekeeper for careers that require a bachelor's degree in a STEM field. These findings speak to the argument that failing calculus is a turning point within a student's trajectory for both their educational attainment and eventual occupational opportunities. However, it is important to note that course failure in calculus is a significant turning point for women interested in pursuing a STEM degree because of the stigma attached to mathematics and female-typed ability. Failing calculus does not appear to be a significant turning point for men in their trajectory, but there are other possible events in the educational career that may push these students away from earning a STEM degree as well.

We can also draw implications from the findings of Chapter 3, which examined student-level and institution-level differences regarding the outcomes of remedial coursetaking and failure. Remediation and academic performance in remediation can be argued as a significant turning point within a college student's trajectory for transitions to adulthood outcomes. Remediation blocks or grants access to college-level coursework in higher education. The access for advanced college-level coursework is a necessary requirement for attaining a bachelor's degree, and is also important for negotiating salary and wages in the labor market for employment post-graduation. This is especially important in light of the results that show that

students who entered two-year college and failed a remedial course earn less over time than their peers, even after accounting for degree completion. In this case, remedial course failure acts as a more significant turning point (more harmful impacts over time) by delaying entry into the labor market for students and the loss of wages because of this delayed entry. Thus, I show that the sorting function of remediation and remedial course failure can have long-lasting implications beyond educational attainment and into their wages over time.

Finally, Chapter 4 provides evidence for how school contexts shape not only the acquisition of skills but also later life outcomes by altering the educational contexts to which students can access. Given that the U.S. contemporary educational system has moved towards using singular indicators of performance to determine advancement or progress, third grade reading assessments is a particularly important turning point for youth in this stage of development. Students begin taking statewide assessments in third grade and they also begin to switch from a "Learning to Read" to a "Reading to Learn" approach in their education. Thus, the findings show that a performance label, with no official state-defined consequences, still impact students' future academic and behavioral outcomes not only in the subsequent year, but also five years later. Understanding schools as a series of formative environments accommodates the findings from Chapter 4, given that the official purpose of performance labels is to provide an intuitive way for third graders to interpret their scores relative to their peers. While the precise mechanisms are still unclear, we can expect that these performance labels play a large role in how students assess their own reading skills, and how teachers or educators assess the reading skills of students, making decisions based on this information. These decisions have lasting implications as students are granted access or blocked from accessing certain educational

opportunities, which may translate into declines in achievement and an increased risk to dropping out of school altogether.

In sum, I find evidence that course failure is a significant turning point within a student's trajectory, and explain one of the ways school contexts shape later life outcomes. My findings also indicate that structural factors play important roles in the developmental cascades that students move through in education and into their adulthood.

Policy Implications

In this section, I discuss the implication of this dissertation's findings related to issues of policy. I focus on the implications of categorical inequality, given that these are crucial junctures to where advantages and disadvantages accumulate over time, although policy concerns should still consider both the short-term and long-term consequences of failure as a turning point in the educational career for students' future trajectories.

First, it is important for educators to understand that one-size-fits-all solutions are not appropriate across institutional, demographic, and timing contexts. For example, findings from Chapter 3 provide evidence that remediation serves different purposes in higher education, and while beneficial to some, can be harmful to others. Given that we know that students need different kinds of interventions and support programs in different formative environments and stages of development, it would be most beneficial for policymakers to focus on what interventions and programs work best for students who are failing, and implement these support structures that best address the needs of students. For example, some institutions may benefit from emulating some of the support systems enacted in certain contexts, and apply this support both efficiently and effectively for students to succeed.

We can also draw policy implications from the longitudinal aspects of these studies. While the results of my analyses are not meant to make causal statements, they indicate that course failure can have lasting implications in the successful transitions into adulthood. This suggests that addressing the key turning points within a students' trajectory could play a role in mitigating the widening gap in inequality over time. Investment in these key turning points also highlights the important role course failure plays in certain contexts but not in others, or why the same life event can have lasting consequences for some, but not others.

Future Directions

While the conclusion is to summarize and complete the explanations for the findings in this dissertation, there are also many questions raised in light of the findings. Given that course failure has not been examined through the life course and categorical inequality lens, these findings represent a novel approach to how we understand educational categories and labels in general. Thus, in this section, I lay out several avenues for future research suggested by this dissertation.

Future studies can build on the regression discontinuity models in Chapter 4 to estimate the causal effects of mathematics course failure in middle school on considerations like attending post-secondary education, labor market outcomes, and family formation decisions. This is possible through unique linkages between Census Bureau data and administrative data that enable researchers to examine the short, medium, and long-term effects associated with course failure. Additionally, the findings from Chapter 4 can be extended to examine the impact of labels on high school graduation and college attendance within the state of Oregon. It would be interesting if these analyses could be replicated in a setting where performance labels do have

officially defined consequences, and thus determine the extent to which these labels impact future academic and behavioral outcomes.

One of the larger unanswered questions from my findings is, what exactly is happening at the school and individual level that contributes to these larger patterns of inequality observed across time and contexts? A potential avenue for future research is to identify and examine the potential mechanisms. These mechanisms may include how individuals self-assess based on larger societal beliefs about their identity vis-à-vis academic performance, which may illuminate how students make educational investment decisions and how they respond to course failure in light of this self-assessment. It could also reveal the underlying processes on how teachers and educators interpret student failure and incorporate this negative signal into their decisions to intervene or place these students on a diverging pathway, where these students likely access different opportunities and resources than their peers.

Lastly, another avenue of research can broaden the analyses of course failure outside of the U.S. context for international comparisons. For example, in Japan, placement into high school and universities is highly dependent on entrance exam scores. This setting has a more obvious application of failure as a turning point for a student's access to future opportunities and resources. However, would a similar application of failure be appropriate in contexts where alternative options exist for students who do not participate in entrance exams, as is the case in Germany? These cross-national comparisons would illuminate other aspects of course failure that contribute to processes of educational inequality on a global scale.

Concluding Remarks

Collectively, these studies provide a nuanced understanding of course failure and its impact on students' future outcomes. My dissertation thus underscores how events in an

educational career, such as course failure, can lead to the accumulation of advantages or disadvantages over time to have a substantial impact on students' future outcomes. That is, to the degree that course failure shapes the educational resources and incentives to which students are exposed, students who failed a course will likely access vastly different opportunities in education, the labor market, and other institutions. My work is careful to ensure that the differences I observe are not being driven by differences in prior achievement, as I use three separate longitudinal datasets and methodologies that closely approach causally estimating the impact of course failure.

In my dissertation, I thus demonstrate that a better way of understanding inequality in education emerges through a consideration of how schooling processes, such as course failure, contribute to the unequal distribution of resources to which students are exposed. Furthermore, in documenting the differences in the consequences of course failure, my findings connect to broader sociological inquiry about how social categories, such as race and gender, interact with schools to shape students' responses to and impacts from course failure. While educational policies primarily focus on ameliorating the short-term impacts of poor academic performance, my scholarly approach identifies how and when students slip through the proverbial cracks within the education system. In doing so, my research provides insight on how educators can support students to obtain the skills needed for future success.

REFERENCES

Adelman, Clifford. 1999. Answers in the toolbox: Academic intensity, attendance patterns, and bachelor's degree attainment. Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education.

Adelman, Clifford. 2004. Principal Indicators of Student Academic Histories in Post-Secondary Education, 1972-2000. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Adolph, Karen E., Whitney G. Cole, Meghana Komati, Jessie S. Garciaguirre, Daryaneh Badaly, Jesse M. Lingeman, Gladys LY Chan, and Rachel B. Sotsky. 2012. "How Do You Learn to Walk? Thousands of Steps and Dozens of Falls per Day." *Psychological Science* 23(11): 1387-1394.

Alexander, Karl L., Doris R. Entwisle and Susan L. Dauber. 2003. *On the Success of Failure: A Reassessment of the Effects of Retention in the Primary School Grades*. 2nd ed. New York, NY: Cambridge University Press.

Alexander, Karl L., Doris R. Entwisle, Susan L. Dauber and Nader Kabbani. 2004. “Dropout in Relation to Grade Retention: An Accounting from the Beginning School Study.” Pp. 5–34 in *Can Unlike Children Learn Together? Grade Retention, Tracking, and Grouping*, edited by H. J. Walberg, A. J. Reynolds, and M. C. Wang. Greenwich, CT: Information Age Publishing.

Alexander, Karl L., Doris R. Entwisle, and Carrie S. Horsey. 1997 "From First Grade Forward: Early Foundations of High School Dropout." *Sociology of Education* 70(2): 87-107.

Allen, Chiharu S., Qi Chen, Victor L. Willson, and Jan N. Hughes. 2009. "Quality of Research Design Moderates Effects of Grade Retention on Achievement: A Meta-Analytic, Multilevel Analysis." *Educational Evaluation and Policy Analysis* 31(4): 480-499.

Allison, Paul D. 2001. *Missing Data*. Sage University Papers Series on Quantitative Applications in the Social Sciences. 07-136. Thousand Oaks, CA: Sage Publications.

Allison, Paul D. 2003. "Missing Data Techniques for Structural Equation Modeling." *Journal of Abnormal Psychology* 112(4): 545-557.

Alon, Sigal, and Marta Tienda. 2007. "Diversity, Opportunity, and the Shifting Meritocracy in Higher Education." *American Sociological Review* 72(4): 487-511.

Andrew, Megan. 2014. "The scarring effects of primary-grade retention? A study of cumulative advantage in the educational career." *Social Forces* 93(2): 653-685.

Attewell, Paul, and Thurston Domina. 2008. "Raising the Bar: Curricular Intensity and Academic Performance." *Educational Evaluation and Policy Analysis* 30(1): 51-71.

Attewell, Paul, David Lavin, Thurston Domina, and Tania Levey. 2006. "New Evidence on College Remediation." *The Journal of Higher Education* 77(5): 886-924.

Bahr, Peter R. 2008. "Does Mathematics Remediation Work?; A Comparative Analysis of Academic Attainment among Community College Students." *Research in Higher Education* 49(5): 420-450.

Bailey, Thomas, Katherine Hughes, and Shanna Smith Jaggars. 2012. "Law Hamstrings College Remedial Programs." *Hartford Courant*, May 18. Retrieved February 5, 2019 (https://www.courant.com/opinion/hc-xpm-2012-05-18-hc-op-bailey-college-remedial-education-bill-too-r-20120518-story.html).

Bailey, Thomas, Dong Wook Jeong, and Sung-Woo Cho. 2010. "Referral, Enrollment, and Completion in Developmental Education Sequences in Community Colleges." *Economics of Education Review* 29(2): 255-270.

Baird, Matthew, and John F. Pane. 2018. "Translating Standardized Effects of Education Programs Into More Interpretable Metrics." RAND Education.

Barreca, Alan I., Melanie Guldi, Jason M. Lindo, and Glen R. Waddell. 2011. "Saving Babies? Revisiting the Effect of Very Low Birth Weight Classification." *The Quarterly Journal of Economics* 126(4): 2117-2123.

Barnett, Elisabeth A., Rachel Hare Bork, Alexander K. Mayer, Joshua Pretlow, Heather D. Wathington, and Madeline Joy Weiss. 2012. "Bridging the Gap: An Impact Study of Eight Developmental Summer Bridge Programs in Texas." New York, NY: National Center for Postsecondary Research, Teachers College, Columbia University.

Bereiter, Carl and Engelman, Siegfried. 1966. *Teaching Disadvantaged Children in the Preschool*. New York, NY: Prentice-Hall.

Bettinger, Eric P., and Bridget T. Long. 2009. "Addressing the Needs of Underprepared Students in Higher Education does College Remediation Work?." *Journal of Human Resources* 44(3): 736-771.

Beyer, Sylvia. 1998. "Gender Differences in Causal Attributions by College Students of Performance on Course Examinations." *Current Psychology* 17(4): 346-358.

Boatman, Angela, and Bridget Terry Long. 2018. "Does Remediation Work for All Students? How the Effects of Postsecondary Remedial and Developmental Courses Vary by Level of Academic Preparation." *Educational Evaluation and Policy Analysis* 40(1): 29-58.

Brehm, Sharon S., and Jack W. Brehm. 1981. *Psychological reactance: A theory of freedom and control.* New York, NY: Academic Press.

Calcagno, Juan C., and Bridget T. Long. 2008. "The Impact of Postsecondary Remediation Using a Regression Discontinuity Approach: Addressing Endogenous Sorting and Noncompliance. Working Paper No. 14194. Cambridge, MA: National Bureau of Economic Research.

Cech, Erin, Brian Rubineau, Susan Silbey, and Caroll Seron. 2011. "Professional Role Confidence and Gendered Persistence in Engineering." *American Sociological Review* 76(5): 641–66.

Centers for Disease Control and Prevention. 2018. CDC'S Developmental Milestones. Washington, DC: U.S. Department of Health and Human Services.

Charles, Maria, and Karen Bradley. 2009. "Indulging Our Gendered Selves? Sex Segregation by Field of Study in 44 Countries." *American Journal of Sociology* 114(4): 924–76.

Chen, Xianglei. 2013. STEM Attrition: College Students' Paths into and out of STEM Fields. Statistical Analysis Report (NCES 2014–001). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

------. 2016. Remedial Coursetaking at US Public 2-and 4-Year Institutions: Scope, Experiences, and Outcomes (NCES 2016-405). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Clifford, Margaret M. 1979. "Effects of Failure: Alternative Explanations and Possible Implications." *Educational Psychologist* 14(1): 44-52.

Clotfelter, Charles T., Helen F. Ladd, Clara Muschkin, and Jacob L. Vigdor. 2015. "Developmental Education in North Carolina Community Colleges." *Educational Evaluation and Policy Analysis* 37(3): 354-375.

Correll, Shelley J. 2001. "Gender and the Career Choice Process: The Role of Biased Self-Assessments." *American Journal of Sociology* 106(6): 1691-1730.

Correll, Shelley J. 2004. "Constraints into Preferences: Gender, Status, and Emerging Career Aspirations." *American Sociological Review* 69(1): 93-113.

Crisp, Gloria, and Chryssa Delgado. 2014. "The Impact of Developmental Education on Community College Persistence and Vertical Transfer." *Community College Review* 42(2): 99-117.

Crisp, Gloria, Amaury Nora, and Amanda Taggart. 2009. "Student characteristics, pre-college, college, and environmental factors as predictors of majoring in and earning a STEM degree: An analysis of students attending a Hispanic serving institution." *American Educational Research Journal* 46(4): 924–42.

Deil-Amen, Regina, and James E. Rosenbaum. 2002 "The Unintended Consequences of Stigma-Free Remediation." *Sociology of Education* 75(3): 249-268.

Denby, David. 2016. "The Limits of Grit." *The New Yorker*, June 21, Retrieved from https://www.newyorker.com/culture/culture-desk/the-limits-of-grit.

Dennebaum, Joanne M., and Janet M. Kulberg. 1994. "Kindergarten Retention and Transition Classrooms: Their Relationship to Achievement." *Psychology in the Schools* 31(1): 5-12.

DiPrete, Thomas A., and Gregory M. Eirich. 2006. "Cumulative Advantage as a Mechanism for Inequality: A Review of Theoretical and Empirical Developments." *Annual Review of Sociology* 32: 271-297.

Domina, Thurston, Andrew Penner, and Emily Penner. 2017. "Categorical Inequality: Schools as Sorting Machines." *Annual Review of Sociology* 43: 311-330.

Duckworth, Angela L. 2013. "The Key to Success? Grit." Retrieved from https://tedsummaries.com/2014/04/08/angela-lee-duckworth-the-key-to-success-grit/

Duckworth, Angela L., Christopher Peterson, Michael D. Matthews, and Dennis R. Kelly. 2007. "Grit: Perseverance and Passion for Long-Term Goals." *Journal of Personality and Social Psychology* 92(6): 1087-1101.

Elder, Glen H. Jr. 1998. "The life course as developmental theory." *Child Development* 69(1):1-12.

Elder, Glen H. Jr., Monica Kirkpatrick Johnson, and Robert Crosnoe. 2007. "The Emergence and Development of Life Course Theory." Pp. 3-19 in *Handbook of the life course*, edited by J.T. Mortimer and M.J. Shanahan. Boston, MA: Springer.

Every Student Succeeds Act of 2015, Public Law 114-95, 192 U.S. Statutes at Large 1802 (2015).

Fishman, Charles. 2016. "Face Time with Meg Whitman." *Fast Company*, April 30. Available online: http://www.webcitation.org/6k6rVdoQW (accessed on 28 August 2016).

Frey, Nancy. 2005. "Retention, Social Promotion, and Academic Redshirting: What Do We Know and Need to Know?" *Remedial and Special Education* 26(6): 332-346.

Gottfried, Michael A. 2013. "Retained Students and Classmates' Absences in Urban Schools." *American Educational Research Journal* 50(6): 1392-1423.

Greene, Jay P., and Marcus A. Winters. 2007. "Revisiting Grade Retention: An Evaluation of Florida's Test-Based Promotion Policy." *Education Finance and Policy* 2(4): 319-340.

Hanushek, Eric A., Paul E. Peterson, and Ludger Woessmann. 2012. "Achievement Growth: International and US State Trends in Student Performance. PEPG Report No.: 12-03." Program on Education Policy and Governance, Harvard University.

Hanushek, Eric A., and Margaret E. Raymond. 2005. "Does School Accountability Lead to Improved Student Performance?" *Journal of Policy Analysis and Management* 24(2): 297-327.

Heckhausen, Jutta, and Richard Schulz. 1995. “A Life-Span Theory of Control.” *Psychological Review* 102(2): 284–304.

Hernandez, Donald J. 2011. "Double Jeopardy: How Third-Grade Reading Skills and Poverty Influence High School Graduation." *Annie E. Casey Foundation*: 1-15.

Hodara, Michelle, and Di Xu. 2016. "Does Developmental Education Improve Labor Market Outcomes? Evidence from Two States." *American Educational Research Journal* 53(3): 781-813.

Holmes, C. Thomas. 1989. “Grade Level Retention Effects: A Meta-Analysis of Research Studies.” Pp. 16-33 in *Flunking Grades: Research and Policies on Retention. Education Policy Perspectives*, edited by L.A. Shepard and M.L. Smith. Breistol, PA: Taylor & Francis.

Hout, Michael. 2012. "Social and Economic Returns to College Education in the United States." *Annual Review of Sociology* 38: 379-400.

Hu, Shouping, Toby J. Park, Chenoa S. Woods, David A. Tandberg, Keith Richard, and Dava Hankerson. 2016. Investigating Developmental and College-Level Course Enrollment and Passing before and after Florida's Developmental Education Reform (REL 2017-203). Washington, DC: U.S. Department of Education, Institute of Education Sciences,

National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast.

Jacob, Brian A., and Lars Lefgren. 2009. "The Effect of Grade Retention on High School Completion." *American Economic Journal: Applied Economics* 1(3): 33-58.

Jacob, Brian A., and Tamara Wilder Linkow. 2011. "Educational Expectations and Attainment." Pp. 133-163 in *Whither Opportunity?: Rising Inequality, Schools, and Children's Life Chances*, edited by G.J. Duncan and R. J. Murnane. New York, NY: Russell Sage Foundation.

Jaggars, Shanna S., and Di Xu. 2016. "How Do Online Course Design Features Influence Student Performance?." *Computers & Education* 95: 270-284.

Jimenez, Laura, Scott Sargrad, Jessica Morales, and Maggie Thompson. 2016. Remedial Education: The Cost of Catching Up. Washington, DC: Center for American Progress.

Jimerson, Shane R. 2001a. "A synthesis of grade retention research: Looking backward and moving forward." *The California School Psychologist* 6(1): 47-59.

------. 2001b. "Meta-Analysis of Grade Retention Research: Implications for Practice in the 21st Century." *School Psychology Review* 30(3): 420-437.

Jimerson, Shane R., Gabrielle E. Anderson, and Angela D. Whipple. 2002. "Winning the Battle and Losing the War: Examining the Relation between Grade Retention and Dropping Out of High School." *Psychology in the Schools* 39(4): 441-457.

Jimerson, Shane R., and Phillip Ferguson. 2007. "A Longitudinal Study of Grade Retention: Academic and Behavioral Outcomes of Retained Students through Adolescence." *School Psychology Quarterly* 22(3): 314-339.

Johnson, Rucker C. 2007. “Wage and Job Dynamics After Welfare Reform: The Importance of Job Skills.” Pp. 231-298 in *Aspects of Worker Well-Being (Research in Labor Economics, Volume 26)*, edited by S.W. Polacheck and O. Bargain. Bingley, West Yorkshire, UK: Emerald Group Publishing Limited.

Kena, Grace, Lauren Musu-Gillette, Jennifer Robinson, Xiaolei Wang, Amy Rathbun, Jijun Zhang, Sidney Wilkinson-Flicker, Amy Barmer, and Erin Dunlop Velez Velez. 2015. The Condition of Education 2015 (NCES 2015-144). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Kohn, Alfie. 2014. "Grit: A Skeptical Look at the Latest Educational Fad." *Independent School* 74(1): 104-08.

Kokkelenberg, Edward C., and Esha Sinha. 2010. “Who Succeeds in STEM Studies? An Analysis of Binghamton University Undergraduate Students.” *Economics of Education Review* 29(6): 935–46.

Kurlaender, Michal, and Jessica S. Howell. 2012. Collegiate Remediation: A Review of the Causes and Consequences. Literature Brief. New York, NY: College Board Advocacy and Policy Center.

Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48(2): 281-355.

Lehr, Camilla A., Mary F. Sinclair, and Sandra L. Christenson. 2004. "Addressing Student Engagement and Truancy Prevention During the Elementary School Years: A Replication Study of the Check & Connect Model." *Journal of Education for Students Placed at Risk* 9(3): 279-301.

Liang, Jian-Hua, Paul E. Heckman, and Jamal Abedi. 2012. "What Do the California Standards Test Results Reveal about the Movement toward Eighth-Grade Algebra for All?." *Educational Evaluation and Policy Analysis* 34(3): 328-343.

Logue, Alexandra W., Mari Watanabe-Rose, and Daniel Douglas. 2016. "Should Students Assessed as Needing Remedial Mathematics Take College-Level Quantitative Courses Instead? A Randomized Controlled Trial." *Educational Evaluation and Policy Analysis* 38(3): 578-598.

Lorence, Jon, and Anthony Gary Dworkin. 2006. "Elementary Grade Retention in Texas and Reading Achievement among Racial Groups: 1994–2002." *Review of Policy Research* 23(5): 999-1033.

Mann, Allison, and Thomas A. Diprete. 2013. “Trends in Gender Segregation in the Choice of Science and Engineering Majors.” *Social Science Research* 42(6): 1519–1541.

Mariano, Louis T., and Paco Martorell. 2013. "The Academic Effects of Summer Instruction and Retention in New York City." *Educational Evaluation and Policy Analysis* 35(1): 96-117.

Martorell, Paco, and Isaac McFarlin Jr. 2011. "Help or Hindrance? The Effects of College Remediation on Academic and Labor Market Outcomes." *The Review of Economics and Statistics* 93(2): 436-454.

Masten, Ann S. 1994. “Resilience in individual development: Successful adaptation despite risk and adversity.” Pp. 3–25 in *Educational Resilience in Inner City America: Challenges and Prospects*, edited by M. C. Wang and E. W. Gordon. Manwah, NJ: Erlbaum Associates Inc.

McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics* 142(2): 698-714.

McDermott, Ray P. 1993. “The Acquisition of a Child by Learning Disability.” Pp. 269-305 in *Understanding Practice: Perspectives on Activity and Context*, edited by S. Chaiklin and J. Lave. Cambridge, England: Cambridge University Press.

Melguizo, Tatiana, Johannes M. Bos, Federick Ngo, Nicholas Mills, and George Prather. 2016. "Using a Regression Discontinuity Design to Estimate the Impact of Placement Decisions in Developmental Math." *Research in Higher Education* 57(2): 123-151.

Merisotis, Jamie P., and Ronald A. Phipps. 2000. "Remedial Education in Colleges and Universities: What's Really Going On?." *The Review of Higher Education* 24(1): 67-85.

Mickelson, Roslyn Arlin. 1989. "Why Does Jane Read and Write So Well? The Anomaly of Women's Achievement." *Sociology of Education* 62(1): 47-63.

Nathan, Linda F. 2017. *When Grit isn't Enough: A High School Principal Examines How Poverty and Inequality Thwart the College-for-All Promise*. Boston, MA: Beacon Press.

NCES. 1988. National Education Longitudinal Study of 1988 (NELS:88). Available online: https://nces.ed.gov/surveys/nels88/ (accessed on 10 May 2017).

NCES. 2000. Postsecondary Education Transcript Study (PETS:2000). Available online: https://nces.ed.gov/surveys/pets/about.asp (accessed on 10 May 2017).

Nelson, Lori J., and Joel Cooper. 1997. "Gender Differences in Children's Reactions to Success and Failure with Computers." *Computers in Human Behavior* 13(2): 247-267.

Ngo, Federick. 2018. "Fractions in College: How Basic Math Remediation Impacts Community College Students." *Research in Higher Education* 60(4): 1-36.

Niederle, Muriel, and Lise Versterlund. 2007. “Do women shy away from competition? Do men compete too much?” *The Quarterly Journal of Economics* 122(3): 1067–101.

No Child Left Behind Act of 2001, Public Law 107-110, 30 U.S. Statutes at Large 750 (2002).

Noel-Levitz. 2014. “2014 National Freshman Attitudes Report.” Available online: http://www.webcitation.org/6qLaQPYCr (accessed on 16 January 2017).

Olson, Steve, and Donna Gerardi Riordan. 2012. “Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics.” *Report to the President*. Available online: http://www.webcitation.org/6q8VSH5BE (accessed on 1 May 2017).

Oregon Department of Education. 2010. *The Oregon Statewide Assessment System Annual Technical Report Volume 1 2009-2010*. Salem, OR: Oregon Department of Education.

Ost, Ben. 2010. “The Role of Peers and Grades in Determining Major Persistence in the Sciences.” *Economics of Education Review* 29(6): 923-934.

Owen, Ann L. 2010. “Grades, Gender, and Encouragement: A Regression Discontinuity Analysis.” *Journal of Economic Education* 41(3): 217-234.

Papay, John P., Richard J. Murnane, and John B. Willett. 2016. "The Impact of Test Score Labels on Human-Capital Investment Decisions." *Journal of Human Resources* 51(2): 357-388.

Parr, Alyssa K., and Verena S. Bonitz. 2015. "Role of Family Background, Student Behaviors, and School-Related Beliefs in Predicting High School Dropout." *The Journal of Educational Research* 108(6): 504-514.

Penner, Andrew M. 2015. “Gender inequality in science.” *Science* 347(6219): 234–235.

Penner, Andrew M., and Robb Willer. 2015. “Refusing to fail: Masculine persistence and the gender gap in science and mathematics.” Department of Sociology, University of California Irvine, Irvine, CA. Unpublished manuscript.

Rask, Kevin. 2010. “Attrition in STEM fields at a liberal arts college: The importance of grades and pre-collegiate preferences.” *Economics of Education Review* 29(6): 892–900.

Rask, Kevin, and Jill Tiefenthaler. 2008. “The role of grade sensitivity in explaining the gender imbalance in undergraduate economics.” *Economics of Education Review* 27(6): 676-687.

Reschly, Amy L., and Sandra L. Christenson. 2013. "Grade retention: historical perspectives and new research." *Journal of School Psychol*ogy 51(3): 319-322.

Reynolds, C. Lockwood. 2012. "Where to Attend? Estimating the Effects of Beginning College at a Two-Year Institution." *Economics of Education Review* 31(4): 345-362.

Riegle-Crumb, Catherine. 2006. “The Path Through Math: Course Sequences and Academic Performance at the Intersection of Race-Ethnicity and Gender.” *American Journal of Education* 113(1): 101-122.

Roderick, Melissa, and Eric Camburn. 1999. "Risk and Recovery from Course Failure in the Early Years of High School." *American Educational Research Journal* 36(2): 303-343.

Rose, Todd, and Ogi Ogas. 2018. *Dark horse: Achieving Success Through the Pursuit of Fulfillment*. New York, NY: Harper Collins.

Roth, Susan, and Larry Kubal. 1975. "Effects of Noncontingent Reinforcement on Tasks of Differing Importance: Facilitation and Learned Helplessness." *Journal of Personality and Social Psychology* 32(4): 680-691.

Rubin, Donald B. 1987. *Multiple Imputation for Non-Response in Surveys*. New York, NY: John Wiley & Sons, Inc.

Rumberger, Russell W., and Katherine A. Larson. 1998. "Student Mobility and the Increased Risk of High School Dropout." *American Journal of Education* 107(1): 1-35.

Rumberger, Russell W., and Gregory J. Palardy. 2005. "Does Segregation Still Matter? The Impact of Student Composition on Academic Achievement in High School." *Teachers College Record* 107(9): 1999-2045.

Rutter, Michael. 1985. “Resilience in the Face of Adversity. Protective Factors and Resistance to Psychiatric Disorder.” *British Journal of Psychiatry* 147(6): 598-611.

Ryckman, David B., and Percy Peckham. 1987. "Gender Differences in Attributions for Success and Failure Situations across Subject Areas." *Journal of Educational Research* 81(2): 120-125.

Saxon, D. Patrick, and Hunter R. Boylan. 2001. "The Cost of Remedial Education in Higher Education." *Journal of Developmental Education* 25(2): 2-9.

Schochet, Peter, Thomas Cook, John Deke, Guido Imbens, J. R. Lockwood, Jack Porter, and Jeffrey Smith. 2010. "Standards for Regression Discontinuity Designs." Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_rd.pdf.

Schwerdt, Guido, Martin R. West, and Marcus A. Winters. 2017. "The Effects of Test-Based Retention on Student Outcomes Over Time: Regression Discontinuity Evidence from Florida." *Journal of Public Economics* 152: 154-169.

Scott-Clayton, Judith, Peter M. Crosta, and Clive R. Belfield. 2014. "Improving the Targeting of Treatment: Evidence from College Remediation." *Educational Evaluation and Policy Analysis* 36(3): 371-393.

Scott-Clayton, Judith, and Olga Rodriguez. 2015. "Development, Discouragement, or Diversion? New Evidence on the Effects of College Remediation Policy." *Education Finance and Policy* 10(1): 4-45.

Seymour, Elaine, and Nancy Hewitt. 1997. *Talking about Leaving: Why Undergraduates Leave the Sciences*. Boulder, CO: Westview Press.

Siegler, Robert S., Greg J. Duncan, Pamela E. Davis-Kean, Kathryn Duckworth, Amy Claessens, Mimi Engel, Maria Ines Susperreguy, and Meichu Chen. 2012. "Early Predictors of High School Mathematics Achievement." *Psychological Science* 23(7): 691-697.

Staying Eligible. (2017, September 22). Retrieved from https://studentaid.ed.gov/sa/eligibility/staying-eligible.

Stearns, Elizabeth, Stephanie Moller, Judith Blau, and Stephanie Potochnick. 2007. "Staying Back and Dropping Out: The relationship Between Grade Retention and School Dropout." *Sociology of Education* 80(3): 210-240.

Sweeney, Paul D., Richard L. Moreland, and Kathy L. Gruber. 1982. "Gender Differences in Performance Attributions: Students' Explanations for Personal Success or Failure." *Sex Roles* 8(4): 359-373.

Tinto, Vincent. 1993. *Leaving college: Rethinking the causes and cures of student attrition* 2nd edition. Chicago, IL: University of Chicago Press.

Valencia, Richard R. 1997. *The Evolution of Deficit Thinking: Educational Thought and Practice*. Abingdon, OX: Routledge.

Visher, Mary G., Michael J. Weiss, Evan Weissman, Timothy Rudd, and Heather D. Wathington. 2012. "The Effects of Learning Communities for Students in Developmental Education: A Synthesis of Findings from Six Community Colleges." New York, NY: National Center for Postsecondary Research, Teachers College, Columbia University.

Wang, Ming-Te, Jacquelynne S. Eccles, and Sarah Kenny. 2013. “Not lack of ability but more choice individual and gender differences in choice of careers in science, technology, engineering, and mathematics.” *Psychological Science* 24(5): 770–775.

Werblow, Jacob, and Luke Duesbery. 2009. "The Impact of High School Size on Math Achievement and Dropout Rate." *The High School Journal* 92(3): 14-23.

Wortman, Camille B., and Jack W. Brehm. 1975. "Responses to Uncontrollable Outcomes: An Integration of Reactance Theory and the Learned Helplessness Model." Pp. 277-336 in *Advances in Experimental Social Psychology*, edited by L. Berkowitz. Orlando, FL: Academic Press.

Wu, Wei, Stephen G. West, and Jan N. Hughes. 2010. "Effect of Grade Retention in First Grade on Psychosocial Outcomes." *Journal of Educational Psychology* 102(1): 135-152.

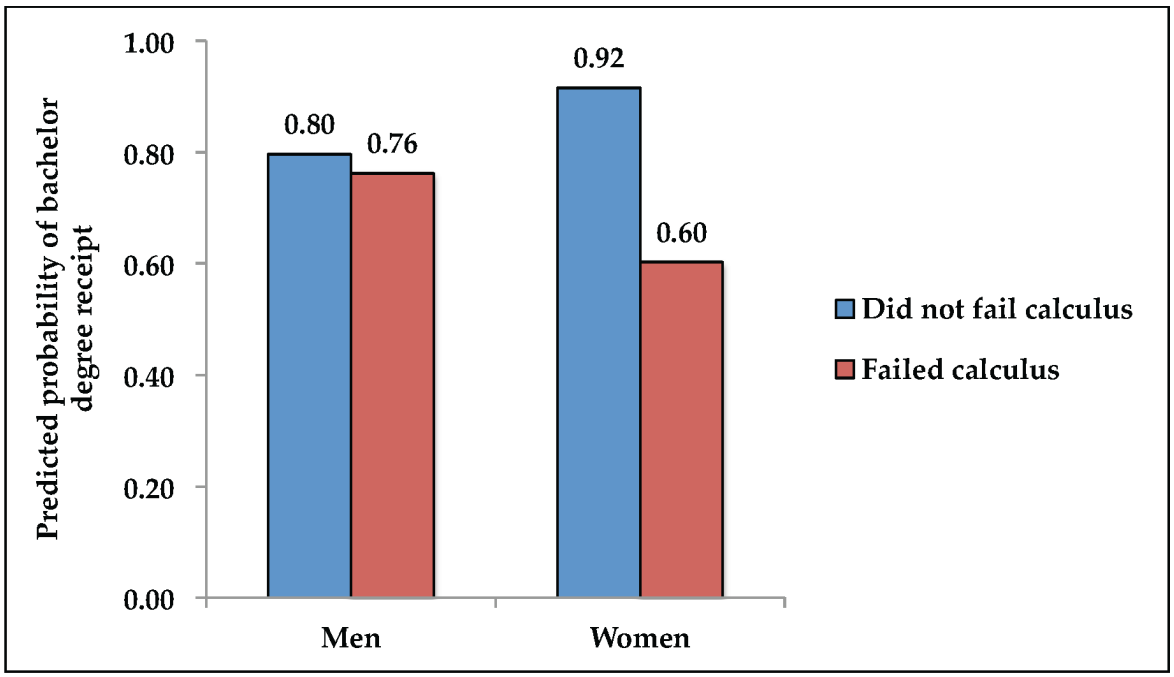
Xu, Di, Shanna S. Jaggars, Jeffrey Fletcher, and John E. Fink. 2018. "Are Community College Transfer Students ‘A Good Bet’ for 4-Year Admissions? Comparing Academic and Labor-Market Outcomes between Transfer and Native 4-year College Students." *Journal of Higher Education* 89(4): 478-502.

Xie, Yu, and Kimberlee A. Shauman. 2007. *Women in Science: Career Processes and Outcomes.* Cambridge, MA: Harvard University Press.

Zakariya, Sally Banks. 2015. "Learning to Read, Reading to Learn: Why Third-Grade is a Pivotal Year for Mastering Literacy." Center for Public Education: 1-15.

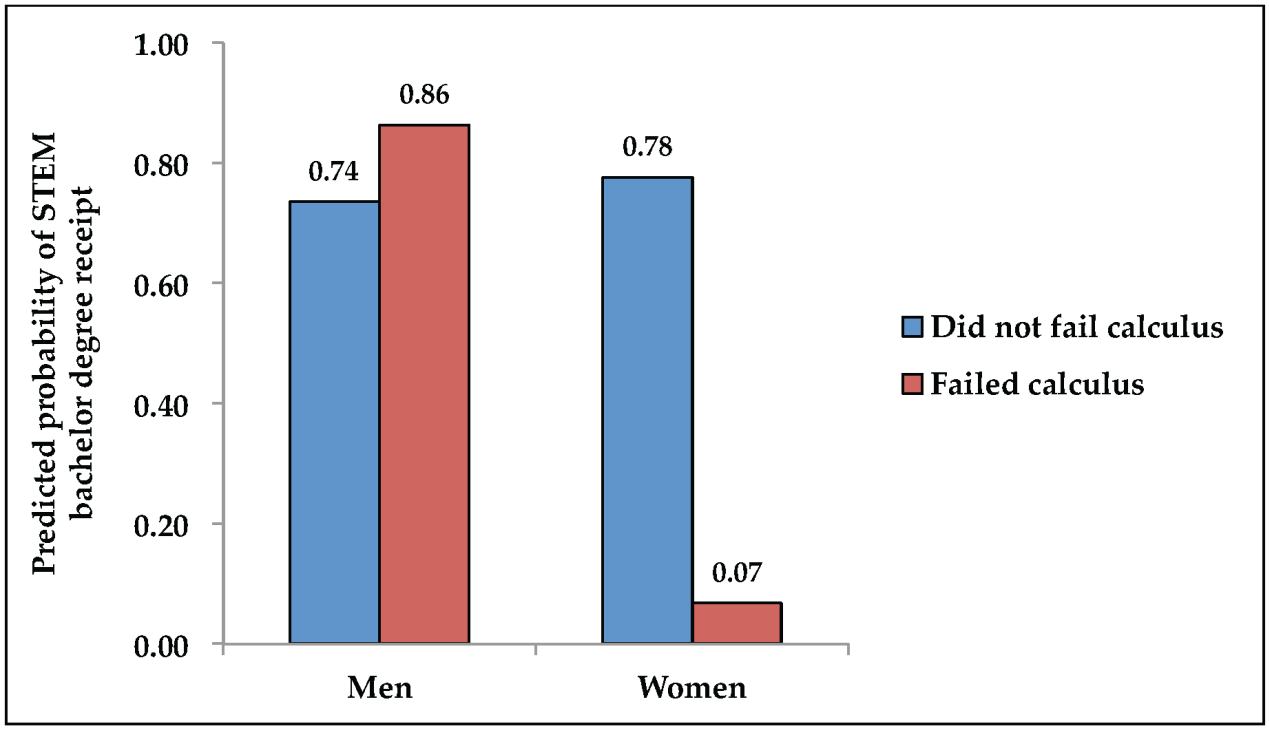
FIGURES

Figure 1.1 Predicted probabilities of bachelor degree receipt by gender.



Source: National Educational Longitudinal Study (NELS:88) and Postsecondary Education Transcript Study (PETS:2000) (NCES 1988; NCES 2000).

Figure 1.2 Predicted probabilities of bachelor degree receipt in a Science, Technology, Engineering, and Mathematics (STEM) field by gender



Source: National Educational Longitudinal Study (NELS:88) and Postsecondary Education Transcript Study (PETS:2000) (NCES 1988; NCES 2000).

Table 1.1. Descriptive statistics of variables used in analyses (n = 3650).

	Full Study Sample		Planned to Major in STEM		Did Not Plan to Major in STEM	
	# valid obs	mean/ %	# valid obs.	mean/ %	# valid obs.	mean/ %
	3650		910		2740	
Gender						
Male	1730	47.4%	450	49.5%	1280	46.7%
Female	1920	52.6%	460	50.5%	1460	53.32%
Race/Ethnicity						
White (Non-Hispanic)	2720	74.5%	640	70.5%	2080	75.8%
Black (Non-Hispanic)	270	7.5%	90	10.2%	180	6.6%
Hispanic	420	11.5%	100	10.6%	320	11.8%
Asian	240	6.6%	80	8.7%	160	5.8%
Age when entered college	3650	18.4	910	18.3	2740	18.4
Socioeconomic status (composite)	3650	0.08	910	0.04	2740	0.09
Prior Ability and Achievement						
NELS test score percentile	3650	60.6	910	62.6	2740	60.0
High School GPA	3650	2.89	910	2.98	2740	2.86
Highest Math Course Taken in High School						
Algebra I or equivalent	380	10.3%	80	8.%	300	11.0%
Geometry	480	13.2%	100	11.0%	380	13.7%
Algebra II	1250	34.2%	260	28.5%	990	36.1%
Trigonometry	550	15.1%	130	14.3%	420	15.3%
Pre-calculus	570	15.6%	170	18.7%	400	14.6%
Calculus	430	11.8%	180	19.8%	260	9.5%
Primary Institution Type						
Public 2 year	1380	37.8%	340	37.3%	1040	38.0%
Private Not-For Profit 4-year	640	17.5%	150	16.5%	490	17.9%
Public 4-year	1630	44.7%	430	47.3%	1200	43.8%
Planned to Major in STEM						
Did not plan to major in STEM	2740	75.1%	--	--	--	--
Planned to major in STEM	910	24.9%	--	--	--	--
Calculus Course						
Taken calculus	560	15.3%	250	27.5%	300	11.0%
Failed calculus	60	1.6%	40	4.4%	30	1.1%
Degree Attainment						
Earned a bacherlor‘s degree	1510	41.4%	360	39.6%	1150	42.0%
Did not earn a bacherlor‘s degree	1660	45.5%	400	52.7%	1260	46.0%
Earned a Bachelor‘s in STEM						
Did not earn bacherlor‘s degree in STEM	1190	32.6%	150	16.5%	1050	38.2%
Earned bacherlor‘s degree in STEM	470	12.9%	250	27.5%	220	8.0%

Source: National Educational Longitudinal Study (NELS:88) and Postsecondary Education Transcript Study (PETS:2000) (NCES 1988; NCES 2000). Sample restricted to students who had valid non-missing information on their postsecondary enrollment status, coursework, institution type, gender, race, age, NELS 12th grade test score percentile, high school GPA, highest math course taken in high school, and orientation towards majoring in a science, technology, engineering or mathematics (STEM) field in college. Degree attainment does not include students who earned an Associate’s Degree. n in models have been rounded to the nearest 10 for disclosure.

Table 1.2 Linear Probability Models (LPM) predicting who takes calculus and who fails calculus.

	Taken Calculus	Failed Calculus
	Compared to Students Who Never Took Calculus	**Only among Students Who Took Calculus**
Demographics		
Female	−0.11 ***	−0.02
	(−8.20)	(−0.44)
Age	−0.38	−0.76
	(0.11)	(−1.46)
Age squared	0.01	0.02
	(0.11)	(1.50)
Black	0.01	−0.01
	(0.68)	(−0.81)
Hispanic	0.01	0.01
	(0.60)	(0.15)
Asian	0.09 *	0.07
	(2.31)	(0.84)
Socio-economic status composite	0.02 *	−0.06 *
	(2.21)	(−2.11)
Prior academic skills and achievement		
NELS 12th grade test score percentile (logged)	0.04 ***	−0.03
	(4.69)	(−0.45)
High school GPA (logged)	0.11 ***	−0.17 +
	(4.27)	(−1.70)
Highest math course taken in High School		
Geometry	−0.03	−0.20
	(−1.64)	(−1.06)
Algebra II	−0.02 +	0.07
	(−1.76)	(−0.35)
Trigonometry	0.04 +	−0.07
	(1.90)	(−0.37)
Pre-calculus	0.10 ***	−0.11
	(3.77)	(−0.59)
Calculus	0.31 ***	−0.12
	(8.75)	(−0.65)
Planned to major in STEM	0.13 ***	0.03
	(7.23)	(0.74)
Institution Type		
Private not-for-profit 4-year	0.06 **	0.05
	(2.62)	(1.23)
Public 4-year	0.03 *	0.11 *
	(2.18)	(2.37)
Constant	3.37	7.53
	(7.23)	(1.55)
R^2	0.24	0.11
n	3490	540

Source: National Educational Longitudinal Study (NELS:88), Postsecondary Education Transcript Study (PETS:2000) (NCES 1988; NCES 2000). *t*-statistics underneath coefficients in parentheses. Controls are in reference to male, White, highest math course taken as Algebra I or other math course in high school, and entered a public two-year college. Sampling weight used in analyses. *n* in models have been rounded to the nearest 10 for disclosure. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 1.3 Linear Probability Models (LPM) predicting receipt of a bachelor's degree and receipt of a bachelor's degree in a STEM field, among students who had taken calculus and planned to major in STEM.

	Bachelor's Degree	**Bachelor's Degree**	**STEM Bachelor's**	**STEM Bachelor's**
Failed calculus	−0.12 $^{+}$		−0.12	
	(−1.66)		(−1.39)	
Gender and Failure Status (Omitted category: men—did not fail calculus)				
Men—failed calculus		−0.03		0.13
		(−0.34)		(1.30)
Women—did not fail calculus		0.12 $^{+}$		0.04
		(1.82)		(0.48)
Women—failed calculus		−0.19		−0.66 ***
		(−1.45)		(−7.40)
Constant	16.43	18.14	−52.37	−44.35
	(0.67)	(0.76)	(−1.52)	(−1.36)
R^2	0.25	0.27	0.31	0.42
n	230	230	190	190

Source: National Educational Longitudinal Study (NELS:88) and Postsecondary Education Transcript Study (PETS:2000) (NCES 1988; NCES 2000). STEM in reference to science, technology, engineering or mathematics fields. *t*-statistics underneath coefficients in parentheses. Reference category for interactions is a male college student who did not fail calculus. Includes demographic, prior achievement/academic skills, and institution controls for doubly robust estimates. *n* in models has been rounded to the nearest 10 for disclosure. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 2.1 *Descriptive statistics of variables used in analyses (n=3.646)*

	Two-Year College				Four-Year College			
	Total	**No R**	**Passed R**	**Failed R**	**Total**	**No R**	**Passed R**	**Failed R**
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
Gender								
Male	0.47	0.51	0.43	0.48	0.45	0.45	0.44	0.51
Female	0.53	0.49	0.57	0.52	0.55	0.55	0.56	0.49
Race/Ethnicity								
White	0.48	0.50	0.55	0.31	0.61	0.69	0.61	0.43
Black	0.25	0.23	0.19	0.39	0.21	0.13	0.22	0.41
Hispanic	0.22	0.23	0.21	0.24	0.14	0.13	0.13	0.16
Asian	0.02	0.01	0.02	0.02	0.03	0.03	0.03	0.00
Other	0.03	0.03	0.03	0.04	0.02	0.02	0.02	0.01
Age at entry (years)	19.8	20.3	19.5	19.7	18.7	18.9	18.5	18.9
Birth Cohort								
1980	0.20	0.20	0.19	0.21	0.17	0.16	0.18	0.20
1981	0.20	0.20	0.19	0.23	0.20	0.21	0.21	0.13
1982	0.20	0.21	0.22	0.15	0.20	0.20	0.19	0.25
1983	0.19	0.18	0.20	0.20	0.22	0.23	0.21	0.19
1984	0.21	0.22	0.20	0.22	0.21	0.20	0.21	0.23
Socioeconomic Status								
Mother's age at first birth	22.9	22.5	23.2	22.5	24.5	25.0	24.2	24.1
Parent's years of education	13.2	12.9	13.5	13.0	14.7	14.9	14.6	14.0
Household income (1997)	$44,650	$43,179	$47,700	$40,752	$62,964	$68,942	$61,460	$50,469
Income per capita	$10,670	$10,231	$11,463	$9,751	$15,582	$16,945	$15,354	$12,344
Academic Background								
ASVAB percentile score	44.9	44.8	49.9	34.9	65.3	71.0	64.8	50.0
High school GPA	2.76	2.71	2.88	2.58	3.18	3.28	3.21	2.84
Employment in College								
Employed students	0.92	0.89	0.95	0.93	0.93	0.92	0.93	0.95
Average hours per week	27.7	27.8	27.2	28.5	24.6	24.2	24.3	26.7
Degree Attainment								
No degree	0.75	0.87	0.60	0.83	0.35	0.34	0.26	0.66
Associate's degree	0.12	0.06	0.18	0.09	0.06	0.03	0.09	0.08
Bachelor's degree	0.13	0.06	0.22	0.08	0.59	0.63	0.65	0.26
Time to Bachelor's Degree								
Earned within 6 years	0.63	0.64	0.68	0.41	0.84	0.88	0.83	0.59
Earned after 6 years	0.37	0.37	0.33	0.59	0.16	0.12	0.18	0.41
Hourly Wage								
Total	$15.41	$16.47	$16.00	$12.22	$18.47	$18.75	$18.83	$16.26
No degree	$14.69	$15.69	$15.33	$11.87	$15.89	$15.01	$17.31	$15.47

BA degree	$18.76	$25.82	$17.98	$13.51	$19.75	$20.32	$19.40	$18.24
n	1,677	580	732	365	1,884	766	847	271

Source: National Longitudinal Survey of Youth (NLSY 1997), Postsecondary Transcript Study, 2011. R = remedial coursetaking. Sample restricted to students who had valid non-missing information on their postsecondary enrollment status and coursework. The sample sizes reported above reports information of the subgroups of students before multiple imputation, which will not add up to the total sample size used for analyses after multiple imputation. Age at entry and mother's age at first birth reported in years. Income per capita is household income divided by the number of residents within household. Average hours per week are reported hours the student worked per week. Hourly wage and degree category do not include students who obtained an associate's degree. Average wages are reported after student's last term in college, which do not include last term years from 2009 and onward.

Table 2.2. Linear Probability Models (LPM) predicting remedial coursetaking and performance among students who entered a two-year college

	Two Year College		
	Remediation (vs no remediation)	Passed Remediation (vs no remediation)	Failed Remediation (vs no remediation)
	Model 1	Model 2	Model 3
Demographics			
Female	0.05^{+}	0.06*	0.02
	(1.80)	(2.07)	(0.60)
Black	0.06^{+}	-0.01	0.19***
	(1.82)	(-0.31)	(4.33)
Hispanic	0.03	0.00	0.10*
	(0.86)	(0.08)	(2.24)
Asian	0.17*	0.15	0.32*
	(1.98)	(1.42)	(2.02)
Other	0.03	-0.03	0.15
	(0.45)	(-0.35)	(1.63)
Age at entry	-0.11^{+}	-0.15*	-0.02
	(-1.75)	(-2.29)	(-0.31)
Age squared	0.00	0.00	-0.01
	(1.40)	(2.03)	(-0.10)
Birth Cohort 1981	0.01	-0.01	0.04
	(0.05)	(-0.20)	(0.88)
Birth Cohort 1982	-0.03	-0.02	-0.05
	(-0.87)	(-0.48)	(-1.12)
Birth Cohort 1983	0.00	0.00	0.02
	(0.06)	(0.09)	(0.44)
Birth Cohort 1984	-0.02	-0.03	0.01
	(-0.48)	(-0.62)	(0.22)
Socio-economic Status			
Household income	0.01	0.01	0.00
	(0.26)	(0.35)	(0.03)
Income per capita	-0.01	-0.01	-0.01
	(-0.31)	(-0.35)	(-0.15)
Parent's years of education	0.01*	0.01*	0.01^{+}
	(2.47)	(2.29)	(1.69)
Mother's age at first birth	-0.01	-0.01	0.00
	(-0.30)	(-0.48)	(0.16)
Mother's age squared	0.00	0.00	-0.00
	(0.50)	(0.65)	(-0.04)
Academic Background			
ASVAB percentile score	-0.01	0.01	-0.07**
	(-0.82)	(0.44)	(-3.07)

High school G.P.A	0.01	0.03	-0.05*
	(0.33)	(1.60)	(-2.18)
Employment Characteristics			
Average hours worked	0.05	0.05	0.06
	(0.98)	(0.94)	(0.99)
Constant	1.65	2.11	0.29
	(2.25)	(2.56)	(0.35)
R^2	0.04	0.05	0.10
n	1722	1354	981

Source: National Longitudinal Survey of Youth (NLSY 1997), Postsecondary Transcript Study, 2011.
Note. T-statistics underneath coefficients in parentheses. Controls are in reference to male, White, and Birth Cohort 1980.
$^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 2.3. Linear Probability Models (LPM) predicting remedial coursetaking and performance among students who entered a four-year college

	Four Year College		
	Remediation (vs no remediation)	Passed Remediation (vs no remediation)	Failed Remediation (vs no remediation)
	Model 1	Model 2	Model 3
Demographics			
Female	-0.04	-0.02	-0.07**
	(-1.38)	(-0.69)	(-2.68)
Black	0.10**	0.09*	0.18***
	(2.92)	(2.24)	(4.04)
Hispanic	0.00	-0.01	0.03
	(0.01)	(-0.25)	(0.63)
Asian	-0.09	-0.05	-0.13***
	(-1.26)	(-0.66)	(-5.29)
Other	-0.04	-0.02	-0.01
	(-0.47)	(-0.25)	(-0.18)
Age at entry	-0.14	-0.13	-0.09
	(-1.63)	(-1.41)	(-1.07)
Age squared	0.00	0.00	0.00
	(1.20)	(0.99)	(0.80)
Birth Cohort 1981	-0.04	-0.02	-0.07+
	(-0.89)	(-0.47)	(-1.88)
Birth Cohort 1982	-0.01	-0.02	0.00
	(-0.31)	(-0.50)	(0.09)
Birth Cohort 1983	-0.05	-0.04	-0.05
	(-1.17)	(-1.03)	(-1.30)
Birth Cohort 1984	-0.02	-0.03	0.01
	(-0.56)	(-0.70)	(0.20)
Socio-economic Status			
Household income (logged)	-0.05	-0.04	-0.04
	(-1.38)	(-1.09)	(-1.19)
Income per capita in household (logged)	0.04	0.04	0.03
	(1.16)	(1.01)	(0.86)
Parent's years of education	-0.00	-0.00	0.00
	(-0.24)	(-0.28)	(0.12)
Mother's age at first birth	0.02	0.02	0.03
	(0.96)	(0.78)	(1.27)
Mother's age squared	-0.01	-0.00	-0.00
	(-1.21)	(-1.08)	(-1.20)
Academic Background			
ASVAB percentile score	-0.09***	-0.08***	-0.08***
	(-4.78)	(-3.86)	(-3.93)
High school GPA	-0.05**	-0.02	-0.11***
	(-2.93)	(-0.97)	(-5.67)
Employment Characteristics			

Average hours worker per week	-0.01	-0.03	0.06
	(-0.22)	(-0.68)	(1.51)
Constant	2.37	2.29	1.12
	(2.42)	(2.16)	(1.51)
R^2	0.07	0.05	0.21
n	1901	1629	1047

Source: National Longitudinal Survey of Youth (NLSY 1997), Postsecondary Transcript Study, 2011.
Note. T-statistics underneath coefficients in parentheses. Controls are in reference to male, White, and Birth Cohort 1980.
$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 2.4. Doubly robust estimates of outcomes associated with remedial coursetaking and failure for students who entered a two-year college

	Model 1	Model 2	Model 3
	Remediation (vs no remediation)	Passed Remediation (vs no remediation)	Failed Remediation (vs no remediation)
Panel A: Predicting bachelor's degree receipt			
Coefficient	0.09***	0.12***	0.05**
T-statistic	(5.80)	(6.99)	(2.98)
N	1515	1181	909
Panel B: Predicting bachelor's degree > 6 years (among degree receivers)			
Coefficient	-0.02	-0.00	0.29+
T-statistic	(-0.22)	(-0.02)	(1.85)
N	219	190	62
Panel C: Wage			
Coefficient	0.00	0.05	-0.14*
T-statistic	(0.07)	(0.94)	(-2.16)
N	1062	849	622
Panel D: Wage (controlling for degree receipt)			
Coefficient	-0.04	0.00	-0.16**
T-statistic	(-0.79)	(0.02)	(-2.67)
N	1062	849	622

Source: National Longitudinal Survey of Youth 1997, Postsecondary Transcript Study, 2011. T-statistics underneath coefficients in parentheses. Includes demographic and prior achievement/academic skills controls for doubly robust estimates. Models in Panel D also control for both associate's and bachelor's degree completion.
$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.5. Doubly robust estimates of outcomes associated with remedial coursetaking and failure for students who entered a four-year college

	Model 1	Model 2	Model 3
	Remediation (vs no remediation)	Passed Remediation (vs no remediation)	Failed Remediation (vs no remediation)
Panel A: Predicting bachelor's degree receipt			
Coefficient	0.02	0.08***	-0.25***
T-statistic	(0.60)	(3.92)	(-9.17)
N	1777	1527	999
Panel B: Predicting bachelor's degree > 6 years (among degree receivers)			
Coefficient	0.05	0.02	0.11**
T-statistic	(1.40)	(1.12)	(3.04)
N	1074	1005	541
Panel C: Wage			
Coefficient	0.03	0.01	-0.09+
T-statistic	(0.70)	(0.19)	(-1.95)
N	1413	1234	777
Panel D: Wage (controlling for degree receipt)			
Coefficient	0.03	-0.01	-0.04
T-statistic	(0.57)	(-0.09)	(-0.75)
N	1413	1234	777

Source: National Longitudinal Survey of Youth 1997, Postsecondary Transcript Study, 2011. T-statistics underneath coefficients in parentheses. Includes demographic and prior achievement/academic skills controls for doubly robust estimates. Models in Panel D also control for both associate's and bachelor's degree completion.
$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3.1

Description of Data Structure, by the Availability of Specific Outcomes

	Fourth grade outcomes			
Third grade cohort	OAKS Reading Scores	OAKS Math Scores	Attendance	Suspension
2004-2005	2005-2006	2005-2006	2005-2006	—
2005-2006	2006-2007	2006-2007	2006-2007	—
2006-2007	2007-2008	2007-2008	2007-2008	—
2007-2008	2008-2009	2008-2009	2008-2009	2008-2009
2008-2009	2009-2010	2009-2010	2009-2010	2009-2010
2009-2010	2010-2011	2010-2011	2010-2011	2010-2011

	Eighth grade outcomes			
Third grade cohort	OAKS Reading Scores	OAKS Math Scores	Attendance	Suspension
2004-2005	2009-2010	2009-2010	2009-2010	2009-2010
2005-2006	2010-2011	2010-2011	2010-2011	2010-2011
2006-2007	2011-2012	2011-2012	2011-2012	2011-2012
2007-2008	2012-2013	2012-2013	2012-2013	2012-2013
2008-2009	2013-2014	2013-2014	2013-2014	2013-2014
2009-2010	—	—	2014-2015	2014-2015

Table 3.2

Descriptive Statistics of Student Characteristics and Outcomes for All Students

	All third grade students
Characteristics	
Percentage White	0.678
Percentage Latinx	0.182
Percentage Black	0.029
Percentage Other	0.111
Percentage Female	0.493
Percentage FRL	0.501
Percentage ELL	0.098
Percentage special education	0.137
Third grade OAKS reading score	213.72
n	281,973
Fourth grade outcomes	
Reading OAKS score (percentile rank)	49.9
Math OAKS score (percentile rank)	49.8
Number of days absent	8.13
Ever suspended	0.028
n	267,682
Eighth grade outcomes	
Reading OAKS score (percentile rank)	49.9
Math OAKS score (percentile rank)	50.02
Number of days absent	10.08
Ever suspended	0.141
n	172,693

Note. Student data are from 2004-2005 to 2014-2015 academic year in Oregon. FRL = eligible for free or reduced price lunches, ELL = English language learner, OAKS = Oregon Assessment of Knowledge and Skills. Third grade OAKS reading score, while not centered here for interpretability, is centered at the cut-off in all analyses. Ever suspended in one year following the initial Grade 3 OAKS reading test only includes third grade students from spring 2008 through 2010 cohorts. The sample is limited to students who have non-missing reading OAKS scores in third grade, student demographic characteristics, and program participation.

Table 3.3

Regression Discontinuity (RD) Estimates for Earning the Negative Performance Label at Different Cutoffs for Students Scoring near each Cut Point, Covariate Balance

Independent Variable	Very Low/ Low	Low/ Nearly Meet the Standard	Nearly Meet / Meet the Standard	Meet / Exceed the Standard
Female	0.032	0.010	-0.004	-0.003
	(0.021)	(0.013)	(0.103)	(0.007)
	14,573	36,038	70,597	99,116
Latinx	-0.008	0.015	0.022*	0.006
	(0.018)	(0.011)	(0.009)	(0.005)
	14,573	36,038	70,597	99,116
Black	-0.001	-0.005	-0.001	0.003
	(0.009)	(0.005)	(0.004)	(0.002)
	14,573	36,038	70,597	99,116
Other	0.009	0.005	-0.014*	-0.004
	(0.012)	(0.008)	(0.006)	(0.004)
	14,573	36,038	70,597	99,116
Free or reduced price lunch	-0.005	-0.006	0.024**	0.005
	(0.015)	(0.011)	(0.009)	(0.006)
	14,573	36,038	70,597	99,116
English language learner	-0.006	0.007	0.023**	0.001
	(0.015)	(0.009)	(0.007)	(0.003)
	14,573	36,038	70,597	99,116
Special education	-0.008	-0.028**	0.018*	-0.001
	(0.020)	(0.011)	(0.008)	(0.004)
	14,573	36,038	70,597	99,116

Note. Each point estimate is from a separate RD regression where the baseline covariate is the dependent variable. Standard errors are reported below coefficient in parentheses.
$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 3.4

Estimated Effect of Earning the Negative Performance Label at Different Cutoffs on Fourth Grade Outcomes for Students Scoring near each Cut Point

	Very Low/ Low	Low/ Nearly Meet the Standard	Nearly Meet / Meet the Standard	Meet / Exceed the Standard
Panel A: Predicting Reading OAKS Score				
Coefficient	1.838*	0.415	-3.671***	-1.114***
Standard Error	(0.772)	(0.462)	(0.374)	(0.287)
n	7,473	24,650	54,714	86,342
Panel B: Predicting Math OAKS Score				
Coefficient	0.024	0.916	-3.194***	-1.443***
Standard Error	(1.178)	(0.638)	(0.494)	(0.330)
n	7,473	24,650	54,714	86,342
Panel C: Predicting Number of Absent Days in School District				
Coefficient	-0.631	0.188	0.294	0.053
Standard Error	(0.562)	(0.263)	(0.189)	(0.102)
n	7,473	24,650	54,714	86,342
Panel D: Predicting Likelihood of Suspension				
Coefficient	-0.009	-0.001	0.009+	0.002
Standard Error	(0.014)	(0.007)	(0.005)	(0.002)
n	7,473	24,650	54,714	86,342

Note. Student data are from 2004-2005 to 2014-2015 academic year in Oregon. Each cell entry includes parameter estimated, standard error (in parentheses), sample size, and approximate p-value. Estimated effects from a local linear regression-discontinuity model from Equation 1 using observations within half standard deviation bandwidth on either side of the cutoff, with the following control predictors: student race/ethnic group, free or reduced lunch participation, was formerly classified as English language learner, and school and academic year fixed effects. All models exclude students identified as special education in third grade. Reading and math OAKS score outcomes transformed into percentile ranks. Standard errors are reported below coefficient in parentheses. +, $p < 0.10$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

Table 3.5

Estimated Effect of Earning the Negative Performance Label at Different Cutoffs on Eighth Grade Outcomes for Students Scoring near each Cut Point

	Very Low/ Low	Low/ Nearly Meet the Standard	Nearly Meet the Standard/ Meet the Standard	Meet the Standard/ Exceed the Standard
Panel A: Predicting Reading OAKS Score				
Coefficient	2.200^{+}	1.049	-3.335***	-1.401***
Standard Error	(1.170)	(0.698)	(0.583)	(0.404)
n	4,987	15,340	33,609	55,310
Panel B: Predicting Math OAKS Score				
Coefficient	2.819^{+}	1.346	-3.133***	-1.567***
Standard Error	(1.504)	(0.862)	(0.682)	(0.442)
n	4,987	15,340	33,609	55,310
Panel C: Predicting Number of Absent Days in School District				
Coefficient	0.384	0.062	0.026	0.531**
Standard Error	(0.998)	(0.486)	(0.335)	(0.175)
n	4,987	15,340	33,609	55,310
Panel D: Predicting Likelihood of Suspension				
Coefficient	-0.077*	-0.011	0.017	0.002
Standard Error	(0.035)	(0.017)	(0.013)	(0.006)
	4,987	15,340	33,609	55,310

Note. Student data are from 2004-2005 to 2014-2015 academic year in Oregon. Each cell entry includes parameter estimated, standard error (in parentheses), sample size, and approximate p-value. Estimated effects from a local linear regression-discontinuity model from Equation 1 using observations within half bandwidth on either side of the cutoff, with the following control predictors: student race/ethnic group, free or reduced lunch participation, was formerly classified as English language learner, and school and academic year fixed effects. All models exclude students identified as special education in third grade. Reading and math OAKS score outcomes transformed into percentile ranks. Standard errors are reported below coefficient in parentheses. $^{+}$, $p < 0.10$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

Appendix A

Table A1. Coding for Expected Majors and Received Majors as STEM.

Planned to Major in STEM	Did Not Plan to Major in STEM
Architecture and Related Programs	Agricultural Business and Production
Biological and Life Sciences	Area, Ethnic and Cultural Studies
Computer and Information Sciences	Business Management
Engineering	Communications
Engineering Related Technologies	Education
Mathematics	Health Professions
Physical Sciences	Humanities
Science Technologies	Law
	Liberal Arts and Sciences
	Public Administration and Services
	Reserve Officers' Training Corp (R.O.T.C)
	Social Sciences
	Vocational Education
	Visual and Performing Arts

Appendix B

Doubly Robust Inverse Probability Weighting for Chapter 2

In the first step of doubly robust IPW, we estimate propensities (*P*) for each student. Using covariates discussed earlier, each student is given a propensity score. An individual variable does not have to be a statistically significant predictor of treatment in the propensity model since the objective is for students in the treated and control categories to be balanced on the covariates. The propensity score equation is a logit model predicting the probability of a student receiving an F in calculus. All individual-level and college-level covariates discussed above were included in the logistic regression equation to predict the probability of treatment:

$$Pr(Fail)_i = \alpha_i + \beta_k \boldsymbol{X}_{ki} + \varepsilon_i. \quad \text{A1}$$

Equation (A1) predicts the probability of a student failing calculus in college and $\mathbf{X}_i$ is a vector of control variables. In the model above, *i* represents the value of an individual in the predictor equation.

After estimating each student's predicted probability of failing calculus in Equation (A1), we then use the probabilities to create inverse probability weights, which we define as the inverse of the probability of receiving or not receiving the treatment given observable characteristics. For students at each category of treatment *t* (failed or passed calculus), we define our inverse probability weight as:

$$W = 1/\hat{P}_t, \quad \text{A2}$$

where $\hat{P}_t$ is the predicted probability that a student received the treatment that he or she received.

For doubly robust IPW estimators, the same covariates used to estimate the probability weights for Equation (A1) are also included as controls in a linear probability model predicting

our degree outcomes. To examine whether the relationship between failing calculus and degree outcomes vary by gender, we estimate models that interact failing calculus with gender. We estimate two sets of these models; the first set predicts bachelor degree completion in any field and the second set predicts STEM bachelor degree completion. Thus, our first model in Table 3 predicts whether students completing a bachelor's degree in any field as a function of failing calculus:

$$Pr(Degree)_i = \alpha_i + \beta_1 Fail_i + \beta_k \boldsymbol{X}_{ki} + \varepsilon_i \,, \qquad \text{A3}$$

where $Fail_i$ is a dummy variable equal to one if a student ever failed calculus and zero otherwise and $\mathbf{X}_i$ is a vector of background controls for doubly robust estimates. The main effect of $Fail_i$ provides information about the association between failing calculus and receiving a bachelor's degree. In the next model, we include an interaction effect between $Fail_i$ and whether the student was female to examine the association any variation between failing calculus and bachelor degree completion by gender. The error term, ε_i, captures characteristics not accounted for in the model that influences the outcome variable. We estimate similar models predicting STEM bachelor degree receipt.

Appendix C

Doubly Robust Inverse Probability Weighting for Chapter 3

In the first step of doubly robust IPW, we estimate treatment propensities (P) for each student. Using covariates discussed in the paper, a propensity score is estimated for each student. An individual variable does not have to be a statistically significant predictor of treatment to be included in the propensity model since the objective is for students in the treated and control categories to be balanced on the covariates. The propensity scores are estimated using a multivariate logistic regression model predicting the probability of a student receiving the treatment (i.e., not taking a remedial course, taking and passing a remedial course, and taking and failing a remedial course). All covariates discussed in the paper were included in the multiple logistic regression equation to predict the probability of treatment:

$$Pr(Remedial\ Group)_i = \alpha_i + \beta_k \boldsymbol{X}_{ki} + \varepsilon_i \quad \text{(A1)}$$

Equation (A1) predicts the probability of student *i* being in one of three groups: never took a remedial course, took and passed remedial coursework, and took and failed a remedial course. $\boldsymbol{X}_{ki}$ is a vector of control variables.

We estimate each student's predicted probability of being in each of the remedial groups in Equation (A1), and use these probabilities to create inverse probability weights. For each treatment category *t* (never took remediation, took and passed remediation, or took and failed remediation), we define our inverse probability weight as:

$$W = 1/\hat{P}_t \quad \text{(A2)}$$

where $\hat{P}_t$ is the predicted probability that a student received the treatment that he or she received.

For doubly robust IPW estimators, the same covariates used to estimate the probability weights for Equation (A1) are also included as controls in a linear probability model predicting

our degree and wage outcomes. We estimate two sets of these models; the first set predicts bachelor's degree completion in any field and the second set predicts the average wage outcomes. Thus, our first model predicts whether students complete a bachelor's degree as a function of being in one of the three remedial groups: never took remediation, took and passed remediation, and took and failed remediation:

$$Pr(Bachelors)_i = \alpha_i + \beta_1 Remedial_i + \beta_k \boldsymbol{X}_{ki} + \varepsilon_i \qquad \text{(A3)}$$

where $Remedial_i$ is a dummy variable equal to one if a student ever took remediation and zero otherwise and $\mathbf{X}_i$ is a vector of background controls for doubly robust estimates. To estimate the relationship between failing remedial coursework and our other outcomes, we estimate additional models that take the same general form as (A3), but instead of $Remedial_i$, we use a dummy variable equal to one if a student took and passed their remedial coursework and zero otherwise, or alternatively a dummy variable equal to one if a student took and failed their remedial coursework and zero otherwise. The error term, ε_i, captures characteristics not accounted in the model that influence the outcome variable. We estimate these models separately for students who entered a two-year and four-year college and we use similar models to predict average post-college wages for the latest five years (2007 through 2011).

Table C1. Linear Probability Models (LPM) predicting failure among remedial course takers

	Two Year College	Four Year College
	Failed Remediation (vs passed remediation)	Failed Remediation (vs passed remediation)
Demographics		
Female	-0.04	-0.04
	(-1.54)	(-1.48)
Black	0.19***	0.08*
	(4.29)	(2.03)
Hispanic	0.08+	0.04
	(1.80)	(0.91)
Asian	0.10	-0.19***
	(0.89)	(-5.60)
Other	0.14+	-0.08
	(1.72)	(-1.01)
Age at entry	0.11+	-0.03
	(1.74)	(-0.31)
Age squared	-0.01+	0.00
	(-1.94)	(0.29)
Birth Cohort 1981	0.07+	-0.06+
	(1.67)	(-1.70)
Birth Cohort 1982	-0.03	0.04
	(-0.74)	(0.93)
Birth Cohort 1983	-0.01	-0.01
	(-0.01)	(-0.21)
Birth Cohort 1984	0.04	0.02
	(0.04)	(0.60)
Socio-economic Status		
Household income	-0.01	-0.03
	(-0.40)	(-0.73)
Income per capita	0.01	0.02
	(0.31)	(0.44)
Parent's years of education	-0.01	-0.00
	(-0.61)	(-0.09)
Mother's age at first birth	0.02	-0.00
	(0.63)	(-0.09)
Mother's age squared	-0.01	0.00
	(-0.62)	(0.34)
Academic Background		
ASVAB percentile score	-0.07***	-0.06**
	(-3.34)	(-2.91)
High school G.P.A	-0.08***	-0.11***
	(-4.14)	(-6.19)
Employment Characteristics		
Average hours worked	0.03	0.07*
	(0.73)	(1.99)

Constant	-1.13	0.47
	(-1.39)	(0.46)
R^2	0.13	0.14
n	1109	1121

Source: National Longitudinal Survey of Youth (NLSY 1997), Postsecondary Transcript Study, 2011.
Note. T-statistics underneath coefficients in parentheses. Controls are in reference to male, White, and Birth Cohort 1980.
$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table C2. Doubly robust estimates of outcomes associated with remedial course failure relative to passing remedial coursework

	Two Year College	Four Year College
	Failed Remediation (vs passed remediation)	Failed Remediation (vs passed remediation)
Panel A: Predicting bachelor's degree receipt		
Coefficient	-0.31***	-0.11***
T-statistic	(-10.64)	(-4.92)
N	1025	940
Panel B: Predicting bachelor's degree > 6 years (among degree receivers)		
Coefficient	0.24***	0.27***
T-statistic	(5.41)	(3.39)
N	602	186
Panel C: Wage		
Coefficient	-0.10*	-0.13**
T-statistic	(-2.23)	(-2.64)
N	814	652
Panel D: Wage (controlling for degree receipt)		
Coefficient	-0.05	-0.11*
T-statistic	(-0.94)	(-2.13)
N	814	652

Source: National Longitudinal Survey of Youth 1997, Postsecondary Transcript Study, 2011. T-statistics underneath coefficients in parentheses. Includes demographic and prior achievement/academic skills controls for doubly robust estimates. Models also control for the total number of remedial courses taken.
$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix D

Grade 3 Reading OAKS Sample Test

Reading and Literature ▼

DIRECTIONS

Read each of the passages. Then read the questions that follow and decide on the BEST answer. There are a lot of different kinds of questions, so read each question carefully before marking an answer on your answer sheet.

THE SQUIRREL'S LOAN

This story of a magpie, a kind of bird, and a squirrel has a great lesson to teach.

THE MAGPIE AND THE SQUIRREL LIVED on the lower slopes of the Himalayas. One cold and snowy winter, the magpie borrowed some nuts from the squirrel, and the squirrel borrowed some feathers to warm his hole in the tree.

In summer, the magpie brought some nuts to return the loan, but the squirrel said, "It's summer and I have plenty of nuts now. You took them from me in winter, so return them to me in winter."

The magpie wondered what he would do because he knew there would be no nuts to be found in winter. So when the squirrel came to return the loan of the feathers, he said, "I have plenty of feathers in my nest now. You took them from me in winter, so return them to me in winter."

"Very well, " said the squirrel, and he stored the feathers along with his horde of nuts.

But it was a hot summer, and the squirrel's house felt like a furnace with all those feathers in it. So he threw the feathers out, thinking he'd pick them up when winter came around and it was time to return the loan.

Reading and Literature ▼

In winter, there was ice and snow everywhere. The feathers were buried underneath. Try as he might, the squirrel could not dig them out.

He said to the magpie, "I'm afraid I can't find feathers in winter."

"Nor can I find nuts at this time," said the magpie.

And the squirrel remembered his loan and his words to the magpie, and he was ashamed. He said, "I should expect you to return the nuts when you can, not when you cannot. A loan is meant to help a friend, not to give him trouble."

From then on they helped each other in winter and repaid their debts in summer. They continued to live happily and became even better friends thereafter.

1

Squirrel added the feathers to his horde. In this story, a synonym or word with the same meaning as *horde* would be

A. friends.
B. food.
C. gifts.
D. collection.

2

Squirrel was ashamed. Which word below best describes *ashamed*?

A. Embarrassed
B. Sarcastic
C. Confused
D. Angry

3

Magpie and Squirrel

A. helped each other.
B. decided not to be friends.
C. argued constantly.
D. liked to play tricks on each other.

4

The story is mostly telling about

A. how to live in winter.
B. cooperation between friends.
C. winters in the Himalayas.
D. paying back debts.

5

What do you think will happen the next winter?

A. Squirrel and Magpie will move to warmer ground.
B. Squirrel and Magpie will not borrow anything.
C. Squirrel and Magpie will not borrow from others.
D. Squirrel and Magpie will loan nuts and feathers again.

(Oregon Department of Education 2008)

Appendix E

Back of the Envelope Calculations for analyses in Chapter 4

For the analyses presented in Chapter 4, I translate the effect sizes to months based on calculations presented in Baird and Pane's study (2018). In their conception, the standardized posttest score for student $\boldsymbol{i}$ can be modeled as a function of treatment status $\boldsymbol{T_i}$, standardized pretest score $\boldsymbol{w_i}$ (if available), a vector of covariates and baseline factors $\boldsymbol{X_i}$, and unobserved factors $\boldsymbol{\varepsilon_i}$:

$$\boldsymbol{z_i = \alpha + \beta T_i + \lambda w_i + X_i \gamma + \varepsilon_i} \quad (1)$$

In the model above, $\boldsymbol{\beta}$ is the standardized treatment effect. Then, the standardized achievement is a function of elapsed time, extending equation $\boldsymbol{D_i}$, the fraction of the school year that has passed:

$$\boldsymbol{z_i = (\alpha + \beta T_i) D_i + \lambda w_i + \gamma X_i + \varepsilon_i} \quad (2)$$

Setting $\boldsymbol{D_i}$ to one (a full year) recovers equation 1. Although the authors note that it is highly unlikely that learning rates are constant across the year, the model above makes simplifying assumptions that achievement accumulates in a linear fashion after controlling for covariates, and that any incremental growth due to a treatment effect (e.g. performance label) also accumulates linearly with time.

The years of learning translation estimates the additional fraction of a year, which must be added to the schooling time of an untreated student (e.g. more positive label) to make their achievement equal to a treated student (e.g. less positive label) who received one year of schooling. Mathematically, it equates it in this way:

$$\boldsymbol{E\{z_i | T_i = 0, D_i = 1 + \phi\} = E\{z_i | T_i = 1, D_i = 1\}} \quad (3)$$

$$\boldsymbol{\rightarrow \alpha(1 + \phi) = \alpha + \beta}$$

Solving for $\boldsymbol{\phi}$,

$$\boldsymbol{\phi} = \frac{\boldsymbol{\beta}}{\boldsymbol{\alpha}} \tag{4}$$

β is the standardized effect size estimated by the evaluation; α is a measure of typical annual growth and can be estimated directly from the regression with the appropriate standardized coefficient by using an external estimate of typical growth.

Appendix F

Robustness Checks for Analyses in Chapter 4

The internal validity of a regression-discontinuity (RD) depends on several assumptions. First, the "treatment" – receiving a particular performance label assigned to students immediately above or below the cut-off – must be assigned exogenously by placing the student according to the cut-off and applied to all students equally. Second, students must not be able to manipulate their position on the "forcing variable" (e.g. third grade reading OAKS score) relative to the threshold. In order for student characteristics to be a smooth function of the forcing variable near the cut-off score, then these conditions must be met.

In this study, these conditions are met because the cut-off scores differ across two specific time points: 2006-2007 and 2008 through 2010, and these cut-off scores were changed after students had taken the test. It would be implausible for students at the margin of receiving a particular label to manipulate their position relative to the distribution of scores while taking the reading assessment in third grade. One of the ways we test this is through examining whether the cut-offs are discontinuous across student demographic characteristics. We show in Table 3.3, that while there are discontinuities for some demographic characteristics, we control for these demographic characteristics to ensure covariate balance.

We also conducted exploratory analyses to check for smoothness in the relationship between observed student characteristics and the forcing variable near the cut-off. We summarize these analyses in a single test, as suggested by Lee and Lemieux (2010) and Papay, Murnane and Willett (2016). This single test is a set of seemingly unrelated regression (SUR) models, each of which contains a local linear RD model from Equation 1 but with a different covariate treated as the outcome. We then test whether these coefficients on the discontinuity

term are equal to zero across all covariates. This procedure was executed four times for each cut-off. Referring to Table F1, we fail to reject the null hypothesis across all cut-offs except the *Nearly Meet the Standard/Meet the Standard* threshold. While the other cut-offs suggests that the state has imposed the cut score exogenously, the cut-off for *Nearly Meet the Standard/Meet the Standard* should be investigated further.

Another key assumption for the RD approach is that we have specified the relationship between the outcome and the forcing variable (third-grade reading OAKS score) correctly, by focusing our analyses within a narrow bandwidth around the cut-scores and using a local linear regression approach. The key decision for these analyses is choosing the bandwidth, which controls the smoothing function. To test the sensitivity of our findings to bandwidth choice, we run the same local linear regression models across a range of bandwidths: full sample, 1 standard deviation bandwidth, 0.5 standard deviation bandwidth, and 0.2 standard deviation bandwidth. In Tables F2 and F3, we present the estimated causal effects for each of our outcomes. While the magnitudes of a few individual estimates are sensitive to these choices, we see a general pattern persist across a wide range of bandwidths. Because of the smaller sample size for the smallest bandwidth choice (0.2 standard deviation), these estimates do not reach traditional levels of statistical significant but the general pattern for the cut-off scores remain.

One concern is that student scores are reported in whole numbers, leading to large clusters of students with reading scores at even-integer values (e.g. 200 and 201, rather than 200.5). Prior research shows that the results from using RD can be biased by "heaping" of the assignment variable (Barreca, Guldi, Lindo, and Waddell 2011). We perform the McCrary (2008) test - a statistical test to assess whether there is a discontinuity in the density of observations at the cut-off score. We can see in Figures F1 through F4 that all cut-offs but one

are robust to the concerns about endogenous sorting. However, we see in Figure F2 that there is a visible discontinuity in the threshold for *Nearly Meet the Standard/Meet the Standard*. This discontinuity will be investigated further.

We present an additional robustness check in Tables F4 and F5 to examine whether our results are being driven by the prevalence of whole-integer reading scores by excluding students with specific values. We model "donut RDs" by excluding students with scores exactly at the cut-off for that particular year. For each of the student outcomes, the point estimates presented in Table F4 and F5 are from individual regressions for the variable third-grade reading score, similar to the point estimates shown in Tables 4 and 5 in Chapter 4. The results in Tables F4 and F5 show that our inferences are robust in specifications that exclude students whose third-grade reading score fell on the heaped value of 0 (the cut-off for that particular year). While the magnitude of the estimates are higher without heaping, we can see that our estimates remain largely robust across the "donut RD" models, however, some estimates become statistically significant without heaping. These models will be examined in future work to determine whether heaping is occurring near the thresholds of interest.

Finally, we provide scatter plots of the mean values of each outcome by each value of the forcing variable (Figures F5-F12), with a solid line denoting the threshold point for each cut-off score examined in the paper. We also provide an additional descriptive table (Table F6) containing the standard deviation for each outcome variable for further information.

Table F1

Results from the Hypothesis Test that the Disruption in each Observed Covariate is Zero in the Population at each of the Four Cut-Scores, from a Seemingly Unrelated Regression (SUR) Regression-Discontinuity Model where each Covariate is Treated as an Outcome for Fully Study Sample and Students Within Half Bandwidth of Cut-Score

	Half bandwidth
Meet/Exceed the Standard cutoff	$\chi^2(7) = 10.33$ $p = 0.171$
Nearly Meet/Meet the Standard cutoff	$\chi^2(7) = 62.57$ $p = 0.000$
Low/Nearly Meet the Standard cutoff	$\chi^2(7) = 11.65$ $p = 0.113$
Very Low/Low the Standard cutoff	$\chi^2(7) = 6.25$ $p = 0.511$

Notes: All inferences from two-tailed hypothesis tests. Covariates treated as outcomes include race, gender, and participation in free or reduced lunch, English language proficiency status, and special education status.

Table F2

Regression Discontinuity Estimates, Academic and Behavioral Fourth Grade Outcomes in Third Grade Sample at Different Cut-Offs, by Bandwidth

Fourth Grade Outcomes	Bandwidth Sample			
	Full Sample	± 1.0	± 0.5	± 0.2
Meet the Standard/Exceed the Standard Cut-Off				
Reading OAKS Score	-8.986***	-2.685***	-1.114***	-1.56**
	(0.138)	(0.186)	(0.287)	(0.570)
Math OAKS Score	-5.656***	-2.081***	-1.443***	-2.078**
	(0.154)	(0.213)	(0.330)	(0.664)
Absent Days	-0.108*	0.011	0.053	0.055
	(0.048)	(0.066)	(0.102)	(0.203)
Ever Suspended	0.002+	0.002	0.002	0.005
	(0.001)	(0.001)	(0.002)	(0.004)
Sample size	231,574	165,052	86,342	42,548
Nearly Meet the Standard/Meet the Standard Cut-Off				
Reading OAKS Score	-6.435***	-3.104***	-3.671***	-4.829***
	(0.197)	(0.250)	(0.374)	(0.796)
Math OAKS Score	-4.547***	-2.885***	-3.194***	-3.265**
	(0.250)	(0.325)	(0.494)	(1.075)
Absent Days	0.246**	0.204+	0.294	0.124
	(0.095)	(0.122)	(0.189)	(0.378)
Ever Suspended	0.009***	0.007*	0.009+	0.010
	(0.002)	(0.003)	(0.005)	(0.010)
Sample size	231,574	112,349	54,714	24,862
Low/Nearly Meet the Standard Cut-Off				
Reading OAKS Score	-2.152***	0.122	0.415	-0.774
	(0.241)	(0.299)	(0.462)	(0.962)

Math OAKS Score	-1.578***	0.389	0.916	0.870
	(0.320)	(0.413)	(0.638)	(1.37)
Absent Days	0.477***	0.065	0.188	-0.463
	(0.130)	(0.165)	(0.263)	(0.587)
Ever Suspended	0.010**	0.002	-0.001	-0.028+
	(0.003)	(0.005)	(0.007)	(0.015)
Sample size	231,574	73,360	24,650	8,892
Very Low/Low Cut-Off				
Reading OAKS Score	6.978***	1.999***	1.838*	2.954+
	(0.455)	(0.533)	(0.772)	(1.618)
Math OAKS Score	3.946***	0.504	0.024	-0.661
	(0.581)	(0.751)	(1.178)	(2.53)
Absent Days	1.135***	-0.146	-0.631	-0.127
	(0.286)	(0.372)	(0.562)	(1.209)
Ever Suspended	0.013*	0.004	-0.009	-0.042
	(0.007)	(0.009)	(0.014)	(0.031)
Sample size	231,574	19,178	7,473	3,320

Note. Student data are from 2004-2005 to 2014-2015 academic year in Oregon. Each cell entry includes parameter estimated, standard error (in parentheses), sample size, and approximate p-value. Estimated effects from a local linear regression-discontinuity model from Equation 1 using observations within different bandwidth samples on either side of the cutoff, with the following control predictors: student race/ethnic group, free or reduced lunch participation, was formerly classified as English language learner, and school and academic year fixed effects. All models exclude students identified as special education in third grade. Reading and math OAKS scores transformed into percentile ranks. Standard errors are reported below coefficient in parentheses. +, $p < 0.10$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

Table F3

Regression Discontinuity Estimates, Academic and Behavioral Fourth Grade Outcomes in Third Grade Sample at Different Cut-Offs, by Bandwidth

Eighth Grade Outcomes	Bandwidth Sample			
	Full Sample	± 1.0	± 0.5	± 0.2
Meet the Standard/Exceed the Standard Cut-Off				
Reading OAKS Score	-7.010***	-2.001***	-1.401***	-1.813*
	(0.182)	(0.258)	(0.404)	(0.806)
Math OAKS Score	-4.689***	-1.745***	-1.567***	-2.777***
	(0.197)	(0.280)	(0.442)	(0.871)
Absent Days	-0.007	0.166	0.531**	0.732*
	(0.081)	(0.113)	(0.175)	(0.362)
Ever Suspended	0.011***	0.002	0.002	0.003
	(0.003)	(0.004)	(0.006)	(0.012)
Sample size	148,660	106,059	55,310	27,794
Nearly Meet the Standard/Meet the Standard Cut-Off				
Reading OAKS Score	-5.497***	-3.124***	-3.335***	-3.103*
	(.287)	(0.370)	(0.583)	(1.270)
Math OAKS Score	-4.369***	-2.901***	-3.133***	-0.126
	(0.324)	(0.429)	(0.682)	(1.494)
Absent Days	0.296+	-0.112	0.026	-1.018
	(0.164)	(0.224)	(0.335)	(0.717)
Ever Suspended	0.022***	0.015+	0.017	-0.011
	(0.006)	(0.008)	(0.013)	(0.028)
Sample size	148,660	70,047	33,609	15,062
Low/Nearly Meet the Standard Cut-Off				
Reading OAKS Score	-1.421***	0.770+	1.049	-1.441
	(0.349)	(0.432)	(0.698)	(1.570)

Math OAKS Score	-1.258**	0.691	1.346	0.417
	(0.397)	(0.531)	(0.862)	(1.918)
Absent Days	0.615**	0.226	0.062	1.569
	(0.226)	(0.303)	(0.486)	(1.153)
Ever Suspended	0.004	-0.012	-0.011	0.054
	(0.008)	(0.011)	(0.017)	(0.038)
Sample size	148,660	45,047	15,340	5,429
Very Low/Low Cut-Off				
Reading OAKS Score	6.189***	2.018***	2.200+	-1.140
	(0.618)	(0.756)	(1.170)	(2.778)
Math OAKS Score	4.564***	3.219**	2.819+	0.476
	(0.738)	(0.966)	(1.504)	(3.237)
Absent Days	0.676	-0.539	0.384	-0.524
	(0.462)	(0.619)	(0.998)	(2.272)
Ever Suspended	0.008	-0.010	-0.077*	-0.110
	(0.016)	(0.022)	(0.035)	(0.070)
Sample size	148,660	12,793	4,987	2,191

Table F4

Regression Discontinuity Estimates, Academic and Behavioral Outcomes in Third Grade Sample With and Without Heaping, Fourth Grade Outcomes

	With Heaping				Without Heaping			
	Very Low/ Low	Low/ Nearly Meet the Standard	Nearly Meet the Standard / Meet the Standard	Meet the Standard/ Exceed the Standard	Very Low/ Low	Low/ Nearly Meet the Standard	Nearly Meet the Standard / Meet the Standard	Meet the Standard / Exceed the Standard
Panel A: Predicting Reading OAKS Score								
Coefficient	1.838*	0.415	-3.671***	-1.114***	1.998*	1.342**	-3.754***	-0.800*
Standard Error	(0.772)	(0.462)	(0.374)	(0.287)	(0.856)	(0.512)	(0.408)	(0.328)
n	7,473	24,650	54,714	86,342	6,861	23,206	47,936	78,408
Panel B: Predicting Math OAKS Score								
Coefficient	0.024	0.916	-3.194***	-1.443***	0.562	1.518*	-3.267***	-1.038**
Standard Error	(1.178)	(0.638)	(0.494)	(0.330)	(1.318)	(0.700)	(0.543)	(0.379)
n	7,473	24,650	54,714	86,342	6,861	23,206	47,936	78,408
Panel C: Predicting Number of Absent Days in School District								
Coefficient	-0.631	0.188	0.294	0.053	-0.724	0.271	0.529**	0.016
Standard Error	(0.562)	(0.263)	(0.189)	(0.102)	(0.624)	(0.286)	(0.199)	(0.116)
n	7,473	24,650	54,714	86,342	6,861	23,206	47,936	78,408
Panel D: Predicting Likelihood of Suspension								
Coefficient	-0.009	-0.001	0.009^{+}	0.002	-0.006	0.002	0.011*	0.000
Standard Error	(0.014)	(0.007)	(0.005)	(0.002)	(0.015)	(0.007)	(0.005)	(0.000)
n	7,473	24,650	54,714	86,342	6,861	23,206	47,936	78,408

Note. Student data are from 2004-2005 to 2014-2015 academic year in Oregon. Each cell entry includes parameter estimated, standard error (in parentheses), sample size, and approximate p-value. Estimated effects from a local linear regression-discontinuity model from Equation 1 using observations within half bandwidth on either side of the cutoff, with the following control predictors: student race/ethnic group, free or reduced lunch participation, was formerly classified as English language learner, and school and academic year fixed effects. All models exclude students identified as special education in third grade. Reading and math OAKS score outcomes transformed into percentile ranks. Standard errors are reported below coefficient in parentheses. $^{+}$, $p < 0.10$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

Table F5

Regression Discontinuity Estimates, Academic and Behavioral Outcomes in Third Grade Sample With and Without Heaping, Eighth Grade Outcomes

	With Heaping				Without Heaping			
	Very Low/ Low	Low/ Nearly Meet the Standard	Nearly Meet the Standard / Meet the Standard	Meet the Standard/ Exceed the Standard	Very Low/ Low	Low/ Nearly Meet the Standard	Nearly Meet the Standard / Meet the Standard	Meet the Standard / Exceed the Standard
Panel A: Predicting Reading OAKS Score								
Coefficient	2.200^{+}	1.049	-3.335***	-1.401***	2.31^{+}	1.749*	-3.470***	-1.006*
Standard Error	(1.170)	(0.698)	(0.583)	(0.404)	(1.298)	(0.770)	(0.648)	(0.467)
n	4,987	15,340	33,609	55,310	4,589	14,493	29,320	50,523
Panel B: Predicting Math OAKS Score								
Coefficient	2.819^{+}	1.346	-3.133***	-1.567***	3.270^{+}	2.284*	-3.207***	-1.230*
Standard Error	(1.504)	(0.862)	(0.682)	(0.442)	(1.675)	(0.943)	(0.737)	(0.500)
n	4,987	15,340	33,609	55,310	4,589	14,493	29,320	50,523
Panel C: Predicting Number of Absent Days in School District								
Coefficient	0.384	0.062	0.026	0.531**	0.402	-0.167	0.187	0.493*
Standard Error	(0.998)	(0.486)	(0.335)	(0.175)	(1.079)	(0.530)	(0.364)	(0.201)
n	4,987	15,340	33,609	55,310	4,589	14,493	29,320	50,523
Panel D: Predicting Likelihood of Suspension								
Coefficient	-0.077*	-0.011	0.017	0.002	-0.053	-0.011	0.020	0.001
Standard Error	(0.035)	(0.017)	(0.013)	(0.006)	(0.038)	(0.019)	(0.014)	(0.007)
n	4,987	15,340	33,609	55,310	4,589	14,493	29,320	50,523

Note. Student data are from 2004-2005 to 2014-2015 academic year in Oregon. Each cell entry includes parameter estimated, standard error (in parentheses), sample size, and approximate p-value. Estimated effects from a local linear regression-discontinuity model from Equation 1 using observations within half bandwidth on either side of the cutoff, with the following control predictors: student race/ethnic group, free or reduced lunch participation, was formerly classified as English language learner, and school and academic year fixed effects. All models exclude students identified as special education in third grade. Reading and math OAKS score outcomes transformed into percentile ranks. Standard errors are reported below coefficient in parentheses. $^{+}$, $p < 0.10$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

Table F6

Descriptive Statistics for Fourth Grade and Eighth Grade Outcomes, by Each Cut-Off for Performance Labels

A. Full sample of third grade students

	All third grade students	
	mean	**std dev**
4th grade outcomes		
Read score	49.9	28.6
Math score	49.8	28.7
Number of days absent	8.13	7.21
Ever suspended	0.028	0.17
n	267,682	267,682
8th grade outcomes		
Read score	49.9	28.6
Math score	50.02	28.6
Number of days absent	10.08	9.9
Ever suspended	0.141	0.35
n	172,693	172,693

Note. Student data are from 2004-2005 to 2014-2015 academic year in Oregon.

B. Third grade students in "Exceed the Standard" group

	Exceed standard (6 points above)		Exceed standard (6 points below)	
	mean	**std dev**	**mean**	**std dev**
4th grade outcomes				
Read score	65.36	20.25	49.86	20.93
Math score	61.53	23.75	49.50	24.27
Number of days absent	7.72	6.80	8.12	7.13
Ever suspended	0.018	0.134	0.027	0.162
n	39,742	39,742	46,278	46,278
8th grade outcomes				
Read score	61.68	22.89	48.89	23.14
Math score	59.42	25.09	48.96	25.13
Number of days absent	9.46	9.28	10.06	9.72
Ever suspended	0.106	0.308	0.139	0.346
n	25,546	25,546	29,946	29,946

Note. Student data are from 2004-2005 to 2014-2015 academic year in Oregon.

C. Third grade students in "Meet the Standard" group

	Meet standard (6 points above)		Meet standard (6 points below)	
	mean	**std dev**	**mean**	**std dev**
4th grade outcomes				
Read score	32.75	18.65	19.74	15.58
Math score	36.45	23.11	25.54	21.08
Number of days absent	8.42	7.47	8.95	7.95
Ever suspended	0.038	0.191	0.046	0.211
n	43,710	43,710	14,514	14,514
8th grade outcomes				
Read score	34.54	21.31	23.41	18.42
Math score	37.68	23.91	28.04	21.73
Number of days absent	10.68	10.50	11.38	11.04
Ever suspended	0.177	0.382	0.219	0.413
n	26,734	26,734	8,877	8,877

Note. Student data are from 2004-2005 to 2014-2015 academic year in Oregon.

D. Third grade students in "Nearly Meet the Standard" group

	Nearly Meet standard (6 points above)		Nearly Meet standard (6 points below)	
	mean	**std dev**	**mean**	**std dev**
4th grade outcomes				
Read score	23.17	16.79	14.23	13.99
Math score	28.45	21.92	20.99	20.07
Number of days absent	8.85	7.81	9.19	8.21
Ever suspended	0.044	0.205	0.050	0.219
n	20,817	20,817	10,452	10,452
8th grade outcomes				
Read score	26.23	19.41	19.11	17.22
Math score	30.42	22.43	24.32	21.02
Number of days absent	11.29	11.11	11.93	11.56
Ever suspended	0.210	0.407	0.234	0.426
n	12,985	12,985	6,526	6,526

Note. Student data are from 2004-2005 to 2014-2015 academic year in Oregon.

E. Third grade students in "Low" group

	Low standard (6 points above)		Low standard (6 points below)	
	mean	**std dev**	**mean**	**std dev**
4th grade outcomes				
Read score	11.67	12.99	8.90	11.84
Math score	18.45	19.08	15.44	17.96
Number of days absent	9.67	8.51	10.23	8.82
Ever suspended	0.054	0.226	0.060	0.238
n	8,156	8,156	3,481	3,481
8th grade outcomes				
Read score	14.47	16.02	14.47	16.82
Math score	19.35	19.05	19.35	20.35
Number of days absent	12.72	12.30	12.72	11.97
Ever suspended	0.270	0.443	0.249	0.432
n	5,359	5,359	2,350	2,350

Note. Student data are from 2004-2005 to 2014-2015 academic year in Oregon.

Figure F1. McCrary Test for Meet the Standard/Exceed the Standard Cut-off

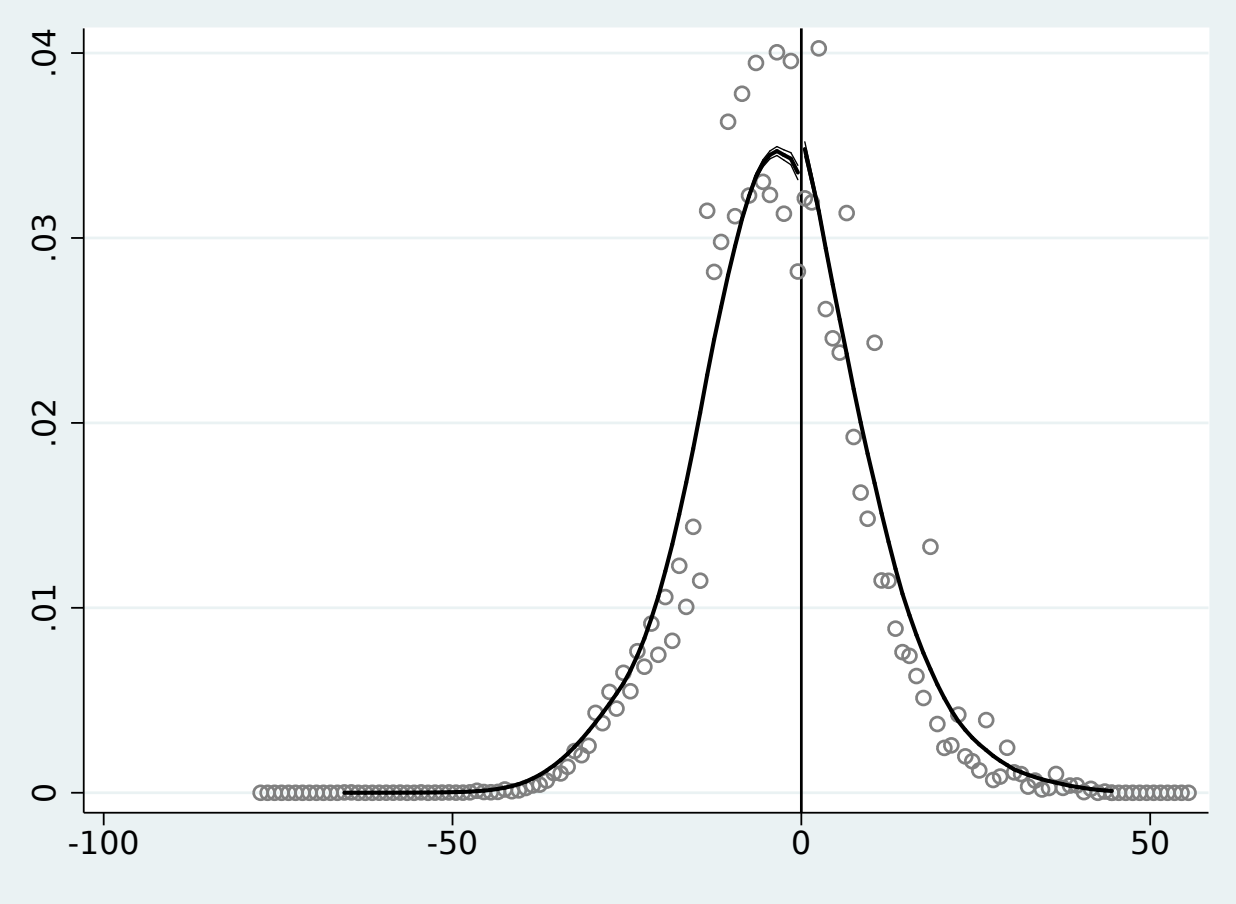


Figure F2. McCrary Test for Nearly Meet the Standard/Meet the Standard Cut-off

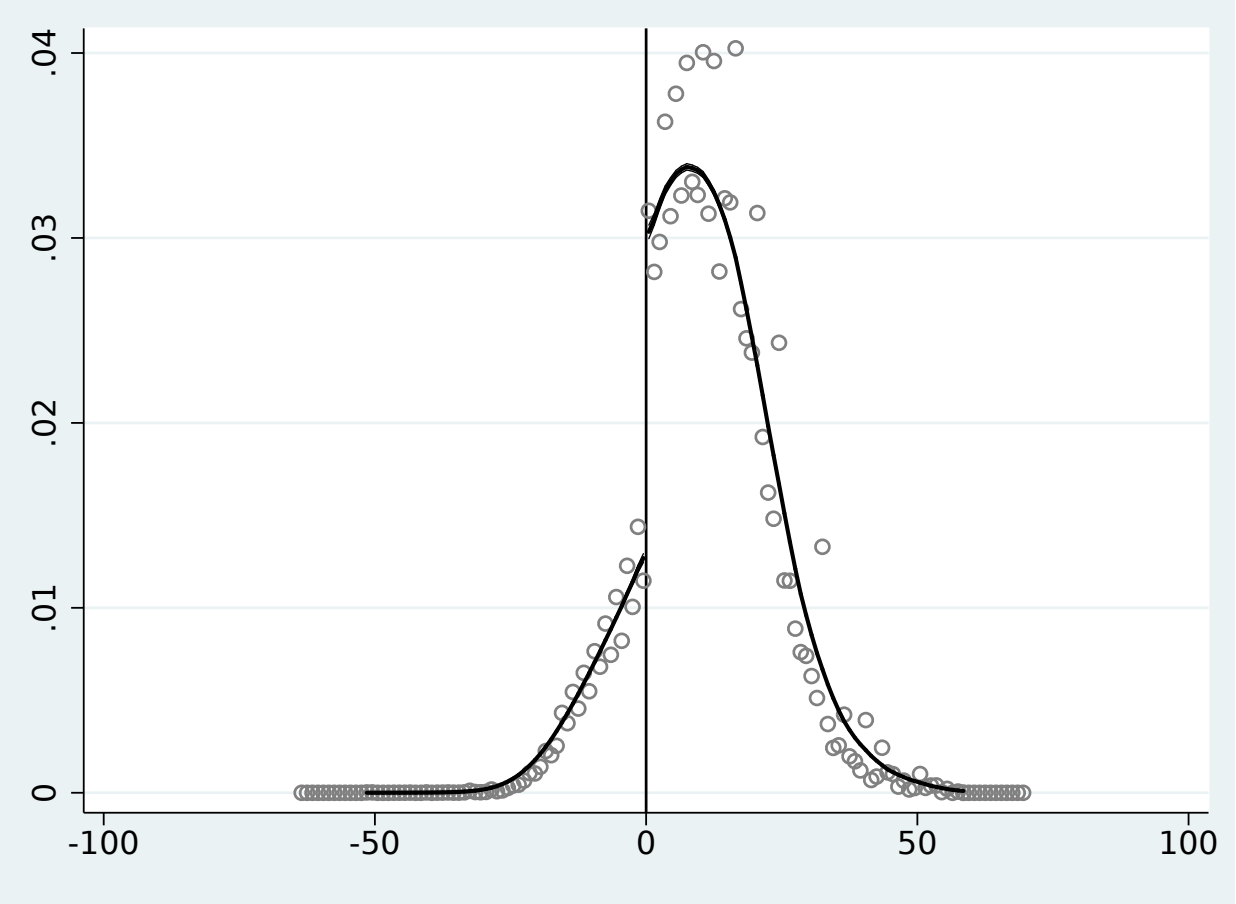


Figure F3. McCrary Test for Low/Nearly Meet the Standard Cut-off

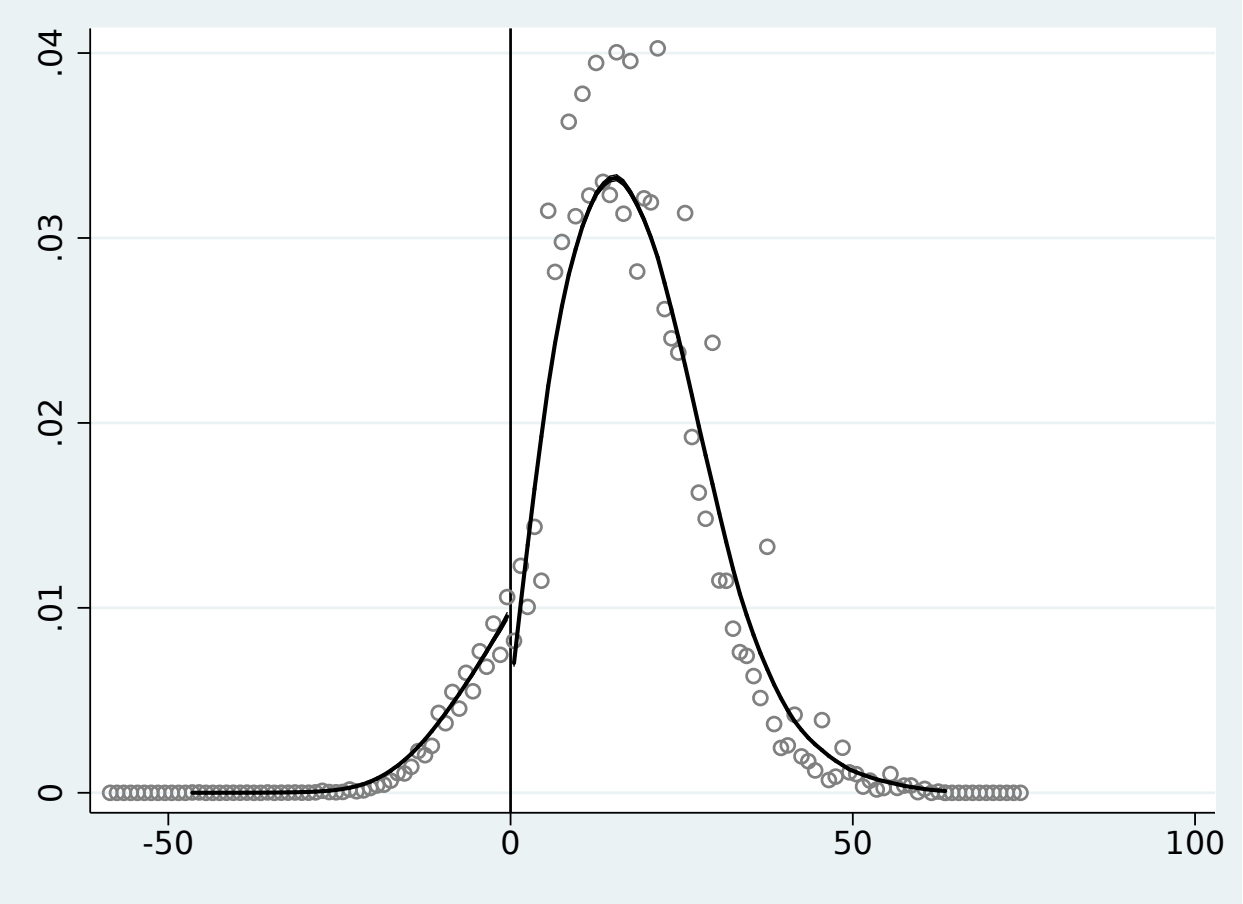


Figure F4. McCrary Test for Very Low/Low Cut-off

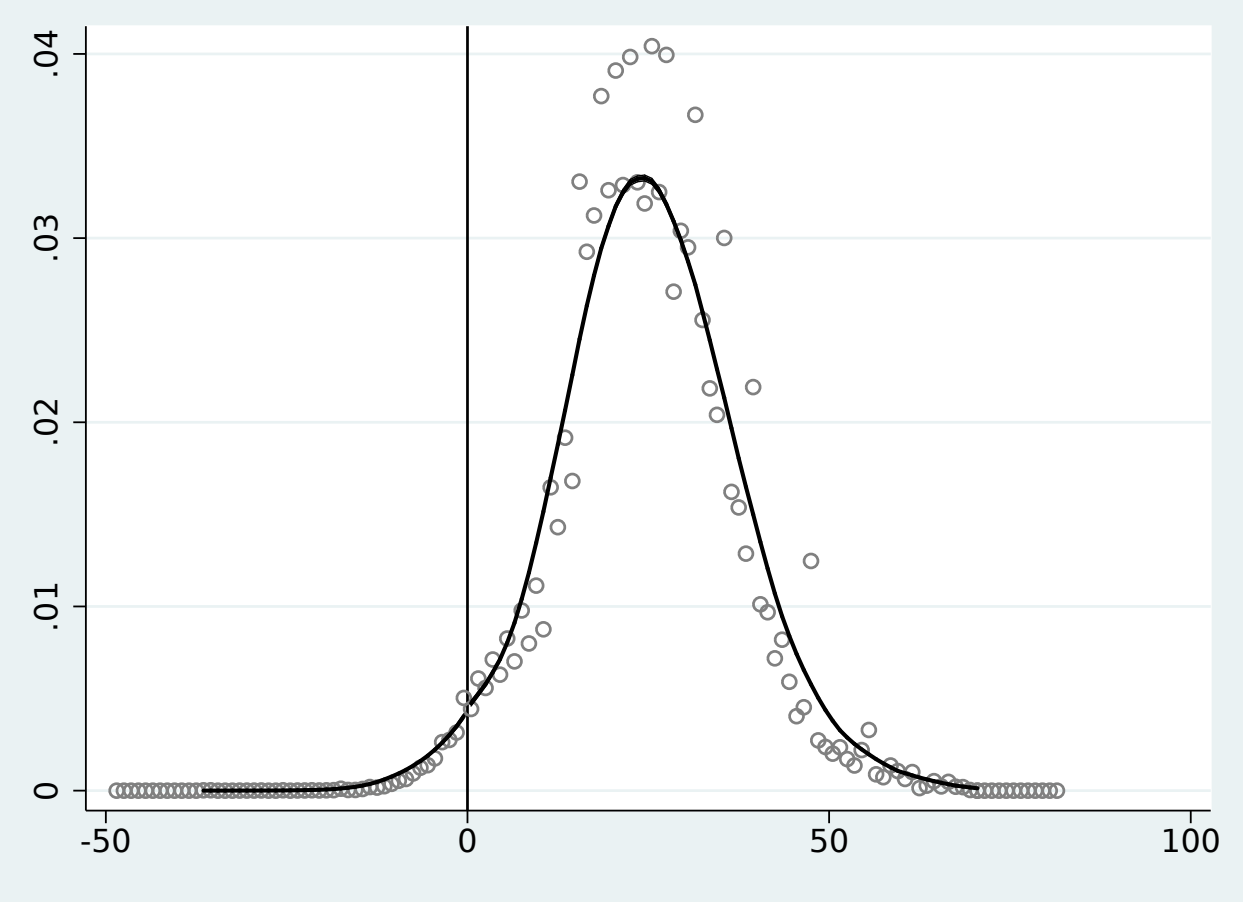
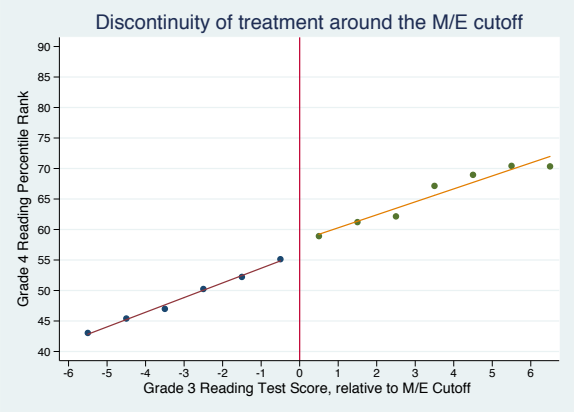
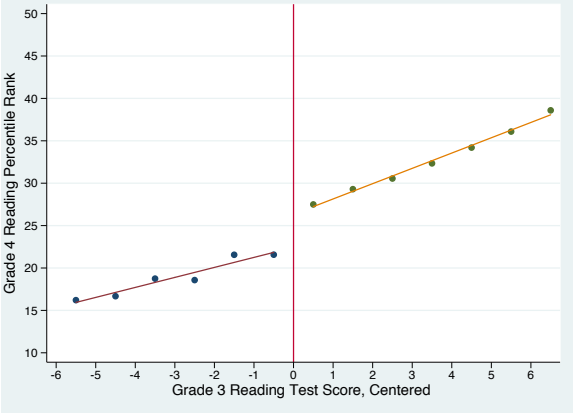


Figure F5. Fourth Grade Reading Percentile Rank across all cut-off scores

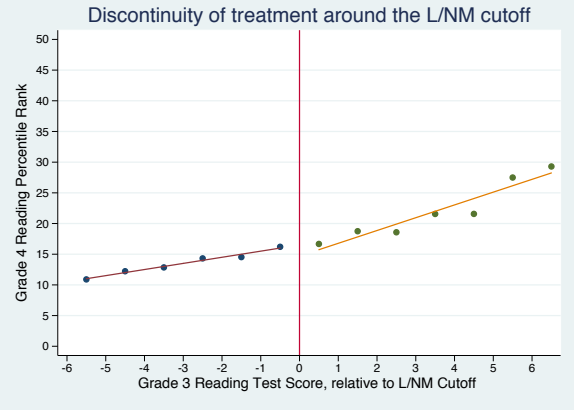
A. Meet the Standard/Exceed the Standard



B. Nearly Meet/Meet the Standard



C. Low/Nearly Meet the Standard



D. Very Low/Low

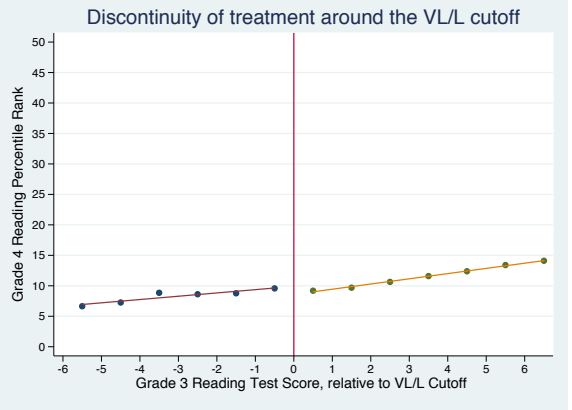
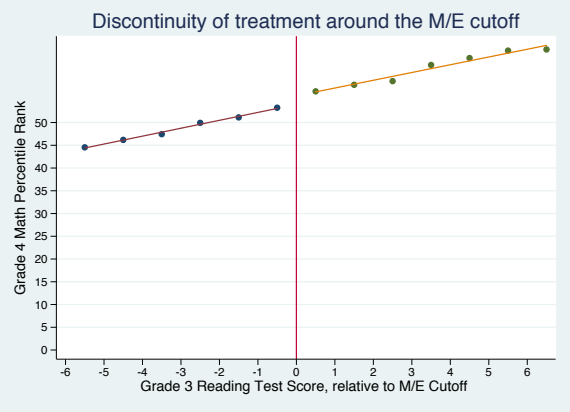
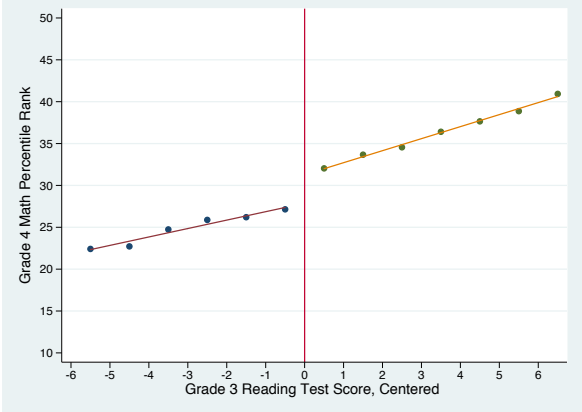


Figure F6. Fourth Grade Math Percentile Rank across all cut-off scores

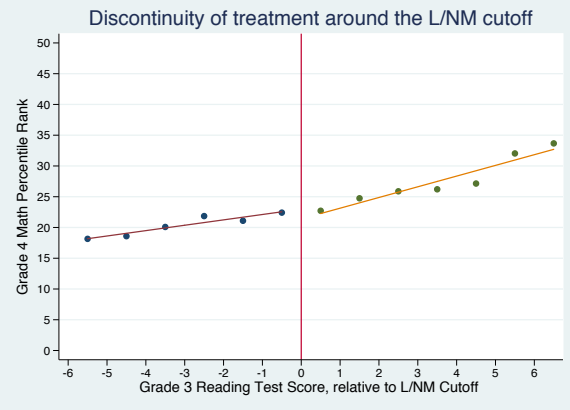
A. Meet the Standard/Exceed the Standard



B. Nearly Meet/Meet the Standard



C. Low/Nearly Meet the Standard



D. Very Low/Low

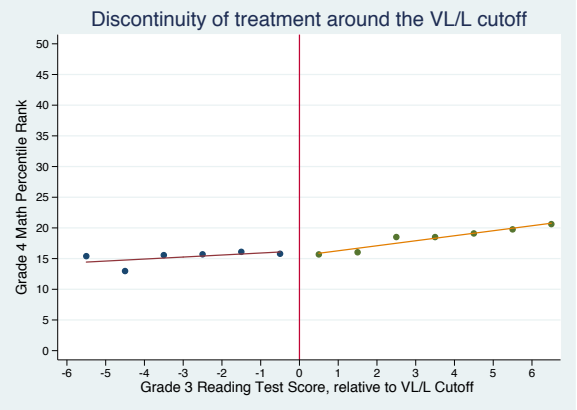
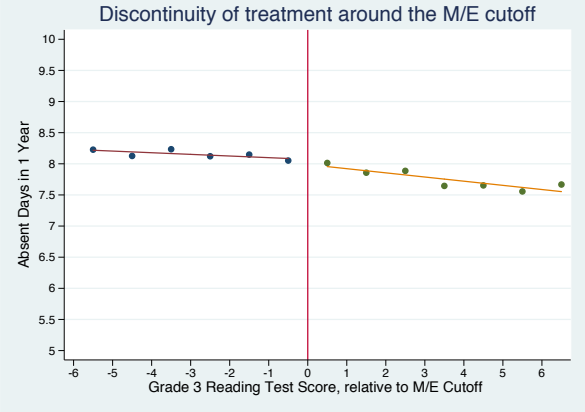
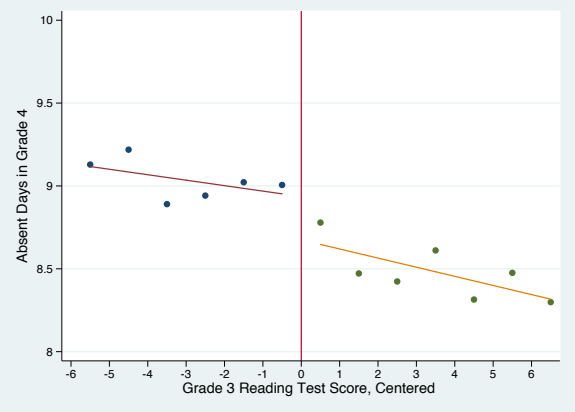


Figure F7. Fourth Grade Number Absent Days across all cut-off scores

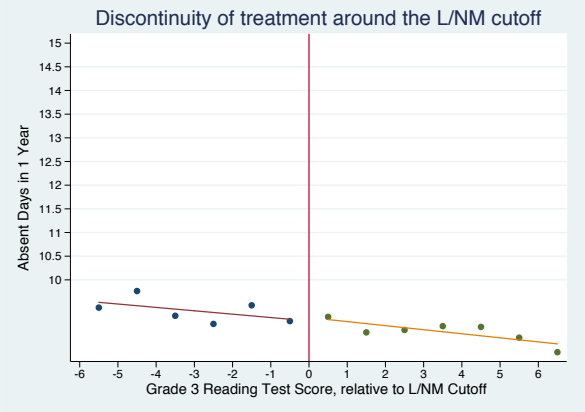
A. Meet the Standard/Exceed the Standard



B. Nearly Meet/Meet the Standard



C. Low/Nearly Meet the Standard



D. Very Low/Low

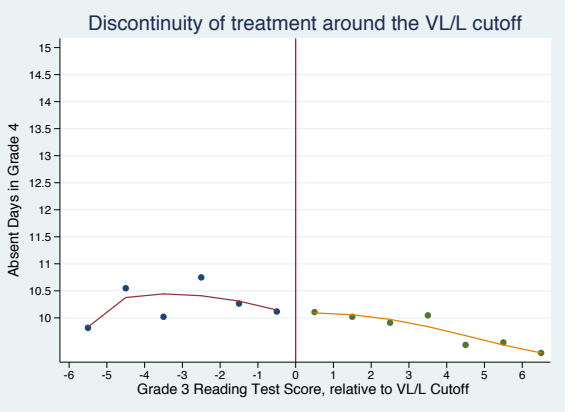
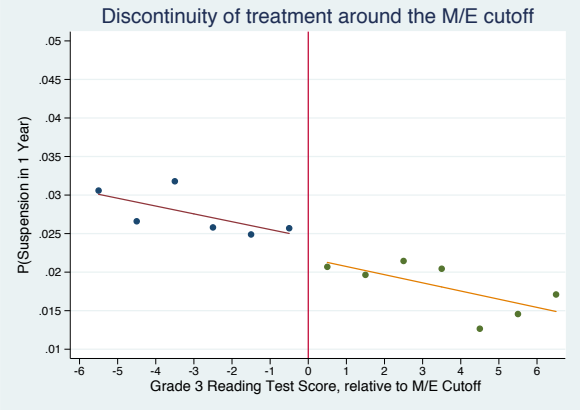
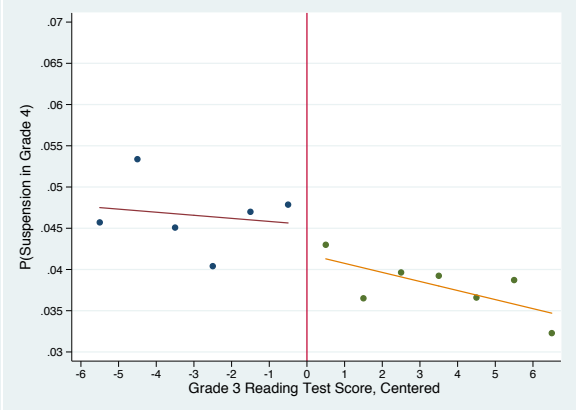


Figure F8. Fourth Grade Probability of Suspension across all cut-off scores

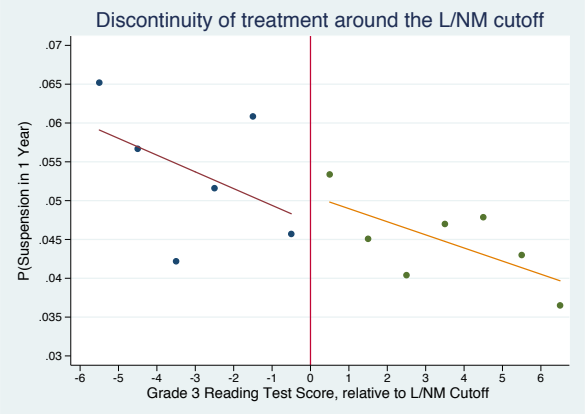
A. Meet the Standard/Exceed the Standard



B. Nearly Meet/Meet the Standard



C. Low/Nearly Meet the Standard



D. Very Low/Low

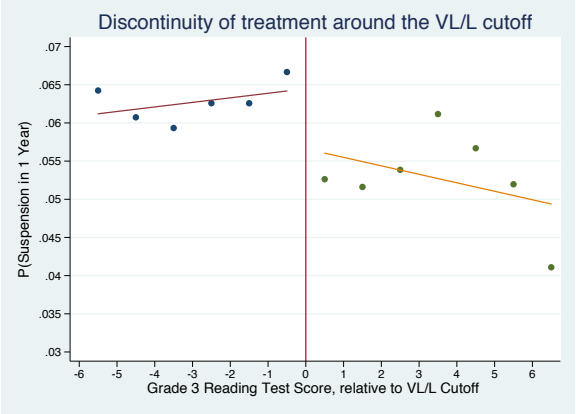
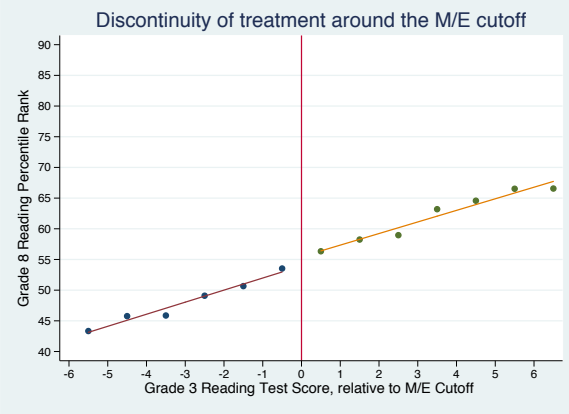
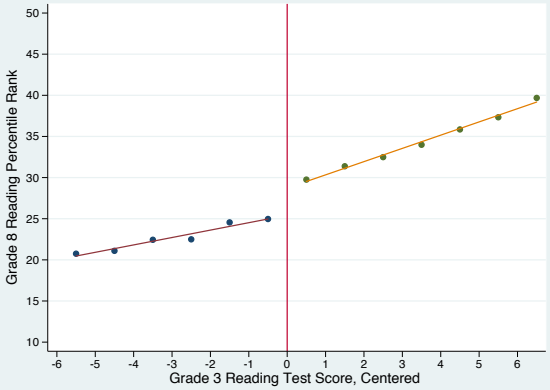


Figure F9. Eighth Grade Reading Percentile Rank across all cut-off scores

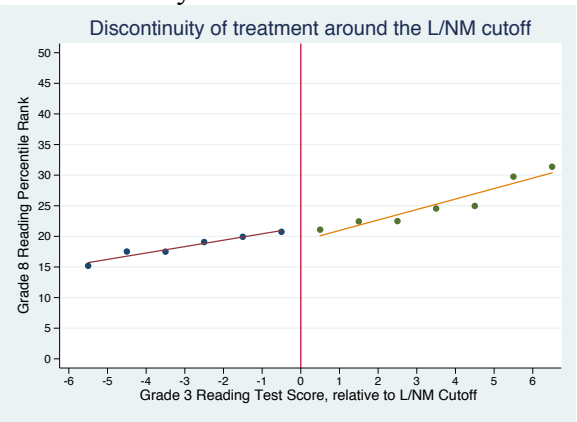
A. Meet the Standard/Exceed the Standard



B. Nearly Meet/Meet the Standard



C. Low/Nearly Meet the Standard



D. Very Low/Low

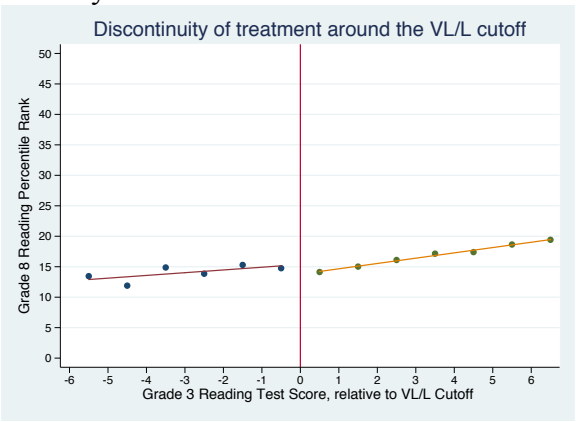
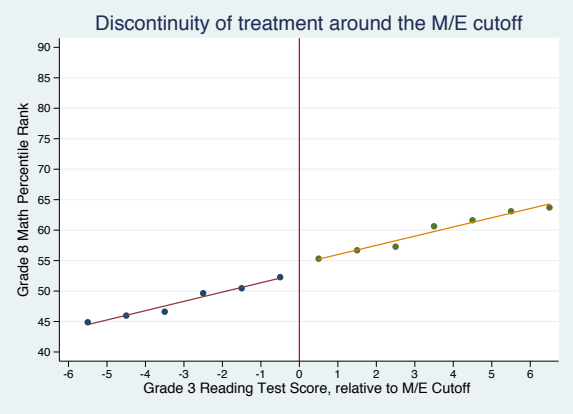
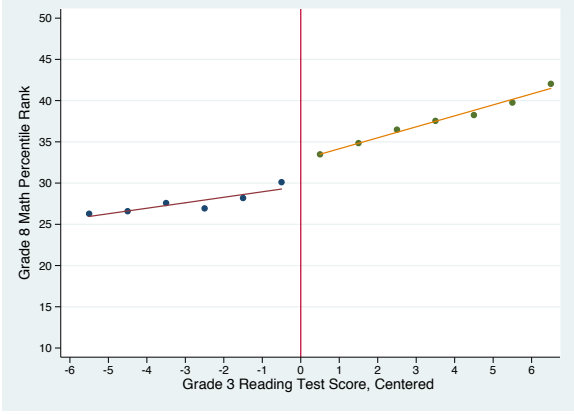


Figure F10. Eighth Grade Math Percentile Rank across all cut-off scores

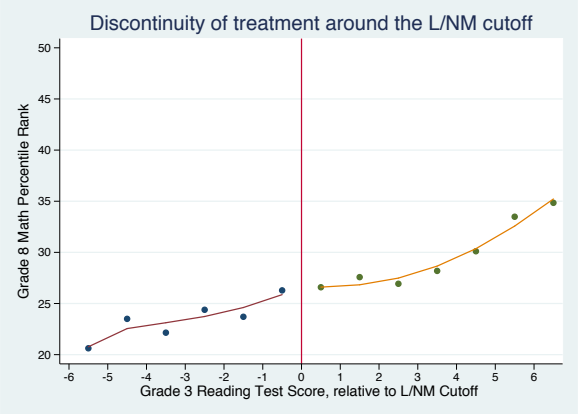
A. Meet the Standard/Exceed the Standard



B. Nearly Meet/Meet the Standard



C. Low/Nearly Meet the Standard



D. Very Low/Low

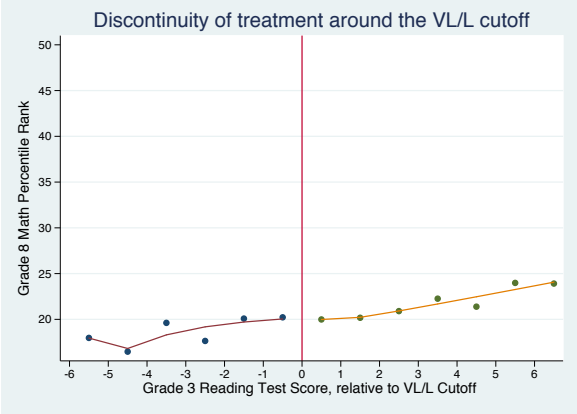
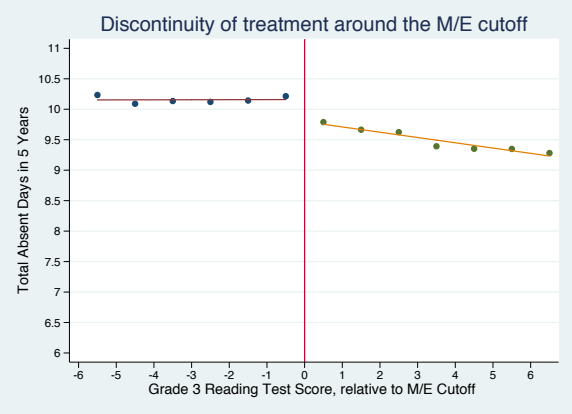
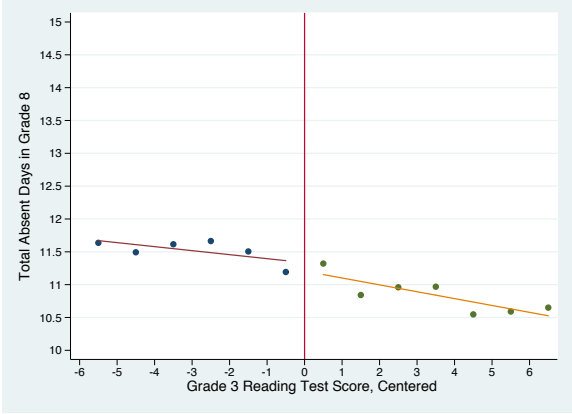


Figure F11. Eighth Grade Number of Absent Days across all cut-off scores

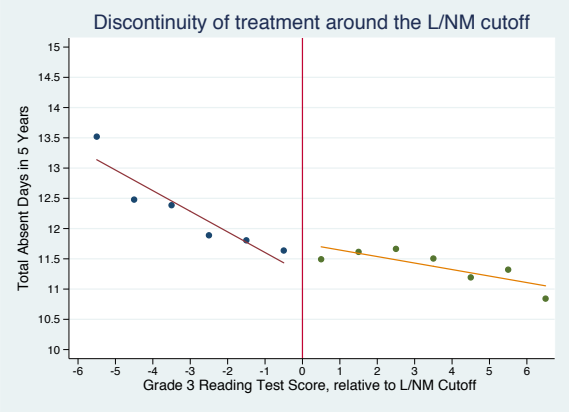
A. Meet the Standard/Exceed the Standard



B. Nearly Meet/Meet the Standard



C. Low/Nearly Meet the Standard



D. Very Low/Low

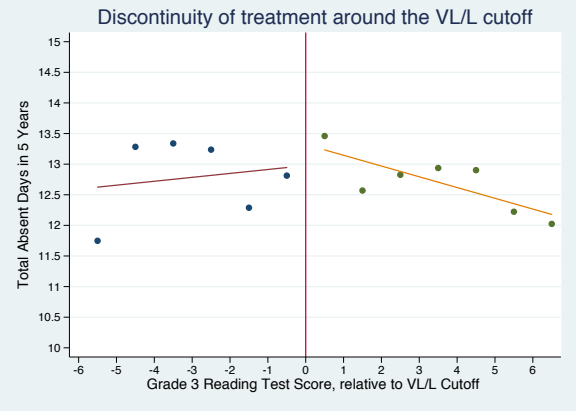
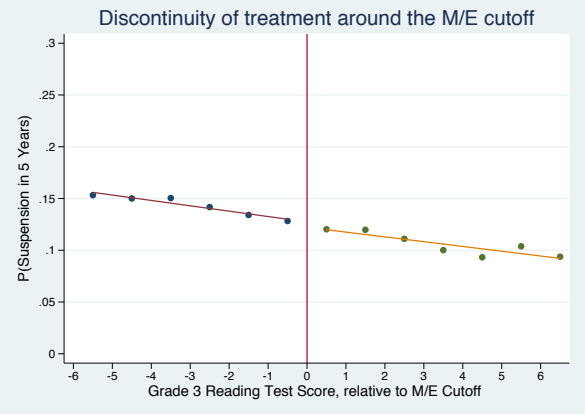
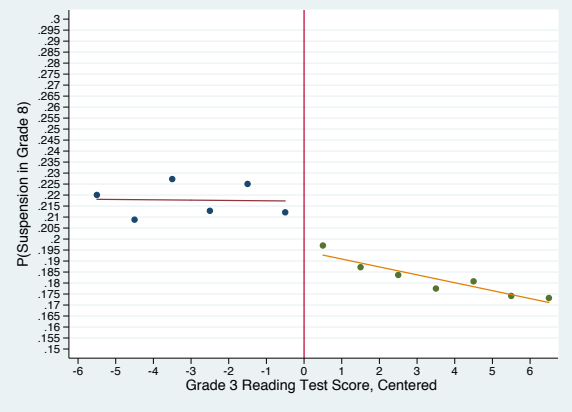


Figure F12. Eighth Grade Probability of Suspension across all cut-off scores

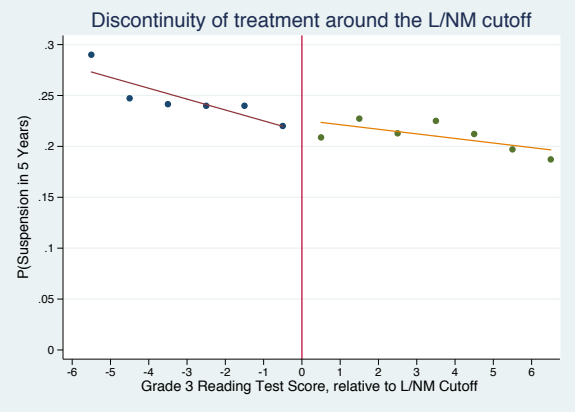
A. Meet the Standard/Exceed the Standard



B. Nearly Meet/Meet the Standard



C. Low/Nearly Meet the Standard



D. Very Low/Low