ESSAYS ON STATE MERIT-BASED FINANCIAL AID: SEEKING STEM, REVISITING METHODOLOGY, AND TEST RETAKERS

A Dissertation presented to
the Faculty of the Graduate School
at the University of Missouri

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

---

by
JUNPENG YAN
Dr. Bradley Curs, Dissertation Supervisor
DECEMBER 2020

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

ESSAYS ON STATE MERIT-BASED FINANCIAL AID:
SEEKING STEM, REVISITING METHODOLOGY,
AND TEST RETAKERS

presented by Junpeng Yan,
a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Bradley Curs

Dr. Rajeev Darolia

Dr. Cory Koedel

Dr. Se Woong Lee

DEDICATION

This dissertation is dedicated to my parents,

for their endless love, support, and encouragement.

谨以此文献给我的父母

感谢他们无尽的爱、支持、和鼓励

ACKNOWLEDGMENTS

Every journey has an end. Now, it is time to say goodbye. I think writing the acknowledgment is a privilege but also the most challenging part of the whole dissertation, since I owe words of gratitude to so many people who have assisted me. I know I cannot go through this long journey without their support.

I would like to express my sincere gratitude to Dr. Bradley Curs, my dissertation advisor. Brad, thank you for your continuous support during my whole Ph.D. journey. Your insightful suggestions and constant encouragement make it possible for me to finish this dissertation, particularly during this pandemic. Thanks again for your leadership and mentorship.

I am also so grateful to have such fantastic committee members, Drs. Rajeev Darolia, Cory Koedel, Se Woong Lee. Raj, I really hope you can stay in Missouri longer than expected. I very appreciate your supervision during the first two years of my Ph.D. life. Your critical comments and constructive suggestions always push me forward. Cory, your feedback is always informative and helps me to validate my empirical research again and again. My dissertation cannot be substantially improved without your comments. Se Woong, you often guide my research from a multi-disciplinary perspective and share your experience as an Asian scholar in the United States. Your valuable advice I will never forget. Thank you all so much for your willingness to serve on my dissertation committee.

I want to give special thanks to Dr. Mark Ehlert (Economics and EPARC) and Mr. Jeremy Kintzel (MDHE). You both give me strong supports in data collection and cleaning. The dissertation cannot be completed without your help. I also miss the time when I was working at EPARC and MDHE.

I also want to mention many kindly faculty members at the University of

Missouri, not limited to Dr. Lisa Dorner (ELPA), Dr. Pilar Mendoza (EPLA), Dr. James Sebastian (ELPA), Dr. Oded Gurantz (TSPA), and Dr. Joan Hermsen (Sociology). Thank you all for your encouragement and support.

Special thanks sent to many of my fellow Ph.D. peers in ELPA, TSPA, and Economics. I value the memories that we studied and worked together. Besides, I want to thank Sheng Zou, my former roommate, for helping me to quickly adapt to my new life in Columbia.

Though the pandemic brings lots of international travel restrictions, it cannot stop me from expressing my most sincere appreciation to two of my professors in China. I want to thank Dr. Zhiyuan Ma and Dr. Changxuan Mao. You are the greatest professors I have ever met. You both not only teach students knowledge but also tell them how to become qualified citizens. Without your encouragement, I cannot choose to pursue such a doctoral degree in higher education policy. This dissertation has your efforts.

Last, I would like to specifically thank my girlfriend, Min Xiao. Your companionship always gives me energy and makes me feel stronger. Thanks for your support and understanding during this hard time.

TABLE OF CONTENTS

CHAPTER

APPENDIX

LIST OF TABLES

Table Page

LIST OF FIGURES

ABSTRACT

This dissertation investigates several topics in state merit-based financial aid research. The main research content includes three separate studies, primarily using the administrative datasets from the Missouri Department of Higher Education. The first study is an empirical paper that examines the effects of the Missouri Bright Flight Scholarship (a merit-based financial aid) on producing STEM graduates in Missouri 4-year public institutions. The fuzzy regression discontinuity estimates indicate that this merit aid program has a negative but statistically insignificant effect on STEM/engineering initial major choice and degree completion. Particularly, the aid still has a negative impact on male students but may encourage female students to enter a STEM field. The second study is a technical note that addresses three ignored issues about regression discontinuity application in merit-based financial aid research: retaking and test score manipulation, rounding errors in the running variable, and misleading statistical inference. I re-examine the practices in recent related studies and provide recommendations for future financial aid research, particularly about dealing with test scores appropriately in regression discontinuity designs. The third study is a policy brief about the consequence of allowing test retaking in merit aid programs. Using the Missouri administrative data, I primarily compare the percentages of aid eligible students based on both the first-time and the highest ACT composite scores in different demographic groups. The empirical results indicate that underrepresented students are less likely to retake the ACT compared to their peers and the acceptance of the highest score brings more inequitable impacts on underrepresented students. The discussion may help policymakers to better understand the retaking behaviors so that they can make related policies to benefit underrepresented students with more effectiveness.

CHAPTER 1

Introduction

State merit-based financial aid has become an increasingly popular policy tool to encourage well-prepared high school graduates to enroll in in-state colleges. State merit-based financial aid programs have been widely adopted since 1990 (Dynarski, 2002; Frisvold & Pitts, 2018; Zhang & Ness, 2010). The rapid growth has also raised much attention in academia. Lots of empirical papers have evaluated many potential impacts of the merit-based financial aid, including college enrollment (Cornwell, Mustard, & Sridhar, 2006; Dynarski, 2000, 2003; Singell et al., 2006), college choice (Bruce & Carruthers, 2014; Cohodes & Goodman, 2014; Dynarski, 2002; Zhang et al., 2016), persistence and graduation (Cohodes & Goodman, 2014; Scott-Clayton, 2011; Welch, 2014), interstate migration (Fitzpatrick & Jones, 2012; Harrington et al., 2016; Leguizamon & Hammond, 2015; Zhang & Ness, 2010), and post-graduation earning (Scott-Clayton & Zafar, 2019; Welch, 2014). Generally, the literature has found that many state merit-based financial aid programs have positive effects on the probability of attending in-state institutions and keeps talents to study and work continuously in the state.

This dissertation includes three separate research papers related to merit-based financial aid research. The first study is an empirical paper that examines the effects of the Missouri Bright Flight Scholarship (a merit-based financial aid) on STEM major choice and degree completion in Missouri 4-year public institutions. The second study is a technical note that addresses three ignored issues about regression discontinuity application in merit-based financial aid research: retaking and test score manipulation, rounding errors in the running variable, and misleading statistical inference. The third study is a policy brief about the impact of allowing test retaking on the educational equity of merit aid programs.

Beyond those traditional student outcomes in the literature, considering the current STEM workforce shortage, I am wondering whether these merit-based financial aid programs can benefit the labor market by producing more STEM graduates. In Chapter 2, my first paper tries to answer this question. This study presents evidence of the effects of the Missouri Bright Flight Scholarship Program, a highly targeted state-funded merit aid program, on postsecondary STEM enrollment and degree attainment. The analytical sample derived from Missouri administrative microdata includes 11 cohorts of first-year students in Missouri public 4-year institutions (roughly 130,000 students) between 1996 and 2006. Utilizing discontinuities in the aid award formula and a fuzzy regression discontinuity research design, I find that this merit aid program has a negative but statistically insignificant effect on STEM/engineering initial major choice and degree completion. However, when disaggregated by gender, the effects differ. After receiving the merit aid award, male students are less likely to enroll in a STEM/engineering field while female students are more likely to enter a STEM/engineering field.

In the literature, regression discontinuity design has been widely used to evaluate the causal effect of financial aid programs on student outcomes. Some of the regression discontinuity studies are derived from policy cut-offs in standardized exams and use test scores as the primary running variables (e.g., the ACT or the SAT scores). In Chapter 3, I revisit this popular quasi-experimental design and discuss three important issues when using the ACT scores as the running variables: retaking and test score manipulation, rounding error in the running variable, and misleading statistical inference. Specifically, I utilize the ignored feature of how the ACT composite score is generated to re-examine the practices in previous studies. Several recommendations are provided to researchers to guide future financial aid research, particularly about dealing with test scores appropriately in regression discontinuity designs.

Educational equity has been raised as another important concern when evaluating state merit-based financial aid programs. Researchers have argued about whether it is equitable and efficient to use state funds to provide discount tuition for relatively affluent students (Heller & Marin, 2004). Empirical evidence has indicated that some merit aid is inequitable overall (Binder & Ganderton, 2004; Cornwell & Mustard, 2004). In Chapter 4, I focus on a less discussed cause of inequities in many merit aid programs. Many aid programs do not have specific restrictions on whether students can retake the standardized tests to increase their chances of eligibility. Due to the different retaking behaviors among different demographic groups, allowing retaking may make the equity problems in merit aid even more serious. Using the Missouri administrative data, I primarily compare the percentages of aid eligible students based on both the first-time and the highest ACT composite scores in different demographic groups. The empirical results indicate that underrepresented students are less likely to retake the ACT compared to their peers. As a result, the acceptance of the highest score brings more inequitable impacts on those underrepresented students. Policymakers are suggested to revisit their current retaking policies and remove the financial barriers to help underrepresented students be well-prepared for standardized tests and financial aid applications.

CHAPTER 2

State Merit-based Financial Aid and Postsecondary STEM Enrollment and Attainment: Evidence from Missouri

With the rapid development of science, technology, engineering, and math (STEM) occupations, STEM educational credentials have become more valuable on the labor market. Between May 2009 and May 2015, employment in STEM occupations grew 10.5% while non-STEM occupations increased by only 5.2%. Moreover, 93% of STEM occupations had wages significantly above the national average wage for all occupations. Specifically, the national average wage for all STEM occupations was $87,570, compared with $45,700 for non-STEM occupations (Fayer et al., 2017).

Given the STEM workforce shortage and associated wage premiums, an important question for higher education policymakers is how to increase college enrollment and degree attainment in STEM fields. In addition to the general workforce shortage, there is a notable gender gap in STEM employment and postsecondary attainment. In 2015, women constituted only 28% of workers in STEM occupations and represented 40% of employed workers with the highest degree in STEM while the corresponding shares in 1993 were 23% and 30%, respectively (National Science Board, 2018). Though the representation of females in post-college STEM fields has increased in recent decades, the gender gap still exists. A key contributor to the STEM occupational gender gap is the fact that the current higher education system is producing fewer female STEM graduates[1]. It is probably because females are less likely to pick a

---

[1] In 2012, women represented 50.5% of all STEM bachelor's degree recipients (almost evenly shared with men), but they only shared 45.6% (master's) and 40.6% (doctorate) of advanced STEM degrees. Moreover, in each gender group, only 28.7% of all female bachelor's degree recipients were in STEM fields, and only 1.5% of them were in engineering fields. The same percentages of all male bachelor's degree recipients were 37.8% and 8.7% (Fiegener, 2015)

STEM major initially and more likely to switch out of STEM (Chen & Weko, 2009). Further, females' motivations to earn a STEM degree differ by discipline. Women are well-represented in social science (70%) and in biological science (62%) while they are minorities in engineering (21%) and in math and computer science (25%) (Malcom & Feder, 2016). Higher education research has posited a number of explanations for this gender gap in universities and colleges, including academic preparation (Griffith, 2010), beliefs and preferences (Zafar, 2013), grading policies (Ahn et al., 2019), and instructors (Price, 2010). Thus, an additional crucial focus for higher education policymakers should be on how to increase the proportion of female college students in STEM degree programs.

Affordability is one of the most vital determinants of higher education choice. With the continual growth in postsecondary tuition and fees, the financial cost has been shown to be a crucial factor when students choose their field of study (Andrews & Stange, 2016; Rothstein & Rouse, 2011; Stange, 2015). Furthermore, the cost of producing STEM degrees are generally more expensive than non-STEM ones (American Institutes for Research, 2013). According to the estimates of "cost per degree" for 28 disciplines, engineering, agriculture, computer, and science-related fields have higher than average production costs. Typically, the "cost per degree" of most STEM programs in public 4-year institutions ranges between \$65,000 and \$80,000[2]. Furthermore, the similar cost of an engineering degree is even higher (nearly \$100,000).

As a consequence, institutions are increasingly implementing differential tuition policies, which often raise the cost of STEM courses and students are paying a higher price that aligns more closely with the actual institutional cost of delivering the degree received (Stange, 2015). In 2010-2011, about 40% of public 4-year institutions with doctoral programs had differential tuitions among different majors (Ehrenberg,

---

[2] To be clarified, the Delta Cost Project report (American Institutes for Research, 2013) separates social, behavioral, and economic sciences (SBE) from STEM disciplines while those majors are often regarded as STEM disciplines in other reports. The costs of SBE disciplines (e.g., social science and psychology) are approximately \$45,000.

2011). Empirically, Stange (2015) found that the differential pricing of engineering and business programs resulted in the decreasing shares of students studying those majors. It indicates that the differential tuition policies may impact students' major choice. This pricing strategy is shifting the financial burden onto students, and in the end, it may create financial incentives for students to choose non-STEM majors.

Compared to other possible determinants of STEM major choice examined in the literature, such as academic preparation/achievement, self-efficacy beliefs, expected earnings, identity, and peer influence (Freeman & Hirsch, 2008; Griffith, 2010; Montmarquette et al., 2002; Price, 2010; Wang, 2013; Wiswall & Zafar, 2015), the affordability of STEM degrees can be directly influenced by federal and state policy. For example, financial aid can be delivered to students from low-income families and underrepresented groups. The financial incentives may strongly motivate students to choose STEM majors and eventually contribute to STEM degree attainment. Particularly, there existed a need-based financial aid program targeting on STEM majors. From the fall of 2006 to the summer of 2011, the SMART program was designed to complement the existing Pell Grant program, and more importantly, it had an emphasis on certain majors including STEM. However, the empirical results do not indicate consistent effects on STEM major choice (Denning & Turley, 2017; Evans, 2017) and more studies are required.

In this paper, I examine the role that merit-based financial aid plays in the decision to pursue a STEM degree. Specifically, I evaluate the treatment effect of receiving the Missouri Bright Flight Scholarship Program, a state-funded merit-based financial aid, on college students' STEM initial major choice and degree completion in Missouri 4-year public universities. I explore the following two research questions: 1) What is the effect of receiving financial aid on STEM attainment, including initial major choice and degree completion? 2) Are there heterogeneous effects of financial aid on STEM attainment by gender?

Background

The Missouri Bright Flight Scholarship Program

The Missouri Bright Flight Scholarship Program was established by the Missouri Legislature in 1986. It is a merit-based financial aid program that encourages top-ranked high school graduates to enroll full-time at a participating[3] Missouri institution, including 13 public 4-year universities, 14 public 2-year institutions, and over 25 private institutions[4]. The eligibility of this financial aid award was intended to be the top 3% of all Missouri ACT takers. Before the 2008-2009 academic year, this program required initial students to have a composite ACT score of 30 or higher. To maintain the top 3% ACT requirement, the minimum ACT score was increased to 31 after the 2008-2009 academic year (Kumar, 2008).

The initial amount of the scholarship was $2,000 per academic year and was not adjusted for inflation until the 2007-2008 academic year. In 2007, the authorizing legislation was revised to add the top 4th and 5th percentiles as eligible for a lesser award, becoming a two-tier award structure. The first tier (top 3%) recipients can get $3,000 while the second tier (4%-5%) recipients are awarded only $1,000. Nevertheless, that change did not become effective until the 2010-2011 academic year. It is important to note that the second-tier award has never received funding through the program and the first-tier award was never fully funded (at the maximum amount of $3,000) until the 2015-2016 academic year. From the 2008-2009 academic year to the 2014-2015 academic year, recipients only received approximately $2,500 per year.

To be eligible for this financial aid for the first time, initial applicants must

---

[3] For a current list of participating institutions, please see here: https://dhewd.mo.gov/ppc/grants/documents/participatingschools.pdf

[4] Though the Bright Flight Scholarship Program cooperates with different types of higher education institutions in Missouri (public, private, 2-year, 4-year, etc.), most of the recipients are probably enrolled in public 4-year institutions. According to the Missouri Department of Higher Education (Missouri Department of Higher Education, 2010), in the fiscal year 2007, the total number of Bright Flight recipients was 8,541, including 6,463 students (75.7%) within the 13 public 4-year universities.

achieve the qualifying score by the June test date immediately following the graduation from high school. Students can retake the ACT multiple times in high school and only their highest ACT scores are considered. Besides the qualified ACT score, the eligibility also requires applicants to enroll full-time at a participating Missouri university and not to pursue a degree or certificate in theology or divinity. Students can renew this financial aid for up to 10 semesters or until completed a bachelor's degree (whichever occurs first) if they keep enrolled full-time and maintain a cumulative GPA over 2.5 and otherwise maintain satisfactory academic progress as defined by enrolled institutions.

Differential Tuition by Disciplines

Nationally, Malcom and Feder (2016) summarized the average net price for a STEM degree in the 2007-2008 academic year. The price ranges from $7,800 for minority students at a public 4-year institution to nearly $30,000 for students at a private research institution. However, the comparable data of the net price for non-STEM degrees are not available. After compared with a report from the College Board (The College Board, 2014), it is still reasonable to claim that the average price for a STEM degree in the 2007-2008 academic was more expensive because it is even higher than the price of a non-STEM degree 6 years later (Malcom & Feder, 2016, p.127).

In my specific setting in Missouri, though Missouri universities and colleges were not mentioned on the list of institutions with differential tuitions (Ehrenberg, 2011), students enrolled in some institutions are still facing differential costs actually due to the enrollment fees varied by disciplines. For example, in the 2006-2007 academic year, the highest course fees for per credit hour in the University of Missouri System (Columbia, Rolla, Saint Louis, Kansas City) were: engineering ($50.5), and health ($49.3), which is at least 20% more expensive than many other courses (except nursing related courses), such as arts & science ($0), business and education

($32), Journalism ($37.5) (University of Missouri System, 2006). Since the 2006-2007 academic year, Missouri University of Science and Technology began to charge an additional $50 supplemental fees for computer science, biological science, chemistry, and so on. Even claiming an initial major in STEM fields is not equal to enrolling more expensive engineering courses immediately, the differential fees at least increase the expected total cost of degree completion in STEM fields and may have a negative effect on STEM attainment finally.

Literature Review

State-funded merit-based financial aid has been widely established since 1990 (Dynarski, 2002; Frisvold & Pitts, 2018; Zhang & Ness, 2010), with examples including the Helping Outstanding Pupils Educationally (HOPE) Scholarship (Georgia, since 1993), and the Bright Future Scholarship (Florida, since 1997). One important rationale for merit aid adoption is the goal of keeping the best and brightest students in state (Heller & Marin, 2004). As a result, the evaluations of merit-based financial aid have primarily focused on the following student outcomes: college enrollment (Cornwell, Mustard, & Sridhar, 2006; Dynarski, 2000, 2003; Singell et al., 2006), college choice (Bruce & Carruthers, 2014; Cohodes & Goodman, 2014; Dynarski, 2002; Zhang et al., 2016), persistence and graduation (Cohodes & Goodman, 2014; Scott-Clayton, 2011; Welch, 2014), interstate migration (Fitzpatrick & Jones, 2012; Harrington et al., 2016; Leguizamon & Hammond, 2015; Zhang & Ness, 2010), and post-graduation earning (Scott-Clayton & Zafar, 2019; Welch, 2014). Generally, these studies conclude that merit aid has a significant positive effect on the probability of attending in-state institutions and keeps talents to study and work continuously in the state.

Because of the STEM workforce shortage, evaluating state-funded merit aid programs with a focus on producing more STEM graduates becomes increasingly necessary. Unlike the topics mentioned above, the effect of merit aid on college major choice has not been fully examined. The following several empirical papers are the primary studies that have evaluated the effects of merit aid on college major choice and degree completion.

Cornwell, Lee, et al. (2006) explored the effect of Georgia HOPE merit aid on students from the University of Georgia through a difference-in-differences approach. They found that HOPE increased freshman GPA while reduced the number of credit hours completed in math and science core curriculum courses in their first year. They also found small significant effects for major choice. The HOPE increased the likelihood of a student choosing an education major by 1.7% but did not affect other majors. To explain the result, they posited a possible argument that the HOPE merit aid had a renewal requirement that students had to maintain a 3.0 GPA. In order to avoid losing HOPE in the future, students might strategically take less complicated courses or even choose some easier subjects to meet the required GPA for renewal.

Stater (2011) examined the effect of student aid on the choice of the first-year major using individual-level data from three flagship public universities (Colorado, Indiana, and Oregon). He found that loans and grants probably had small or insignificant effects on major choice while merit aid showed a large positive effect on claiming a major in humanities and science. Although he controlled student characteristics, the manipulation to receive merit aid, as an endogeneity, was not considered. For example, students who want to receive merit aid may retake standardized exams or choose easier high school courses to maintain a high GPA. Those unobserved behaviors could affect major choice.

Zhang (2011) focused on the effect of the Florida and Georgia merit aid on annual statewide conferred STEM degrees using the Integrated Postsecondary Education Data System's (IPEDS) Completion Survey. The empirical results showed that HOPE (GA) and Bright Future (FL) programs increased the number of both STEM and non-STEM baccalaureate degrees conferred by 4-year institutions in Georgia and Florida. But no significant effect was found on the percentage of STEM majors, except a 1.6% increase in Florida private institutions. It might be problematic since this study used an aggregated dataset instead of individual-level microdata. Merit aid can attract students with a better academic background to stay in their home state for higher education. The increased student quality may also affect students' major choice, which cannot be excluded without student-level information.

Sjoquist and Winters (2015a) investigated national microdata from the American Community Survey (ACS) and evaluated the effect of state merit aids on students' major choice. Unlike previous studies with limited samples, they examined the individual-level data with a national-wide sample, making their results with stronger external validity. They also considered the different scales among different state merit aid programs. Nine states merit aid programs were defined as strong merit aid programs according to the eligibility criteria, the number of recipients, and the size of the award. Their results suggested that adopting a strong merit aid program reduced the number of STEM graduates from the state by 6.5%. As a complementary to Sjoquist and Winters (2015a), Sjoquist and Winters (2015b) used administrative data from the University System of Georgia to examine whether Georgia's HOPE Scholarship had affected students' college major decisions. Similarly, they also found the negative effect of merit aid on STEM degree production.

With respect to the extant literature, this study contributes in three primary ways. First, many analytic samples of state-funded financial aid evaluation are derived from Georgia and Florida (Castleman et al., 2018; Cornwell, Lee, et al., 2006; Singell et al., 2006; Sjoquist & Winters, 2015a, 2015b; Zhang, 2011; Zhang et al., 2016). Though financial aid programs in Georgia and Florida have larger targeted subjects and make sizable impacts on students, as a supplementary to the literature, it is still necessary to have more studies with different state-funded financial aid programs. Hence, this study explores a different state-funded financial aid program and can enlarge the literature with additional empirical evidence. Unlike other state-funded merit aids, the Bright Flight Scholarship Program only targets a few Missouri students with very high ACT scores (30 or above). Others usually have a lower ACT requirement but may include a requirement of high school GPA (HS GPA) : Florida (two tiers: ACT 20, HS GPA 3.0; ACT 28, HS GPA 3.5); Georgia (HS GPA 3.0); Louisiana (three tiers: ACT 20, HS GPA 3.0; ACT 23, HS GPA 3.0; ACT 27, HS GPA 3.0); Mississippi (ACT 29, HS GPA 3.5) (Dynarski, 2002; Frisvold & Pitts, 2018; Zhang & Ness, 2010).

Second, many studies focus on the difference between pre-aid and post-aid periods, using the difference-in-differences framework to identify the treatment effect (Cornwell, Lee, et al., 2006; Singell et al., 2006; Sjoquist & Winters, 2015a, 2015b; Zhang, 2011). Since many merit-based financial aid programs were introduced in the 1990s (Dynarski, 2002), the comparison between pre-aid and post-aid periods usually requires a sample of students enrolled in colleges more than 25 years ago, which may be a little bit out-of-date when considering the quick changes on the current STEM labor market. Further, under the difference-in-differences framework, the empirical strategy utilizes the variation of whether a merit aid program was implemented or not. Instead, the method is not required to identify whether a student was eligible or received the merit aid. That could partially explain why researchers are more interested in large merit aid programs so that sizable impacts can be concluded. The effect of small merit aid programs may be easily disturbed by other policy changes so that the effect could be too hard to be observed clearly from a state-level when the difference-in-differences framework is applied. Besides, improper control groups can create biased estimators under the difference-in-differences framework. Sjoquist and Winters (2015b) argued that non-resident students were an imperfect control group. They also only used time difference estimates as a supplementary to difference-in-differences estimates. Multiple robustness checks have to be provided to make their results more convincing. Because the aid eligibility is based on the ACT score, it is a good opportunity to utilize the discontinuities in the aid award formula to evaluate the treatment effect of the Bright Flight scholarship. By using a regression discontinuity framework, my study can directly compare subjects from treatment and control groups in recent cohorts (from the 1996-1997 academic year to the 2006-2007 academic year). Also, compared to the difference-in-differences framework, the regression discontinuity design is a strong quasi-experimental design with local randomization around the policy threshold (if several key assumptions are satisfied). With the unique policy threshold and proper model settings, the causal effect can be easily identified without being disturbed by other policies. Though the regression discontinuity design has become more popular in recent financial aid studies (Castleman et al., 2018; Denning & Turley, 2017; Evans, 2017; Zhang et al., 2016), many of them focus on the need-based aid programs, including the National SMART Grant (Denning & Turley, 2017; Evans, 2017). This paper can be considered as a new application of the regression discontinuity design on merit-aid evaluation.

Third, the gender-differentiated effects of financial aid on major choice do not have highly consistent estimates in these studies. Cornwell, Lee, et al. (2006) compared the effects of HOPE (GA) on major choice by gender and only found that the financial aid affects more positively on females to choose education majors. Zhang (2011) showed both HOPE (GA) and Bright Future (FL) programs had significant positive effects on the percentage of STEM degrees only among male students from private institutions. Sjoquist and Winters (2015a) found that male students experienced a significant decrease in the probability of receiving a STEM major while females did not. Though these scholars have adopted similar difference-in-differences frameworks, their results may still differ by various sample sizes, aid programs, and state contexts. In this paper, the heterogeneity in gender is measured differently through a regression discontinuity approach. The estimates capture the variation among a specific group of individual male and female students who achieve very high ACT scores. This approach can provide new empirical evidence and may increase the robustness of the previous conclusions in the literature. Therefore, I believe my evaluation of a state-funded merit aid's impacts on STEM attainment can be a good supplement to the literature.

Conceptual Framework

The conceptual framework is developed from the theory of consumer choice. Going to college or choosing certain majors can be regarded as a consumption behavior in education. Offering students financial aid could theoretically affect STEM attainment in college on either the extensive (e.g., enrollment) or intensive (e.g., major choice) margins (Castleman et al., 2018). Each margin is a combination of income effect and substitution effect. The income effect expresses the impact of increased purchasing power on consumption, while the substitution effect describes how consumption is changed due to a change in relative prices. On the extensive margin, financial aid can increase schooling investment by reducing budget constraints (income effect) and by decreasing the cost of attendance relative to non-school options (substitution effect). Overall, the combined effect is usually positive. On the intensive margin, the impact of financial aid is not obvious. It is unclear whether the income effect on the intensive margin can encourage or discourage students to study STEM majors. Besides, since the financial aid is not assigned according to specific majors, there should be no substitution effect. However, due to the renewal requirement of some financial aid programs, students study rigorous majors may be more difficult to get renewed. With less expected financial aid, the renewal requirement may increase the price of certain majors and generate a substitution effect. More details about these margins will be discussed later.

In this paper, due to data availability[5], the extensive margin is reframed as the proportion of total enrollment in Missouri public 4-year institutions based on all Missouri ACT takers. Then the intensive margin represents the proportion of students with different majors conditional on being enrolled in Missouri public 4-year institutions. Because of the data limitation, I plan to simplify my analysis and primarily measure the intensive margin. In other words, I only explore the effects on students restricted to Missouri public 4-year institutions. But since many previous studies address the significance of the extensive margin, including general enrollment (Cornwell, Lee, et al., 2006; Dynarski, 2000, 2003) and in-state college choice (Bruce & Carruthers, 2014; Cohodes & Goodman, 2014; Dynarski, 2002; Zhang et al., 2016). The restricted sample can bring selection bias. To ensure the validity of my study, I am going to verify whether the financial aid produces any significant extensive margin at first. Technically, the extensive margin will be examined by using a sample of all Missouri ACT takers. More discussions will be presented in the section of research design.

Theoretical and Empirical Analysis

The effect of intensive margin is a little bit complicated and unclear. Conditional on being enrolled in higher education institutions, in order to choose the proper college major, m, a student is assumed to face the following utility maximization problem:

$$\underset{m=S,N}{U} = U_m(R_m, C_m) \tag{2.1}$$

---

[5] The analytic dataset only has enrollment information of students in public institutions and cannot identify each student's final college choice, either enrolled in in-state private institutions, out-of-state ones, or even not in any colleges. Further, the ACT score is usually not mandated in public 2-year institutions while most of the students in public 4-year institutions have the ACT scores. Thus, the extensive margin in this paper is defined as whether a student enrolls in a Missouri 4-year public institution, which may be slightly different from a more common definition that captures the margin between students who go to college and who do not.

Subject to the constraint:

$$C_m + A_{m,t} + A_{m,t+1} = T_m \tag{2.2}$$

A student will choose a proper major (STEM, S, or non-STEM, N) to maximize his or her utility[6], which is a function of future revenue, R_m, and personal cost, C_m. The personal cost, C_m, initial aid, $A_{m,t}$, and expected renewed aid in the following years, $A_{m,t+1}$, cover the whole cost of major m, T_m. Naturally, I assume that higher future income and lower personal cost increase the utility, which means $\frac{\partial U_m}{\partial R_m} > 0$ and $\frac{\partial U_m}{\partial C_m} < 0$. It is also reasonable to assume that graduates with STEM degrees have higher future income, so $R_S = R_N + DR_S$.

To decide whether to choose a STEM major or not, the utility function can be rewritten like this:

$$\Delta U_S = U_S(R_S, C_S) - U_N(R_N, C_N) \tag{2.3}$$

If $\Delta U_S > 0$, a student is more likely to choose a STEM major otherwise a non-STEM major is preferred.

The Missouri Bright Flight Scholarship Program is likely to affect major choice through an income effect and a substitution effect. Generally, this scholarship is not major-targeted, which means the initial aid is indifferent between STEM and non-STEM majors. Thus, I assume that $A_t = A_{S,t} = A_{N,t}$. I also notice that this scholarship has a renewal GPA requirement. STEM courses are usually more rigorous than non-STEM courses, making it harder to maintain a high GPA. Considering the eligibility of receiving the scholarship in the future, choosing a STEM major reduces the total expected amount of the scholarship that a student can get in the following

---

[6] I assume that male and female students have the same utility function but will test this hypothesis through examining the heterogeneity with empirical data.

years. I assume that $A_{t+1} = A_{S,t+1} + DA_{S,t+1} = A_{N,t+1}$ and the positive difference, $DA_{S,t+1}$, shows the potential difficulties in aid renewal for STEM students.

Therefore, Equation 2.3 is equal to:

$$\Delta U_S = U_S(R_m, T_S - A + DA_{S,t+1}) - U_N(R_N, T_N - A) \tag{2.4}$$

While $A = A_t + A_{t+1} = A_{S,t} + A_{S,t+1} + DA_{S,t+1} = A_{N,t} + A_{N,t+1}$. Also, R_m and T_m can be regarded as exogenous variables and uncorrelated with A and $DA_{S,t+1}$. I assume that the amount of financial aid is not associate with cost and future income.

The income effect could be:

$$\frac{\partial \Delta U_S}{\partial A}\Delta A = (\frac{\partial U_S}{\partial R_S} \cdot \frac{\partial R_S}{\partial A} + \frac{\partial U_S}{\partial C_S} \cdot \frac{\partial C_S}{\partial A} - \frac{\partial U_N}{\partial C_N} \cdot \frac{\partial R_N}{\partial A} - \frac{\partial U_N}{\partial C_N} \cdot \frac{\partial C_N}{\partial A})\Delta A \tag{2.5}$$

Since R_m is not related to A so that $\frac{\partial R_m}{\partial A} = 0$. So the income effect could be simplified as:

$$\frac{\partial \Delta U_S}{\partial A}\Delta A = (\frac{\partial U_S}{\partial C_S} \cdot \frac{\partial (T_S - A + DA_{S,t+1})}{\partial A} - \frac{\partial U_N}{\partial C_N} \cdot \frac{\partial (T_N - A)}{\partial A})\Delta A \tag{2.6}$$

Because $\frac{\partial U_m}{\partial C_m} < 0$, both $\frac{\partial (T_S - A + DA_{S,t+1})}{\partial A}$ and $\frac{\partial (T_N - A)}{\partial A}$ are also negative, making the income effect either positive or negative.

Similarly, the substitution effect could be:

$$\frac{\partial \Delta U_S}{\partial S}\Delta S = (\frac{\partial U_S}{\partial R_S} \cdot \frac{\partial (R_N - DR_s)}{\partial S} + \frac{\partial U_S}{\partial C_S} \cdot \frac{\partial (T_S - A + DA_{S,t+1})}{\partial S})\Delta S \tag{2.7}$$

According to my previous assumptions, I have $\Delta S > 0$, $\frac{\partial U_S}{\partial R_S} > 0$, $\frac{\partial (R_N - DR_s)}{\partial S} > 0$, $\frac{\partial U_S}{\partial C_S} < 0$, and $\frac{\partial (T_S - A + DA_{S,t+1})}{\partial S} > 0$. As a result, the substitution effect is unclear. Choosing a STEM major has two types of substitution effect. The first type is positive due to the higher expected future income of STEM occupations. On the contrary, the second type is negative because it is easier to lose more financial aid in STEM programs and increases the expected cost of STEM majors. As a result, the threat of losing financial aid can encourage students to select non-STEM majors. Theoretically, when combined, the predicted effect of the financial aid program on STEM outcomes is ambiguous.

Empirically, the income effect is also not clear. On one side, the income effect could be positive in choosing STEM majors. The Bright Flight scholarship can cover additional costs of STEM majors, such as higher course fees, laboratory or material fees (American Institutes for Research, 2013; Stange, 2015). With a higher salary expectation in STEM occupations, the financial aid lowers the cost and makes STEM majors much more cost-effective. The cost-effectiveness of STEM degrees also indicates a positive substitution effect. On the opposite side, receiving the Bright Flight scholarship can remove the pressure to pursue majors with higher expected earnings, such as STEM (Andrews & Stange, 2016; Rothstein & Rouse, 2011; Stater, 2011).

Meanwhile, the higher expected income can make STEM majors more attractive and contribute to a positive substitution effect. The expected earning is essential in the choice of a college major (Montmarquette et al., 2002) and the choice of college major is responsive to wage changes (Freeman & Hirsch, 2008). Oppositely, the negative substitution effect caused by the renewal requirement is also mentioned by previous empirical studies. The threat of losing the scholarship may make students choose a safer major to maintain the required GPA (Cornwell, Lee, et al., 2006; Sjoquist & Winters, 2015a). Since STEM courses are usually more rigorous with less GPA inflation, it is harder to maintain the required GPA in STEM fields than other disciplines, which would be a deterrent for recipients to choose STEM majors.

Hypothesis

Based on my discussion above, due to the uncertain effect of financial aid on STEM attainment, I want to test the following competing hypotheses for the intensive margin:

1. The Bright Flight Scholarship Program has a stronger positive income effect with weaker negative income and substitution effects, making the combined effect on STEM attainment become positive.
2. The Bright Flight Scholarship Program has stronger negative income and substitution effects with a weaker positive income effect, making the combined effect on STEM attainment become negative.

The completing hypotheses above has a hidden assumption that all subjects are homogenous. However, male and female students may have inconsistent utility functions so that the income effects and the substitution effects can differ by gender. The combined effect could be gender-differentiated. I also examine the additional completing hypotheses about the existence of the heterogeneity in gender:

1. The Bright Flight Scholarship Program has similar combined effects on both male and female recipients.
2. The Bright Flight Scholarship Program has different combined effects on both male and female recipients.

Research Design

Overview

I use a regression discontinuity approach to identify the treatment effects of the Bright Flight Scholarship Program. In the regression discontinuity design, the running variable is the ACT composite score and the treatment is assigned to students according to whether they can achieve the highest ACT score of 30 or above. Following the practice in the literature (Bruce & Carruthers, 2014; Harrington et al., 2016; Zhang et al., 2016), I apply a framework of intent-to-treat that measures the effect of the aid eligibility. Furthermore, considering the potential manipulation of the highest ACT score, I adopt a fuzzy regression discontinuity design with the first-time ACT score as the primary running variable. The estimation strategy is a two-stage local linear regression specification (2SLS):

The first stage:

$$\begin{aligned} BF\_eligible_i = \alpha Above_i &+ \sum_{j=1}^{p} \delta_j^{-}(ACT\_dist_i)^j * Below_i \\ &+ \sum_{j=1}^{p} \delta_j^{+}(ACT\_dist_i)^j * Above_i + \eta' X_i + \varepsilon_i \end{aligned} \tag{2.8}$$

The second stage:

$$\begin{aligned} Y_i = \beta \widehat{BF\_eligible}_i &+ \sum_{j=1}^{p} \gamma_j^{-}(ACT\_dist_i)^j * Below_i \\ &+ \sum_{j=1}^{p} \gamma_j^{+}(ACT\_dist_i)^j * Above_i + \theta' X_i + \epsilon_i \end{aligned} \tag{2.9}$$

In the equations above, $BF\_eligible_i$ represents the probability of being Bright Flight eligible, which means student i's highest ACT is 30 or above. $Above_i$ or $Below_i$ indicates that the subject's ACT score is above or below the threshold, ACT_dist_i is the distance between the ACT score and the cut-off score, X_i is a vector of student-level control variables.

In the first stage, the probability of being Bright Flight eligible ($BF\_eligible_i$) is estimated based on whether the subject is Bright Flight eligible on the first attempt, functions of the gap between the ACT and the threshold, and a set of control variables. Preferred specifications include linear ($p = 1$) or quadratic functions ($p = 2$) and are allowed to differ on either side of the threshold. In the second stage, the predicted probability of being Bright Flight eligible ($\widehat{BF\_eligible_i}$) is substituted for being Bright Flight eligible ($BF\_eligible_i$) and is used to estimate the effect of the Bright Flight eligibility on STEM/Engineering major choice/degree completion. The key coefficient β can be interpreted as the effect of the treatment on the treated[7]. In my setting, it is the effect of being Bright Flight eligible for those who become Bright Flight eligible based on their first-time ACT composite score.

Data

The analytic dataset is derived from three administrative datasets provided by the Missouri Department of Higher Education (MDHE). The first dataset is derived from Enhanced Missouri Student Achievement Study (EMSAS). It contains first-time, full-time, degree-seeking students from Missouri public colleges and universities, which includes each student's demographic information, high school performance, and college information (major, credits, GPA, etc.). The second dataset has every Missouri ACT taker's historical records, including each one's both highest ACT composite score and first-time ACT composite score. The last dataset provides the information about every recipient of the state's financial aid, such as which state grant he or she receives.

To measure the intensive margin, the analytic sample is restricted to first-time, full-time, degree-seeking students. To avoid inconsistent policies after the 2007-2008 academic year, only students who enrolled in Missouri public 4-year institutions from the 1996-1997 academic year to the 2006-2007 academic year. Students from 2-year community colleges are not considered because many 2-year institutions do not require ACT scores and most of the Bright Flight recipients are enrolled in 4-year public institutions[8]. The whole sample includes about 130,000 students. Table 2.1

---

[7] More accurately, it is the treatment on the intent-to-treat (being aid eligible instead of real treated).

[8] According to Missouri Department of Higher Education (Missouri Department of Higher

Table 2.1

Data Construction

Sample	Total	Lost
First-time, degree-seeking students from Missouri 4-yr public institutions	148,654	
Full-time students	143,316	5,338
With historical ACT data matched	138,356	4,960
With high school course taken and high school class rank	125,883	12,473

shows the process of data construction.

Outcomes and STEM/Engineering Categories

In this paper, there are four outcome variables as the measures of STEM attainment, including whether a student claims an initial major in STEM/engineering fields and whether a student can finish a STEM/engineering degree within six years. The Classification of Instructional Programs (CIP) code stored in the EMSAS dataset is used to identify the specific STEM/engineering programs and degrees. In this study, engineering majors/degrees have three main categories: engineering, computer science, and technology. STEM majors/degrees have additional categories apart from engineering, including agricultural and animal sciences, natural science, biological science, mathematics, military/security science, physical science, psychology, business, social science, and health science (Darolia et al., 2020). For detail information about the definition of each STEM/engineering category and its CIP code, please see Table A.1 in Appendix A.

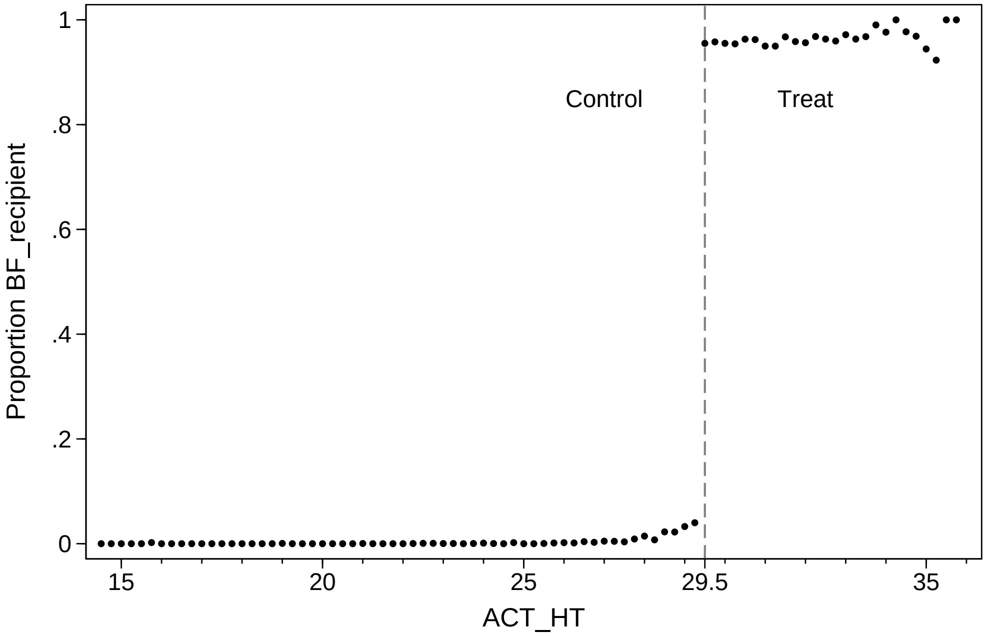
ACT as the Running Variable

The ACT score is used as the main running variable. According to MDHE's policy, students with the highest ACT score of 30 or above possess the Bright Flight eligibility. In my regression discontinuity setting, I use the raw value of the ACT

Education, 2010), in the fiscal year 2007, the total number of the Bright Flight recipients was 8,541, including 6,463 students (75.7%) within the 13 public 4-year universities.

Figure 2.1

The Proportion of Bright Flight Recipient by the Highest ACT Composite Score



composite score, which is the average of four subject scores (English, reading, math, and science), rounded to the nearest whole number. As a result, the new equivalent cut-off point is 29.5. Students with an average score of 29.5 (rounded to 30) are coded as being Bright Flight eligible in the MDHE data system. The big advantage of using this unrounded score is that the interval of the running variable would be reduced to 0.25 instead of 1, making the running variable more "continuous" while the rounded running variables may lead to inconsistent estimates of treatment effects (Dong, 2015) and misleading inference (McCall & Bielby, 2012) due to the larger interval. As a visual examination, Figure 2.1 shows the proportion of the actual Bright Flight recipients by different highest ACT composite score. The huge jump of the proportion indicates the existence of significant discontinuity around the policy threshold though a slight fuzziness also exits.

Control Variables

Ideally, many unobserved sorting behaviors should be controlled under the strong assumptions of regression discontinuity design and there is no need to include X_i in the equations above. But to get more precise estimates with the large bandwidth, following the suggestions from D. S. Lee and Card (2008), I introduce some control variables in my regression discontinuity model, such as students' demographic information (race, gender), high school performance (class rank, math and science course taken), family income level, and the fixed effect of each enrollment year. Descriptive statistics of these covariates are presented in Table 2.2 and Table 2.3.

Sample Selection

One of the biggest threats to the validity of this study is that the restricted sample creates a selection bias. Since the preferred dataset is limited to students enrolled in Missouri public 4-year institutions, it can only measure the intensive margin about major choice and degree completion among those institutions. But in addition to the shifts in the proportion of STEM/engineering majors, producing more STEM/engineering graduates to meet the needs of the labor market is also crucial for policymakers. In other words, policymakers are more interested in finding the impact of this financial aid program on the STEM outcomes in aggregation rather than just the shifts within the public institutions. Intuitively, financial aid programs can increase college enrollment and ignoring the extensive margin may bring bias in evaluating the effects of the Bright Flight Scholarship Program on total STEM attainment.

Before beginning the main empirical analysis, it is necessary to verify my previous argument that the extensive margin in Missouri context is not significant. Theoretically, the extensive margin in this paper can be decomposed into two parts.

Table 2.2

Descriptive Statistics

Variable	All sample		ACT27-31.75	
	Mean	SD	Mean	SD
Engineering major	0.11	0.32	0.22	0.41
STEM major	0.21	0.41	0.37	0.48
Major unclaimed	0.18	0.38	0.11	0.32
Engineering degree in 6 year	0.06	0.24	0.14	0.35
STEM degree in 6 year	0.12	0.33	0.26	0.44
Bright Flight eligibility	0.12	0.32	0.68	0.47
Race and Gender				
Black	0.07	0.25	0.01	0.09
Hispanic	0.01	0.12	0.01	0.11
Asian	0.02	0.14	0.02	0.15
Others	0.03	0.17	0.04	0.19
Female	0.55	0.50	0.46	0.50
HS Performance				
Math course taken	3.96	1.51	4.78	1.82
Science course taken	3.40	1.76	4.32	2.14
HS class rank	72.86	81.11	41.24	58.29
Family Income				
Missing	0.16	0.36	0.17	0.38
Income zero	0.05	0.21	0.03	0.16
Less than $24,000	0.05	0.21	0.03	0.18
$24,000 to $36,000	0.05	0.22	0.04	0.20
$36,000 to $50,000	0.06	0.24	0.05	0.23
$50,000 to $60,000	0.07	0.26	0.07	0.25
$60,000 to $80,000	0.09	0.29	0.09	0.29
$80,000 to $100,000	0.11	0.31	0.11	0.31
$100,000 to $120,000	0.15	0.36	0.17	0.37
$120,000 to $150,000	0.10	0.30	0.11	0.31
Greater than $150,000	0.11	0.32	0.13	0.34
Observation	125,883		14,425	

Note: The ACT refers to the first-time ACT. HS=High School.

Table 2.3

Sample Comparison

Variable	ACT27-31.75		T-value
	Below (< 29.5)	Above ($\geqslant 29.5$)	
Engineering major	0.202	0.251	-6.439***
STEM major	0.356	0.424	-7.535***
Major unclaimed	0.118	0.099	3.251***
Engineering degree in 6 year	0.132	0.171	-5.995***
STEM degree in 6 year	0.240	0.306	-8.149***
Bright Flight eligibility	0.562	1.000	-55.663***
Race and Gender			
Black	0.009	0.005	2.473**
Hispanic	0.012	0.010	1.307
Asian	0.022	0.023	-0.433
Others	0.037	0.045	-2.372**
Female	0.471	0.417	5.824***
HS Performance			
Math course taken	4.698	4.993	-8.716***
Science course taken	4.222	4.564	-8.625***
HS class rank	43.399	35.566	7.223***
Family Income			
Missing	0.165	0.181	-2.329**
Income zero	0.026	0.024	0.814
Less than $24,000	0.034	0.031	0.741
$24,000 to $36,000	0.042	0.040	0.796
$36,000 to $50,000	0.054	0.053	0.135
$50,000 to $60,000	0.068	0.059	1.904*
$60,000 to $80,000	0.091	0.088	0.589
$80,000 to $100,000	0.109	0.105	0.820
$100,000 to $120,000	0.167	0.169	-0.259
$120,000 to $150,000	0.114	0.106	1.355
Greater than $150,000	0.130	0.144	-2.354**
Observation	10,452	3,973	

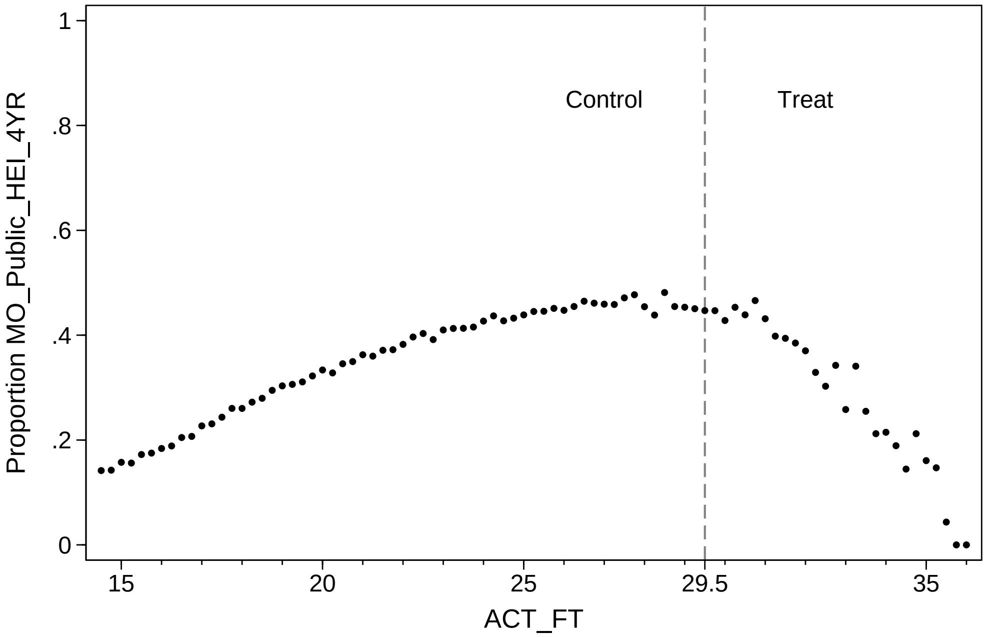
*** $p<0.01$, ** $p<0.05$, * $p<0.10$

First, a state-fund merit-based financial aid increases the total enrollment in any in-state college due to the positive income and substitution effects. The positive effect is also addressed by many previous empirical studies (Cornwell, Lee, et al., 2006; Dynarski, 2000, 2003). Second, it affects students' in-state college choice (Bruce & Carruthers, 2014; Cohodes & Goodman, 2014; Dynarski, 2002; Zhang et al., 2016). Students are more likely to shift to public 4-year institutions since the merit-based financial aid in many states usually can cover the major part of the cost in public 4-year institutions instead of the far more expensive tuition and fees in private ones. The nearly free and more affordable higher education creates a strong substitution effect and encourages students to stay longer in public institutions. For example, in several merit-based aid programs studied in the literature, the percentages of the average amount in tuition and fees of in-state 4-year public universities are 131.19% (Florida), 102.27% (Georgia), and 86.13% (Tennessee) (Frisvold & Pitts, 2018).

Hence, when combined, it is usually expected to find a positive extensive margin, especially in public 4-year institutions. To examine whether the extensive margin in Missouri context exists or not, I utilize the administrative dataset which contains all Missouri ACT takers who graduated from high school between 1996 and 2006. Figure 2.2 shows the proportion of Missouri 4-year public institution enrollment by the first-time ACT score. Visually, there is no jump around the cut-off point, which implies the merit aid may not affect public 4-year institution enrollment. Next, I apply a fuzzy regression discontinuity approach with a similar setting in my primary analysis of the intensive margin to measure the effect of the Bright Flight eligibility on enrollment in public 4-year institutions. Due to the data limitation, the control variables only include demographic information (race, gender) and family income level. Table 2.4 reports the estimates of the extensive margin with different bandwidths and model settings. Panel 1 and Panel 2 report the local linear and quadratic models, separately. Column 1-4 present estimates with different bandwidths and my preferred

Figure 2.2

The Proportion of Missouri 4-year Institution Enrollment by the First-time ACT Composite Score



one is in Column 3 (27-31.5).

However, the empirical results are not compatible with previous studies (Bruce & Carruthers, 2014; Zhang et al., 2016). A possible explanation is that the fuzzy regression discontinuity estimate is the local treatment effect of the students near the threshold and the identification highly relies on the variation of outcome variables (e.g., college enrollment/choice) among compliers. Unlike many other state-funded merit-based aid programs, the Bright Flight scholarship is targeting on Missouri high school graduates with the top 3% ACT scores. It has an ACT threshold of 30 or above, which is significantly higher compared to many merit aids in other states (Frisvold & Pitts, 2018). It may be fair to assume that high school graduates who have the potential to get very high ACT scores are usually more well-prepared and self-motivated. Many of them have already decided to go to colleges (especially 4-

Table 2.4

Fuzzy Regression Discontinuity Estimates of Extensive Margin

	(1) ACT 24.5-34.25	(2) ACT 25.75-33	(3) ACT 27-31.75*	(4) ACT 28.25-30.5
Linear				
BF eligibility	0.002	0.018	0.002	-0.044
	(0.023)	(0.030)	(0.038)	(0.055)
ACTdist_above	-0.046***	-0.041***	-0.022***	-0.003
	(0.003)	(0.004)	(0.007)	(0.017)
ACTdist_below	0.006*	-0.001	-0.005	0.009
	(0.004)	(0.006)	(0.010)	(0.022)
Control variables	X	X	X	X
N	87,670	59,435	36,445	17,227
R-squared	0.008	0.012	0.009	
Quadratic				
BF eligibility	0.000	-0.013	-0.019	0.087
	(0.046)	(0.049)	(0.058)	(0.078)
ACTdist_above	0.001	0.019	0.028	-0.016
	(0.011)	(0.015)	(0.023)	(0.059)
ACTdist_below	-0.015	-0.016	-0.016	-0.131*
	(0.015)	(0.018)	(0.029)	(0.069)
ACTdist_sq_above	-0.012***	-0.019***	-0.024**	0.014
	(0.003)	(0.004)	(0.011)	(0.058)
ACTdist_sq_below	-0.004**	-0.005*	-0.005	-0.077**
	(0.002)	(0.003)	(0.007)	(0.038)
Control variables	X	X	X	X
Observations	87,670	59,435	36,445	17,227
R-squared	0.008	0.005	0.003	0.026

Robust standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.10$

year institutions other than 2-year ones) and finalized their dream colleges. As a result, their decisions are probably not affected by whether receiving the Bright Flight scholarship. In other words, the high policy cut-off point may reduce the variation of college enrollment/choice in compliers and finally makes the extensive margin very small and insignificant.

Therefore, empirically, the financial aid could have no significant impact on increasing enrollment in Missouri public 4-year universities and my previous assumption that the extensive margin is pretty small can still be held. The sample selection does not severely threaten the validity of this study and it is acceptable to exclude the extensive margin and only focuses on the intensive margin in the following analysis.

Assumptions of Regression Discontinuity

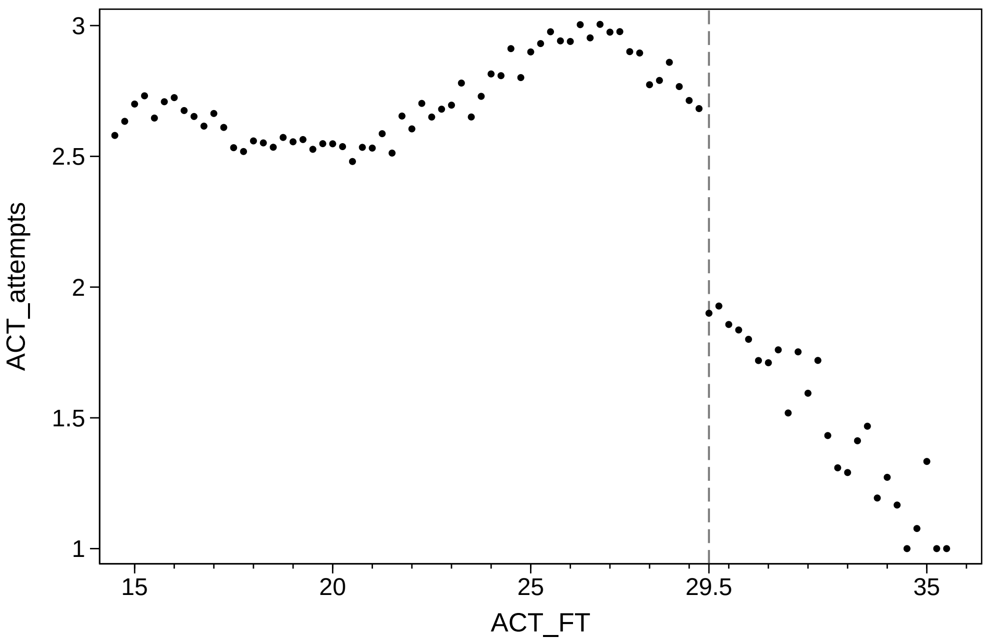
To secure the internal validity of a regression discontinuity design, several assumptions should be satisfied. In this section, I plan to discuss some issues that may threaten the internal validity in this regression discontinuity design and attempt to make my estimates more robust.

Bandwidth Selection

In order to get a more robust estimation, enough observations closed to the cut-off are required in both treatment and control groups, making it very crucial to pick up a proper bandwidth. However, my running variable, the ACT score, is not continuous, which means those commonly used bandwidth selection procedures (Calonico et al., 2014; Imbens & Kalyanaraman, 2012) cannot be applied. In practice, it often requires a least five data points in each group. In related literature, Bruce and Carruthers (2014) used quarter-points of the ACT scores but with 5 whole values in both treatment group (16-21) and control group (21-26). Harrington et al. (2016) chose ACT from 24 to 35 as a bandwidth, with 6 data points in both treatment group (30-35) and control group (24-29). Zhang et al. (2016) preferred ACT from 15 to 24, with 5 data points in both treatment group (15-19) and control group (20-24). Because of the larger interval (1.0) of the ACT score, the bandwidth used in some previous studies are too broader and may jeopardize the assumptions of the regression discontinuity design. Fortunately, similar to Bruce and Carruthers (2014), my ACT score is not rounded and has a smaller interval of 0.25 instead of 1. But due to the very high ACT threshold, data points are limited above the cut-off. Hence, I narrow

Figure 2.3

The Average Number of Retakes by the First-time ACT Composite Score



down my preferred bandwidth ranging from 27 to 31.75, with 10 data points in each group. Other bandwidths for robust check include 5 (28.25-30.5), 15 (25.75-33), and 20 (24.5-34.25) data points on each side.

Manipulation of the Running Variable

To get an unbiased estimation through a regression discontinuity design, the key assumption is that the running variable cannot be manipulated. Unfortunately, my running variable, the highest ACT composite score, can be easily controlled through retaking the ACT exams multiple times. Students may retake the ACT if they cannot get enough points at the first attempt. The retaking behaviors can result in a manipulation of the running variable and create biased estimates. To exam the existence of manipulation, the following figures are provided for visual inspection (Harrington et al., 2016; McCrary, 2008). Figure 2.3 presents the average number of retakes by the first-time ACT score a student received. It implies that the average number of retakes increased until passing the cut-off point of the Bright Flight eligibility and then subsequently declined. Figure 2.4 shows the density distribution of the highest ACT composite score. It is obvious to notice a density increase around the cut-off point (29.5). Figure 2.5 shows the density distribution of the first-time ACT composite score. As a comparison, there is no significant change around the cut-off point in the distribution of the first-time ACT score and the density function is more likely to be continuous. To sum up, I can conclude that the highest ACT score is manipulated and cannot be used as the running variable in a regression discontinuity design. Students who really want to get the financial aid would probably retake the ACT exams multiple times, which creates an endogenous issue.

The first-time ACT score seems to a better option and is adopted by many empirical papers (Bruce & Carruthers, 2014; Harrington et al., 2016; Welch, 2014; Zhang et al., 2016). Manipulation of a student's first-time ACT is unlikely because the ACT is a standardized test and students are supposed to try to score as high as they can. Figure 2.6 presents the proportion of Bright Flight recipients by different first-time ACT composite scores. Even if there is no sharp discontinuity, a significant jump in the probability of receiving the Bright Flight scholarship around the threshold still exists. In this case, a fuzzy regression discontinuity framework has a more proper fit with the dataset and it can solve the noncompliance bias through the two-stage procedure mentioned above.

Continuity of the Outcome-Forcing Variable Relationship

In an ideal regression discontinuity design, the covariates should be very similar around the threshold. Any significant jumps of other covariates at the cut-off point should not be expected. Then similar to the practice in the literature (Bruce & Carruthers, 2014; Harrington et al., 2016), I re-estimate the effect of the Bright Flight eligibility on control variables using the same model but replacing the dependent variables with each covariate. Table 2.5 reports the fuzzy regression discontinuity

Figure 2.4

Histogram of the Highest ACT Composite Score

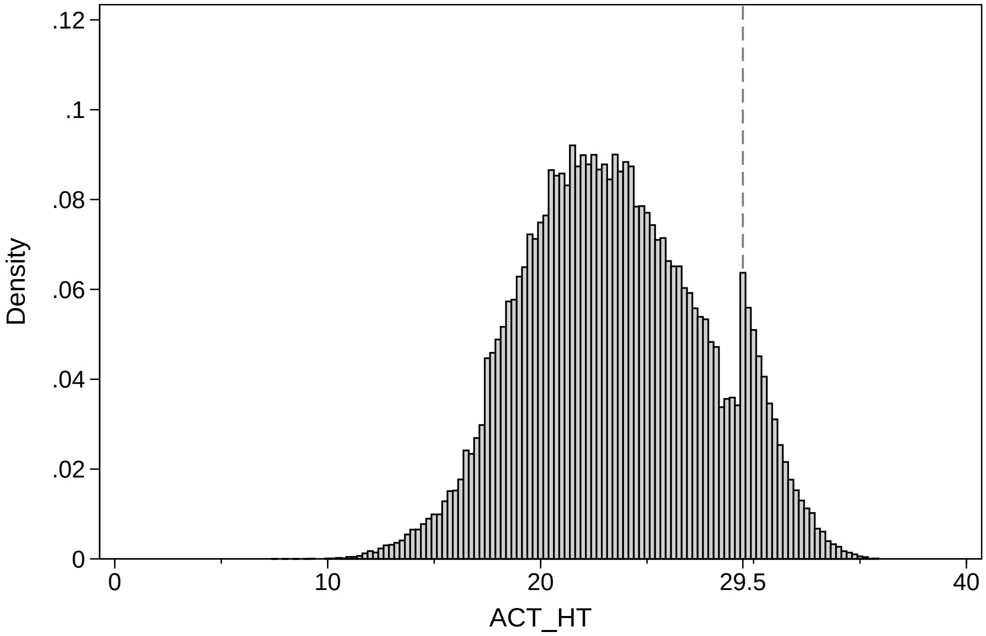


Figure 2.5

Histogram of the First-time ACT Composite Score

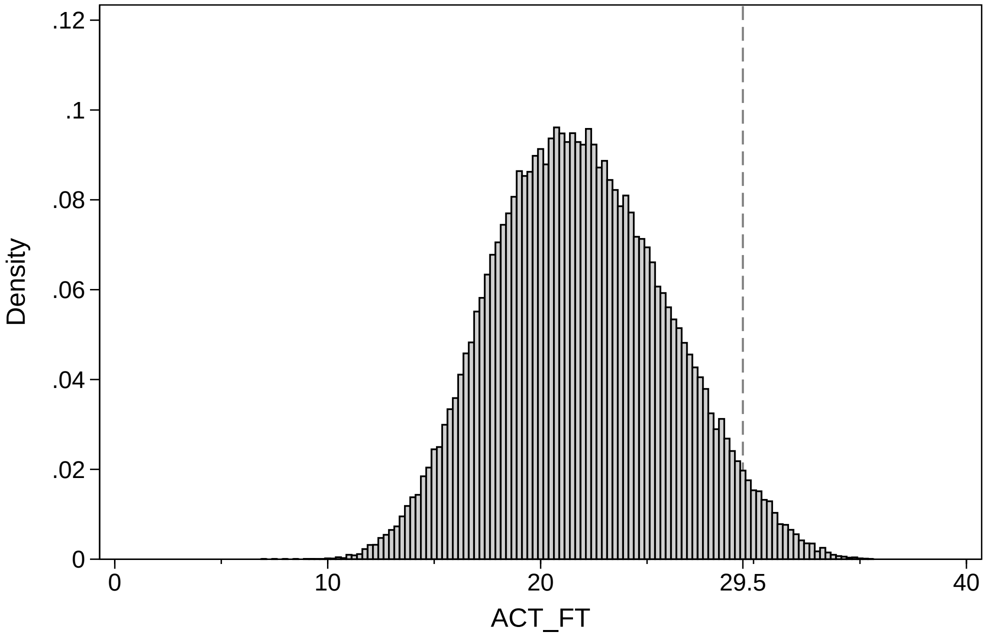
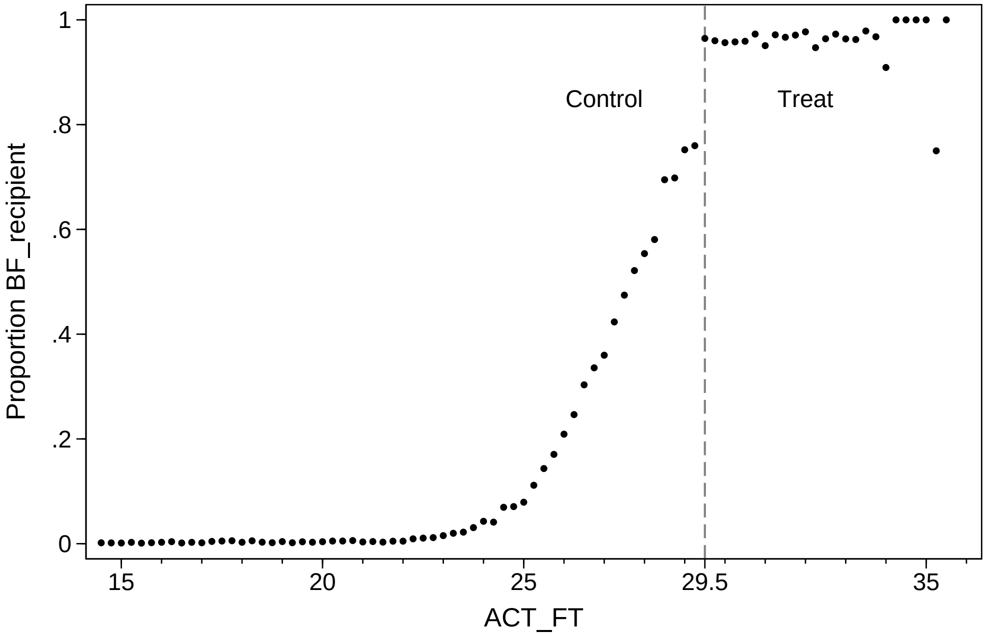


Figure 2.6

The Proportion of Bright Flight Recipient by the First-time ACT Composite Score



estimates of the aid eligibility on control variables. Most control variables have no significant jump around the cut-off point. I believe that including these covariates do not strongly affect the fuzzy regression discontinuity estimates.

Results

Summary Statistics

Table 2.2 presents descriptive statistics for my preferred analytic sample with the first-time ACT score between 27 and 31.75. Compared to the full sample, my selected sample has more students choosing STEM/engineering majors, higher STEM/engineering completion rates. In race and gender compositions, my selected sample has fewer minorities and female students. Particularly, the percentage of black students drops from 7% to 1%. In high school performance, students in my analytic sample usually have finished more math and science courses with better class ranks.

Table 2.5

Treatment Effect on Control Variables

	(1) ACT 24.5-34.25	(2) ACT 25.75-33	(3) ACT 27-31.75*	(4) ACT 28.25-30.5
Race and Gender				
Black	0.008	0.007	-0.002	-0.004
	(0.012)	(0.017)	(0.023)	(0.030)
Hispanic	-0.014	-0.018	0.007	-0.021
	(0.014)	(0.020)	(0.026)	(0.033)
Asian	-0.029	-0.036	-0.052	-0.056
	(0.018)	(0.026)	(0.034)	(0.046)
Others	-0.000	-0.018	-0.047	-0.098
	(0.027)	(0.037)	(0.049)	(0.067)
Female	-0.065	-0.075	-0.134	-0.130
	(0.066)	(0.093)	(0.123)	(0.164)
HS Performance				
Math course taken	0.341	0.291	0.686	0.192
	(0.246)	(0.350)	(0.457)	(0.608)
Science course taken	0.383	0.381	0.920*	0.700
	(0.287)	(0.408)	(0.533)	(0.708)
HS class rank	5.394	5.142	3.645	-2.607
	(7.538)	(10.612)	(14.040)	(18.632)
Family Income				
Missing	0.066	0.046	0.021	0.008
	(0.050)	(0.071)	(0.093)	(0.124)
Income zero	-0.007	0.007	-0.010	-0.009
	(0.020)	(0.029)	(0.039)	(0.052)
Less than $24,000	0.022	0.041	0.014	0.036
	(0.024)	(0.035)	(0.046)	(0.060)
$24,000 to $36,000	-0.009	-0.023	-0.043	-0.116*
	(0.026)	(0.036)	(0.048)	(0.064)
$36,000 to $50,000	0.002	0.001	-0.003	0.056
	(0.030)	(0.042)	(0.055)	(0.071)
$50,000 to $60,000	-0.040	-0.044	-0.059	-0.066
	(0.033)	(0.046)	(0.062)	(0.083)
$60,000 to $80,000	-0.002	0.012	0.022	0.114
	(0.038)	(0.054)	(0.071)	(0.096)
$80,000 to $100,000	-0.041	-0.063	-0.069	-0.004
	(0.041)	(0.059)	(0.077)	(0.105)
$100,000 to $120,000	0.016	-0.026	-0.046	-0.059
	(0.049)	(0.070)	(0.092)	(0.123)
$120,000 to $150,000	-0.044	-0.033	0.045	-0.002
	(0.041)	(0.058)	(0.076)	(0.101)
Greater than $150,000	0.037	0.081	0.129	0.044
	(0.046)	(0.065)	(0.086)	(0.115)

Robust standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.10$

Due to the high policy threshold, the analytic data is very different from the full sample, and it should be very careful to interpret the conclusions when being generalized to the full sample.

In the regression discontinuity design, students around the cut-off point are usually assumed to have similar characteristics in order to be comparable. Table 2.3 compares my preferred sample above and below the cut-off point. There are fewer black students and female students above the cut-off point. And students who pass the minimum ACT requirement at the first time are those who finish more high school credits in math and science. Differently, some merit-based scholarships in other states have an additional high school GPA requirement. As a result, high school students can be motivated to choose easier courses to maintain a higher high school GPA. But in Missouri, the only requirement on ACT composite score may encourage students to do well in all four subjects instead of giving up more difficult math or science courses. There is supposed to be no correlation between high school course-taking and the Bright Flight eligibility.

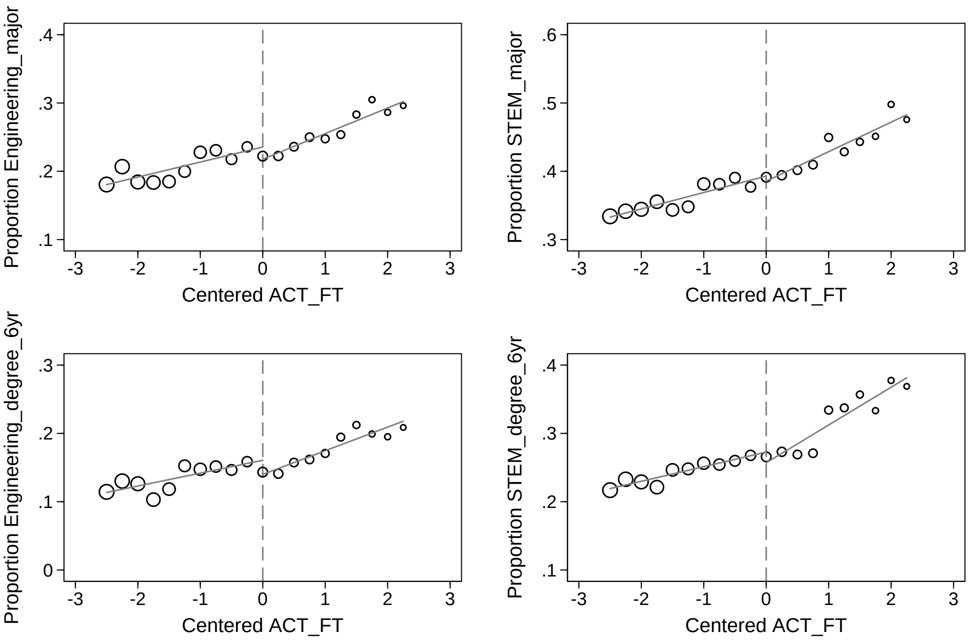
Usually, it is not necessary to add any control variables under the assumption of the regression discontinuity design. However, the discrete running variable requires a relatively larger bandwidth, leading to bigger differences in gender, race, high school performance, and family income between treatment and control groups. Luckily, as Table 2.5 reported, these control variables are more likely to be continuous around the threshold, fitted with the local linear model. The differences between treatment and control groups can be controlled through including these covariates and it also does not bias the fuzzy regression discontinuity estimates. Therefore, I put those variables as control variables in my primary models to get more precise estimates.

Fuzzy Regression Discontinuity Estimates

Figure 2.7 illustrates the regression discontinuity for students who enrolled in Missouri 4-year institutions. Visually, students from the treatment group are more

Figure 2.7

Regression Discontinuity Estimates (Linear)



likely to choose non-engineering majors or finish their college with non-engineering degrees in six years. More precise estimates with covariates using the fuzzy regression discontinuity approach are reported in Table 2.6 and Table 2.7.

Table 2.6 and Table 2.7 display the fuzzy regression discontinuity estimates with the linear probability model. Only robust standard errors instead of errors clustered by the ACT are being reported[9]. Table 2.6 includes the results of the first-stage regression. The first-stage result is identical among the four outcome variables since there is no change of both dependent and independent variables in the first-

---

[9] As Kolesár and Rothe (2018) pointed out recently, if the running variable was discrete, the commonly used confidence intervals based on standard errors that were clustered by the running variable had poor coverage properties. In my scenario, the clustered errors are much smaller than robust standard errors. A possible explanation is that my setting only has a few clusters (maximum number is about 40) and can over-reject the hypothesis (Cameron et al., 2008). I believe my running variable with a quarter-point is more "continuous" and has richer support than the rounded ACT used in previous papers (Harrington et al., 2016; Zhang et al., 2016). Hence, as suggested by the literature (Kolesár & Rothe, 2018), since my bandwidth is smaller and contains fewer clusters, it is better to use the robust standard error to make the proper statistical inference.

stage regression model. The four panels in Table 2.7 represent the estimates of my four outcome variables. To make my results more robust, in Column 1-4, I provide estimates of the effect of being the Bright Flight eligible with multiple bandwidths, ranging from 24.5 to 34.25.

The first-stage result shows the strong correlation between being Bright Flight eligible finally and being Bright Flight eligible with the first-time ACT score. Estimates of the F-test among all bandwidths are above 20, indicating that the first stage is creating good predictions for the second stage. The second stage contains the coefficients of the effect of the Bright Flight eligibility on STEM/engineering major choice or degree completion. In general, there is no positive effect on all four outcome variables across different bandwidths. In my preferred bandwidth, Column 3 (27-31.75), the results imply that the Bright Flight eligibility can reduce the probability of choosing STEM/engineering majors by 9.5% /16.0% and reduce the completion rate of STEM/engineering degrees in 6 years by 13.8%/16.6%. These coefficients show the negative treatment effect on all four outcomes and only weak statistical significance is found on engineering degree completion.

Though the stronger statistical significance does not exist among all models, the estimates demonstrate considerable practical significance that cannot be ignored (10%-15%). Besides, the large standard errors make the estimates to be very imprecise. It is probably caused by the disadvantages of the fuzzy regression discontinuity approach (or the instrument variable). The fuzzy regression discontinuity provides an estimation of the local average treatment effect (LATE) and only utilizes the compliers around the cut-off point to identify the effect. The compliers are captured by the instrument variable which, in my case, is the Bright Flight eligibility based on the first-time ACT. The first-stage result is statistically significant, but the coefficient of the Bright Flight eligibility on the first-time ACT is only 13.9%, indicating that the percentage of the compliers around the threshold is very small. The small percent-

Table 2.6

Fuzzy Regression Discontinuity Estimates with Different Bandwidths (First Stage)

	(1) ACT 24.5-34.25	(2) ACT 25.75-33	(3) ACT 27-31.75*	(4) ACT 28.25-30.5
First stage (Y=Bright Flight eligibility)				
Above	0.200***	0.157***	0.139***	0.149***
	(0.006)	(0.008)	(0.010)	(0.016)
ACTdist_above	-0.006***	-0.008***	-0.006***	-0.006
	(0.001)	(0.001)	(0.002)	(0.004)
ACTdist_below	0.151***	0.172***	0.185***	0.175***
	(0.002)	(0.003)	(0.006)	(0.019)
Control variables	X	X	X	X
N	34,329	23,453	14,425	6,733
R-squared	0.433	0.359	0.281	0.191
F-test	4123	1236	337	62.36
p-value	0	0	0	0

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

age of the compliers may result in insufficient statistical power and large standard errors. In addition, these compliers are less likely to retake the ACT[10] and probably from low-income families. Table 2.8 compares the retaking times and other control variables between the “always-takers” and the “compliers” when their first-time ACT score is just below the threshold[11]. It shows that “compliers” are more likely to be minority and female students from low-income families. They also take the ACT with fewer times and are less well-prepared in high school math and science courses even though they perform pretty well in the ACT. So the “compliers” can be more sensitive to financial incentives and less prepared in STEM fields when compared

---

[10] Intuitively, if students retake the ACT multiple times, they have a higher probability to be eligible finally though their first-time ACT is below the threshold. They will be noted as the “always-takers” in the fuzzy regression discontinuity design. This indicates the negative correlation between retaking behaviors and compliance.

[11] Compliers may usually refer to students who never retake the test and be compliant with the policy. However, in my case, ineligible students cannot make themselves become 100% aid eligible through retaking, indicating that some of them may still be “compliant” with the policy after additional attempts. To be clear, conditioned on students who are not aid eligible after the first attempt, I define “compliers” to be students who are never aid eligible and “always-takers” to be students who are aid eligible finally.

Table 2.7

Fuzzy Regression Discontinuity Estimates with Different Bandwidths (Second Stage)

	(1) ACT 24.5-34.25	(2) ACT 25.75-33	(3) ACT 27-31.75*	(4) ACT 28.25-30.5
Second stage (Y=engineering major choice)				
BF eligibility	-0.072	-0.124*	-0.160	-0.199
	(0.053)	(0.075)	(0.101)	(0.136)
ACTdist_above	0.029***	0.029***	0.031***	0.032
	(0.006)	(0.006)	(0.009)	(0.022)
ACTdist_below	0.026***	0.040***	0.048**	0.056
	(0.009)	(0.015)	(0.022)	(0.039)
Control variables	X	X	X	X
N	34,329	23,453	14,425	6,733
R-squared	0.103	0.082	0.074	0.077
Second stage (Y=STEM major choice)				
BF eligibility	-0.076	-0.116	-0.095	-0.162
	(0.062)	(0.088)	(0.119)	(0.160)
ACTdist_above	0.034***	0.036***	0.036***	0.051*
	(0.006)	(0.007)	(0.011)	(0.026)
ACTdist_below	0.032***	0.041**	0.034	0.052
	(0.010)	(0.017)	(0.026)	(0.045)
Control variables	X	X	X	X
N	34,329	23,453	14,425	6,733
R-squared	0.085	0.066	0.075	0.061
Second stage (Y=engineering degree completion in 6 years)				
BF eligibility	-0.076	-0.122*	-0.166*	-0.129
	(0.046)	(0.065)	(0.088)	(0.117)
ACTdist_above	0.027***	0.025***	0.027***	0.026
	(0.005)	(0.006)	(0.008)	(0.019)
ACTdist_below	0.023***	0.035***	0.046**	0.022
	(0.008)	(0.013)	(0.020)	(0.034)
Control variables	X	X	X	X
N	34,329	23,453	14,425	6,733
R-squared	0.059	0.038	0.022	0.056
Second stage (Y=STEM degree completion in 6 years)				
BF eligibility	-0.070	-0.131	-0.138	-0.133
	(0.057)	(0.081)	(0.109)	(0.146)
ACTdist_above	0.039***	0.038***	0.046***	0.045*
	(0.006)	(0.007)	(0.010)	(0.024)
ACTdist_below	0.028***	0.043***	0.040*	0.036
	(0.010)	(0.016)	(0.024)	(0.041)
Control variables	X	X	X	X
N	34,329	23,453	14,425	6,733
R-squared	0.065	0.040	0.040	0.052

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

to the “always-takers”. As a result, the Bright Flight scholarship may cause a more significant impact on their STEM decisions. When the fuzzy regression discontinuity approach applied, the estimates can be magnified because the sample is restricted to the compliers. This explains why the fuzzy regression discontinuity estimates can be practically significant but not pass the statistical significance test. Overall, the results should be carefully addressed to policymakers with notifications about the limited external validity provided by the fuzzy regression discontinuity approach.

Heterogeneity in Gender

In this section, I evaluate the impact of the Bright Flight Scholarship Program by gender. Considering the difference in preference and the academic preparation of STEM/engineering majors (Griffith, 2010; Price, 2010), female students may react to financial aid differently from their male peers. It may be an alternative option to adopt merit-based financial aid to encourage female students to choose STEM/engineering majors. Similar to previous procedures, Figure 2.8 and Figure 2.9 provide a visual examination and Table 2.9 reports the fuzzy regression discontinuity results based on subsamples grouped by gender with the consistent bandwidths in previous tables.

Interestingly, financial aid may have gender-differentiated effects. Unlike Figure 2.7, Figure 2.8 implies very weak but positive effects of financial aid and Figure 2.9 shows more negative and significant effects on male students. Coefficients in Table 2.9 are also correspondent to those figures. Based on the preferred bandwidth in Column 3, the financial aid eligibility can discourage male students from choosing STEM/engineering fields significantly and dramatically. Meantime, female students may be more likely to participate in STEM/engineering programs slightly by being financial aid eligible. Specifically, for male students, the treatment has a strong negative marginal effect on choosing STEM/engineering majors (26.7%/34.5%) and completing STEM/engineering degrees (37.4%/31.6%), which probably contribute to the majority part of the general negative effects reported in Table 2.7.

Table 2.8

"Always-takers" and "Compliers"

	ACT27-29.25		T-value
	Always-takers	Compliers	
Retaking times	3.121	2.516	25.734***
Race and Gender			
Black	0.004	0.017	-6.857***
Hispanic	0.010	0.016	-2.497**
Asian	0.024	0.019	1.551
Others	0.033	0.041	-2.068**
Female	0.461	0.484	-2.317**
HS Performance			
Math course taken	4.901	4.439	13.208***
Science course taken	4.497	3.870	15.266***
HS class rank	32.033	57.972	-22.663***
Family Income			
Missing	0.165	0.165	0.024
Income zero	0.020	0.034	-4.252***
Less than $24,000	0.030	0.038	-2.378**
$24,000 to $36,000	0.038	0.048	-2.390**
$36,000 to $50,000	0.050	0.059	-2.115**
$50,000 to $60,000	0.068	0.067	0.166
$60,000 to $80,000	0.096	0.084	2.038**
$80,000 to $100,000	0.114	0.104	1.597
$100,000 to $120,000	0.174	0.158	2.187**
$120,000 to $150,000	0.111	0.117	-1.082
Greater than $150,000	0.134	0.124	1.369
Observation	5,872	4,580	

*** $p<0.01$, ** $p<0.05$, * $p<0.10$

Note: This is a rough comparison between "always-takers" and "compliers". As noted, "always-takers" refer to students who are not eligible after the first attempt but are finally eligible after retaking. Opposite, "compliers" refer to students who are always ineligible. In the control group, fuzzy regression discontinuity only uses the compliers while sharp regression discontinuity uses the whole sample. The difference in the control group can probably explain the magnified fuzzy regression discontinuity estimates.

Figure 2.8

Regression Discontinuity Estimates of Female Student (Linear)

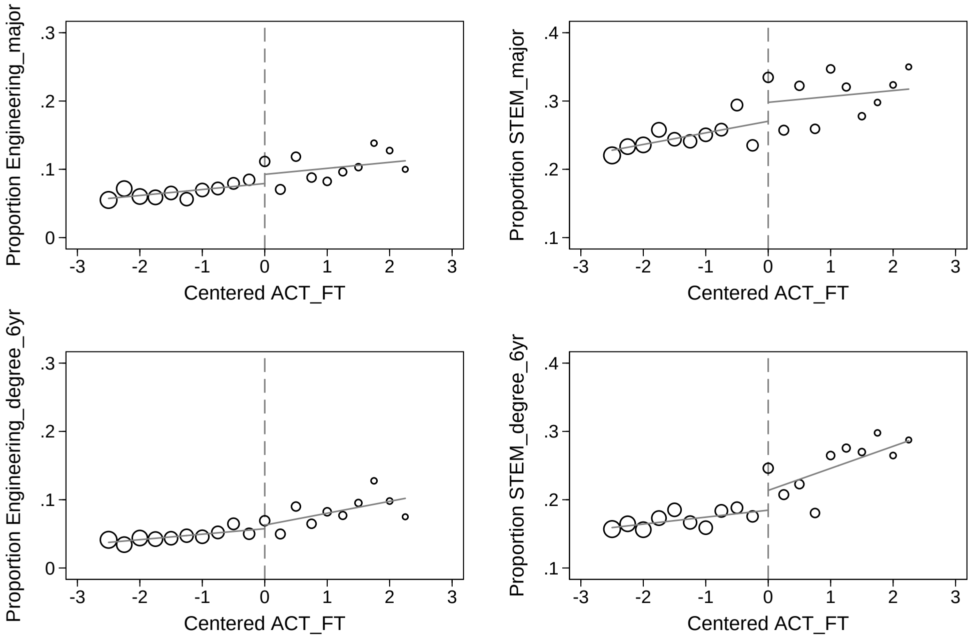


Figure 2.9

Regression Discontinuity Estimates of Male Students (Linear)

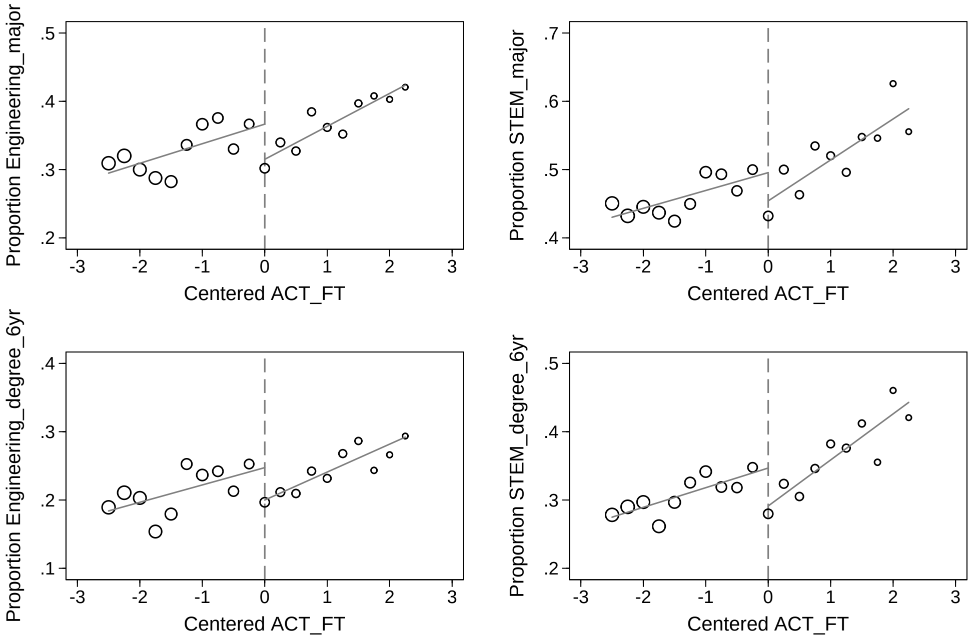


Table 2.9

Fuzzy Regression Discontinuity Estimates by Gender

	(1) ACT 24.5-34.25	(2) ACT 25.75-33	(3) ACT 27-31.75*	(4) ACT 28.25-30.5
Second stage (Y=engineering major choice)				
BF eligibility (Female)	0.045	0.051	0.080	0.043
	(0.054)	(0.079)	(0.106)	(0.150)
BF eligibility (Male)	-0.175**	-0.280**	-0.345**	-0.371*
	(0.087)	(0.122)	(0.164)	(0.213)
Second stage (Y=STEM major choice)				
BF eligibility (Female)	0.013	0.076	0.119	0.153
	(0.084)	(0.121)	(0.165)	(0.233)
BF eligibility (Male)	-0.145	-0.281**	-0.267	-0.405*
	(0.091)	(0.128)	(0.170)	(0.225)
Second stage (Y=engineering degree completion in 6 years)				
BF eligibility (Female)	0.040	0.018	0.029	-0.009
	(0.047)	(0.068)	(0.090)	(0.124)
BF eligibility (Male)	-0.186**	-0.251**	-0.316**	-0.211
	(0.077)	(0.107)	(0.143)	(0.183)
Second stage (Y=STEM degree completion in 6 years)				
BF eligibility (Female)	0.065	0.103	0.151	0.176
	(0.077)	(0.109)	(0.149)	(0.210)
BF eligibility (Male)	-0.186**	-0.330***	-0.374**	-0.356*
	(0.085)	(0.120)	(0.160)	(0.208)

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

Further, to explore the substitute in non-STEM majors for both female and male students, I pick three representative non-STEM majors with the most observations under my preferred bandwidth (ACT 27-31.75): Journalism (CIP2 09), interdisciplinary studies (CIP2 30), and business (CIP2 52)[12]. Using the same procedure, I compare the effect of the Bright Flight scholarship on choosing those majors. Figure 2.10 and 2.11 indicate that the financial aid has positive effects on choosing journalism- and business-related majors for male students while the aid also discour-

---

[12] See Appendix A for more information about CIP codes.

ages female students from entering those non-STEM fields.

To sum up, similar to Sjoquist and Winters (2015a), I also find that the financial aid has a stronger negative effect on male students' STEM major choice. But the effects are slightly varied among female students. Though the estimates of females are also not statistically significant, the financial aid has a positive effect on female students' STEM/engineering major choice and degree completion. Besides, business majors become an important alternative option for male students when they plan to quit the STEM majors. To be integrated with the literature, a possible explanation is that the financial aid may have a weak substitution effect on female students. Women study substantially more than men (Arcidiacono et al., 2012; Stinebrickner & Stinebrickner, 2004), so that female recipients are more likely to maintain a higher GPA because of their efforts[13]. Females are also less sensitive to future net earnings (Freeman & Hirsch, 2008; Montmarquette et al., 2002; Wiswall & Zafar, 2015). The salary premium of STEM degrees may only generate a little positive effect on female students. So without being worried about aid renewal and future net income, the financial aid has positive but insignificant effects on the STEM outcomes of female students. Conversely, male students may be more influenced by the GPA requirement and the threat of losing the aid determines a stronger substitution effect and make the combined effect negative. Several non-STEM majors, such as journalism and business, can be better options for male recipients.

---

[13] A recent study indicates that the harsher grading policies in STEM courses disproportionately affect women (Ahn et al., 2019). Since female students usually get better scores with more study time, the harsher grading policies make taking STEM courses become less cost-effective and discourage females more than males to choose STEM majors. Their results may have some conflicts with my explanation. Unlike their general equilibrium analysis, my estimates are the local treatment effects of students with very high ACT scores. These students are well-prepared for colleges and should have higher GPAs than average. Roughly estimated in my analytic data, among students in the preferred bandwidth (first-time ACT 27-31.75), female students' first-year accumulative GPAs would be more than 3.3 while their male peers' GPAs are only about 3.0. Though the accumulative GPA is a post-choice estimation and may have already reflected the consequence of the rigorous STEM courses, considering the big gap, I think these female students with high ACT scores should not feel concerned about the financial aid renewal or cost-effectiveness of the STEM majors.

Figure 2.10

Effect of Bright Flight on Choosing Non-STEM Majors (Female Students)

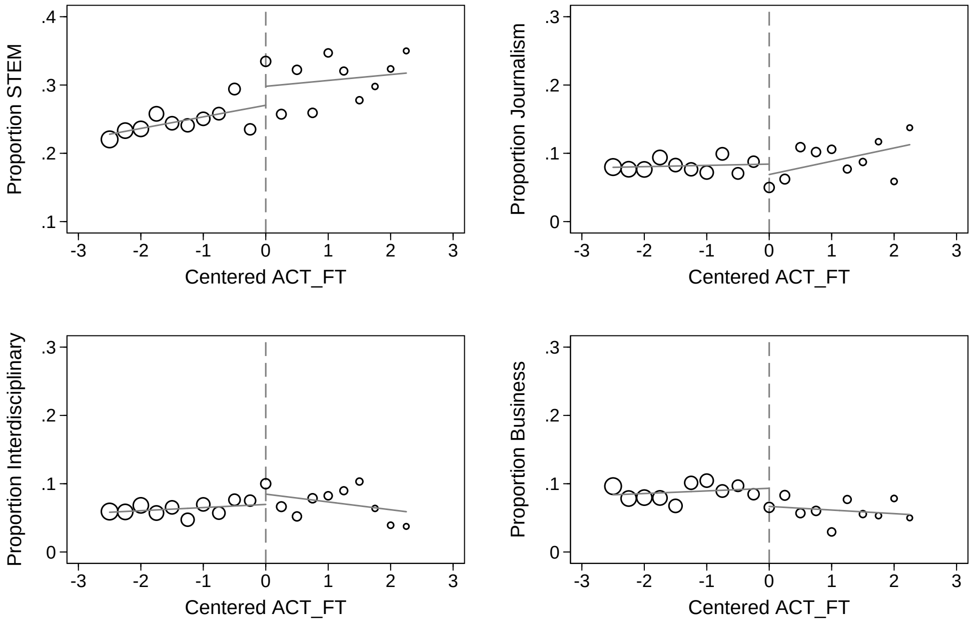
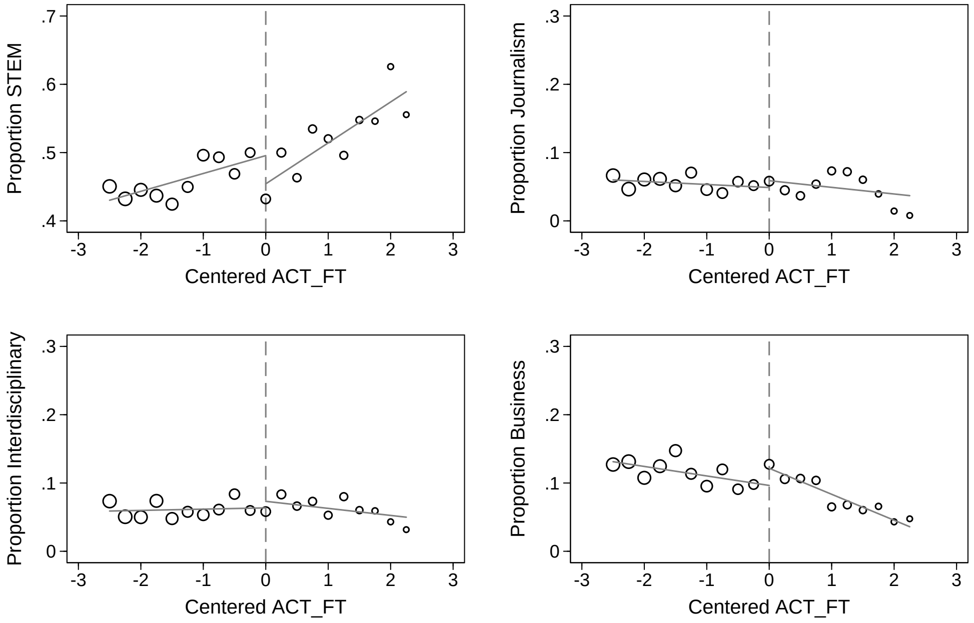


Figure 2.11

Effect of Bright Flight on Choosing Non-STEM Majors (Male Students)



Sensitivity

Alternative Functional Forms

Choosing a proper function form in a parametric regression discontinuity design is very important since the misspecification of the functional form typically generates a bias in the treatment effect (D. S. Lee & Lemieux, 2010). Table 2.10 reports an additional robustness check with alternative functional forms. First, Column 1 is the local linear specification, which is the same as the one in Column 3, Table 2.7. Second, Column 2 is the local quadratic regression specification. Both specifications do not include slope restrictions. Other high-order polynomial functions are not reported, following the guidance of Gelman and Imbens (2019).

Similarly, the alternative specification also points out the negative effects on all four outcome variables while the quadratic specification provides stronger significant estimates on engineering major choice and degree completion. However, the coefficients of quadratic terms presented in Column 2 are all very small and statistically insignificant while the coefficients of linear terms in Column 1 are more significant. The scatter plots in Figure 2.7 also do not strongly indicate the existence of the quadratic term. Therefore, I believe the linear functional form with covariates I have applied is acceptable and properly specified.

Sorting in Public Institutions

The STEM/engineering programs in Missouri public 4-year institutions are not evenly distributed. Noticed that the University of Missouri-Columbia and Missouri University of Science and Technology are the two biggest STEM/engineering providers in Missouri, the major choice could be associated with college choice. For example, if students choose Missouri University of Science and Technology, they are more likely to choose engineering majors because of the fewer non-engineering programs that Missouri University of Science and Technology can provide. It is necessary to verify

Table 2.10

Different Functional Forms

	Eengineering		STEM	
	(1) Linear	(2) Quadratic	(1) Linear	(2) Quadratic
Second stage (Y=engineering/STEM major choice)				
BF eligibility	-0.160	-0.295**	-0.095	-0.156
	(0.101)	(0.138)	(0.119)	(0.158)
ACTdist_above	0.031***	0.029	0.036***	0.032
	(0.009)	(0.031)	(0.011)	(0.036)
ACTdist_below	0.048**	0.109**	0.034	0.063
	(0.022)	(0.044)	(0.026)	(0.050)
ACTdist_sq_above		0.001		0.002
		(0.015)		(0.017)
ACTdist_sq_below		0.012		0.006
		(0.008)		(0.009)
Control variables	X	X	X	X
N	14,425	14,425	14,425	14,425
R-squared	0.074	0.001	0.075	0.056
Second stage (Y=engineering/STEM degree completion in 6 years)				
BF eligibility	-0.166*	-0.269**	-0.138	-0.163
	(0.088)	(0.119)	(0.109)	(0.145)
ACTdist_above	0.027***	0.041	0.046***	0.047
	(0.008)	(0.027)	(0.010)	(0.034)
ACTdist_below	0.046**	0.084**	0.040*	0.051
	(0.020)	(0.038)	(0.024)	(0.046)
ACTdist_sq_above		-0.007		-0.000
		(0.013)		(0.016)
ACTdist_sq_below		0.007		0.002
		(0.007)		(0.008)
Control variables	X	X	X	X
N	14,425	14,425	14,425	14,425
R-squared	0.022		0.040	0.029

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.10

Table 2.11

Treatment Effect on Public Institution Choice

	(1) ACT 24.5-34.25	(2) ACT 25.75-33	(3) ACT 27-31.75*	(4) ACT 28.25-30.5
Truman	0.026	0.032	0.007	0.069
	(0.034)	(0.047)	(0.064)	(0.085)
UMC	0.044	0.029	0.036	-0.049
	(0.064)	(0.090)	(0.121)	(0.164)
UMR	-0.053	-0.074	-0.073	-0.089
	(0.043)	(0.060)	(0.080)	(0.108)

Robust standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.10$

the empirical evidence if the major choice is related to public college choice.

Since the high policy threshold of the Bright Flight Scholarship Program, many Bright Flight recipients are enrolled in the following three public selective institutions: the University of Missouri-Columbia (UMC), Missouri University of Science and Technology (UMR), and Truman State University (Truman). Considering that Truman State University only hosts a few STEM/engineering programs, the decreasing STEM/engineering attainment may be explained by the increasing enrollment at Truman State University. Furthermore, Truman State University has an additional institutional scholarship that awards students with a top 3% ACT score, which has a similar target group of the Bright Flight Scholarship Program[14]. Those merit aids may overestimate the effects of the Bright Flight aid. To exclude those potential compounding effects, Table 2.11 reports the fuzzy regression discontinuity estimates of the Bright Flight eligibility, replacing the outcome variables with public institution choice. Most of the estimates are not statistically significant and indicate that there is no sorting in public college related to the Bright Flight eligibility.

---

[14] The recent information implies that this scholarship only has twelve awards annually and I believe it may have few impacts on my previous analysis. See https://www.truman.edu/admission-cost/cost-aid/scholarships/competitive-scholarships/.

Conclusion

My empirical result is similar to Sjoquist and Winters (2015a), the Bright Flight financial aid has no positive effect on STEM/engineering major choice or degree completion. Particularly, male students are more likely to be negatively impacted by the merit aid in choosing STEM/engineering majors. As Sjoquist and Winters (2015a) argued, the mechanism beyond merit-based financial aid and major choice is still unclear. Additional qualitative research may be helpful to explore the mechanism in detail. My study implies that non-targeted financial aid cannot promote postsecondary STEM outcomes for policymakers.

Additionally, the gender-differentiated effects of the financial aid provide an interesting policy implication. These results may encourage policymakers to redesign financial aid programs in the future. Due to the underrepresentation of females in STEM, using financial aid, especially targeted on women, may motivate more female students to switch to STEM programs through a more direct and positive substitution effect. It could be very influential to the whole society that more female STEM workers are produced by the higher education system. I hope my analysis of heterogeneity in gender is a good start and it is still necessary to have further theoretical and empirical studies that continuously evaluate the relationship between policy tools and women STEM participation.

Here are also some caveats of my empirical results. First, my analytic sample is restricted and the estimates only measure the local treatment effect of the compliers nearby the threshold. Unlike merit aids in Georgia and Florida, the Missouri Bright Flight Scholarship Program is a highly targeted program, with a focus on top-ranked students. Though top-ranked students are probably from families with higher social-economic status, the compliers among them imply that they are less likely to retake the ACT. The local treatment effect reminds policymakers to extrapolate the results and implications for a different sample carefully. Second, due to the high tuition and fees in private institutions and the small amount of the scholarship, the financial aid may not specifically provide financial incentives for students to shift between Missouri public and private institutions. However, the merit aid may stop the brain drain (Harrington et al., 2016; Zhang & Ness, 2010) and create a substitution effect when choosing in-state private institutions compared to out-of-state ones. For example, the aid may encourage students to choose Washington University in Saint Louis rather than Vanderbilt University in Nashville. It finally contributes to producing more STEM graduates in Missouri, which can also be regarded as an important policy implication of the financial aid. Unfortunately, due to the data limitation, the margins between in-state and out-of-state private institutions cannot be measured. Third, compared to claiming a specific major, enrolled credits are more directly related to students' financial costs. Students can enroll as engineering students but take more non-engineering courses or other equivalent cheap credits. The data does not include related variables to allow me to examine the effect of financial aid on students' course taken behaviors, even though I notice that STEM credits may be a better proxy associated with the cost of attendance directly.

CHAPTER 3

ACT Score as the Running Variable: Revisiting Regression Discontinuity Applications in Financial Aid Research

Empirical research in higher education policy has focused on establishing the causal link between policy-relevant variables and academic outcomes. As a quasi-experimental design, regression discontinuity (RD) design can create a local-randomized sample and provide unbiased estimates of the treatment effects, leading to a growing popularity as a way to evaluate higher education policy, especially in financial aid research (McCall & Bielby, 2012; Nguyen et al., 2019). For example, many merit-based financial aid programs usually have requirements of standardized test scores, such as the ACT and the SAT (Dynarski, 2002; Frisvold & Pitts, 2018; Zhang & Ness, 2010). Hence, researchers can utilize these policy cut-offs in standardized tests to apply regression discontinuity design to evaluate the treatment effects of financial aid programs on student outcomes. As a result, test scores become popular running variables. Subjects who just pass the threshold will be assigned to the treatment group (receiving a merit-based financial aid) while the rest will stay in the comparison group. Within a very narrow range, there is no reason to believe that students who just below the cut-off are systematically different (on both observed and unobserved factors) from those who just above the cut-off, which means that the sample around the cut-off can be regarded as a random experiment (if certain assumptions are satisfied). The only mean difference should be the treatment variable, whether the subject receives the treatment or not. Then the causal effect of the treatment can be identified by the difference in the outcomes of the students who barely pass the threshold against those who do not.

However, test scores are not perfect running variables in regression discontinuity designs and the following three issues should be highlighted when applying regression discontinuity designs with test scores. First, the behaviors of retaking tests widely exist among different groups of students (Goodman et al., 2018; Vigdor & Clotfelter, 2003) and may also be influenced by specific financial aid programs (Bruce & Carruthers, 2014; Harrington et al., 2016; Welch, 2014; Zhang et al., 2016), indicating that the test scores can be often manipulated because of the incentives from financial aid. Therefore, the randomization assumption of the regression discontinuity design may not be satisfied due to the manipulation of the running variable. Moreover, the selection bias can finally result in biased estimates and make the causal inference less convincing.

Second, test scores are usually reported as discrete variables. For example, the ACT is officially reported as scaled scores, including all the whole values that range from 1 to 36. As a result, when applying regression discontinuity designs, due to the discretization of the reported scores, researchers have to extrapolate from the largest value of the score just below the cut-off value to the cut-off value. The extrapolation may create potential bias in point estimation, called the "rounding errors" (Dong, 2015).

Third, the extrapolation may also lead to misleading inference (McCall & Bielby, 2012). But the procedures to calculate the proper statistical inference recommended by the literature are not consistent. D. S. Lee and Card (2008) suggested using standard errors clustered by the running variable (CRV) so that the statistical inference would be more robust. Oppositely, Kolesár and Rothe (2018) showed the confidence intervals based on clustered standard errors had poor coverage properties through simulations. Instead, they recommended choosing the conventional Eicker-Huber-White (EHW) heteroskedasticity-robust standard errors when the discrete running variables could provide rich support within a smaller bandwidth. For applied researchers, using the improper standard errors may lead to wrong statistical inference and impact the hypothesis testing, particularly when the p-value is very close to the thresholds of preferred significance levels.

Given the concerns of test scores in regression discontinuity designs and their empirical applications, in this paper, I will focus on the ACT[1] scores and discuss the related issues when using them as the running variables in regression discontinuity designs. I will explore the following three issues, retaking and test score manipulation, rounding errors in the running variable, and misleading statistical inference, in the evaluations of merit-based financial aid programs. Particularly, I will summarize and compare the practices in recent empirical articles with the ACT scores as the primary running variables and provide recommendations to researchers to guide their future financial aid research.

An Introduction to Regression Discontinuity Design

The regression discontinuity design is a popular quasi-experimental design to estimate the causal treatment effect of policies or interventions on subjects. In a general setting of sharp regression discontinuity, every subject will be assigned a continuous random number X. This continuous number, X, is also known as the running variable. The assignment to the treatment group or the control group is based on the running variable X and the threshold c. Individuals with $X \geqslant c$ will be given the treatment and the others will be assigned to the control group. For example, Figure 3.1 shows the treatment status in a sharp regression discontinuous design. All the subjects who pass the threshold will receive the treatment and the others will not, indicating a significant change in the treatment status around the threshold. Figure 3.2 displays the scatter plot between the outcome Y and the running variable X. The fitted lines are generated by the data points below and above the threshold, separately.

---

[1] Without specific notification, the ACT score in this paper usually refers to the ACT composite score.

The outcomes of individuals above the threshold can be impacted by the treatment. It is reasonable to assume that subjects around the threshold are comparable since they have similar values of X. Hence, the difference in the outcomes of those subjects can be explained as the causal effect of policies or interventions. Furthermore, the gap between the two fitted lines at the threshold can be interpreted as the local average treatment effect (LATE).

Based on the guidelines from What Works Clearinghouse (Schochet et al., 2010), to make the regression discontinuity estimates more convincing, the following four points should be carefully discussed, and certain standards must be satisfied:

1. Integrity of the forcing variable (running variable).
2. Attrition.
3. Continuity of the outcome-forcing variable relationship.
4. Functional form and bandwidth.

Among all the four points above, the running variable's integrity is the most crucial condition for regression discontinuity designs to produce unbiased estimates of effects of an intervention, which means there is no systematic manipulation of the running variable. Particularly, the running variable cannot be precisely manipulated due to the threshold[2].

Hence, in the three issues mentioned previously, the first issue, retaking and test score manipulation, will be the most important one since it is associated with the internal validity of the regression discontinuity design. The rest two issues only happen when the running variable is discrete and they may also not have a severe impact on the empirical results.

---

[2] In some cases, though the running variable is manipulated, those manipulated data points may be far away from the interested thresholds. It indicates that the manipulation is not directly associated with the thresholds. In this situation, the manipulation and selection bias may not brother the causal inference and the internal validity of the regression discontinuity design can still be secured.

Figure 3.1

Treatment Status in a Sharp Regression Discontinuity Design

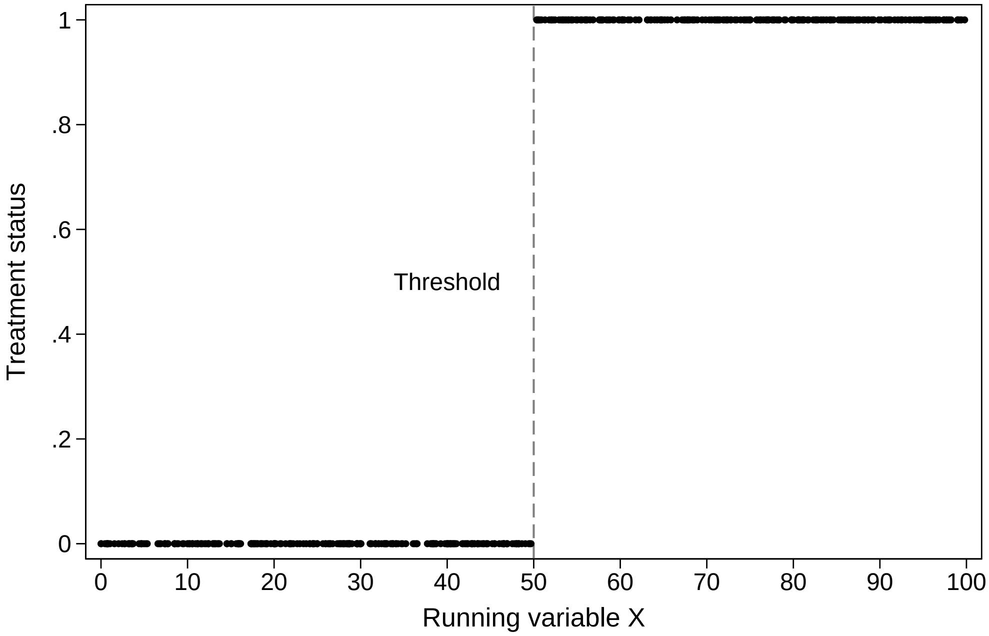
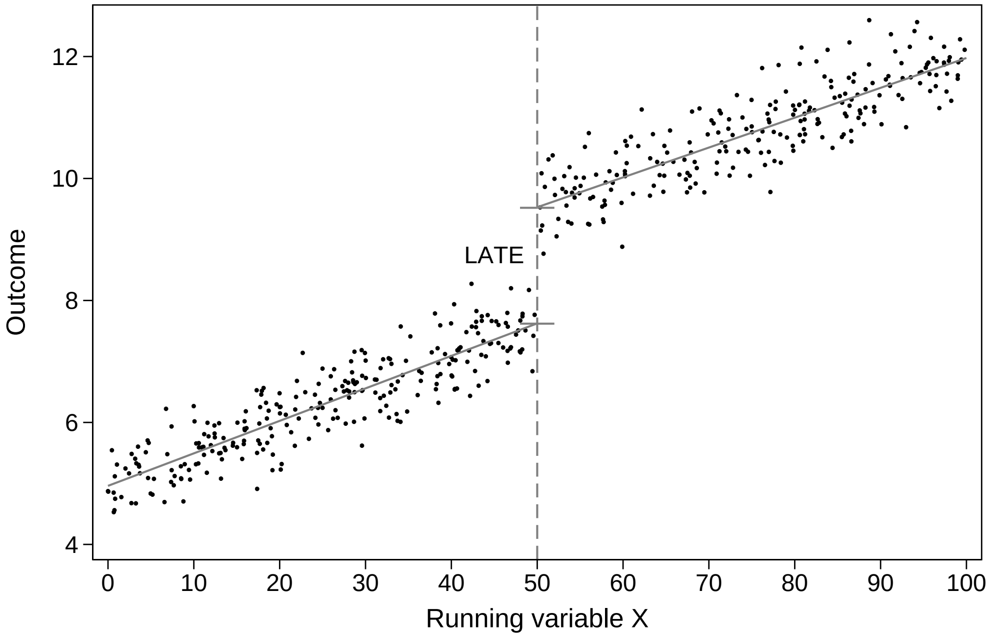


Figure 3.2

Estimated Local Average Treatment Effect



Retaking and Test Score Manipulation

In this section, I will focus on the first point, the integrity of the running variable. The manipulation of the running variable is the biggest threat to the internal validity of the regression discontinuity design. Luckily, not all the running variable can be manipulated. For example, many time-related running variables are considered as not manipulatable, such as age and birthday[3].

However, several running variables in education studies, especially variables like test scores, are potentially manipulatable. As a standardized test, the ACT takers are legally allowed to retake the ACT for up to 12 times. Further, higher education institutions and state agencies usually use students' highest scores[4] as the reference for admission or scholarship. Hence, students who are close to the cut-off are impacted by the incentives from the treatment (being admitted or receiving aid) and are more likely to retake the exam, which will invalidate the assumption of the local randomization. Particularly, previous empirical studies have verified that the behaviors of retaking standardized tests are associated with thresholds of financial aid programs. Harrington et al. (2016) showed that in Missouri, the average times of retaking the ACT among students with their first ACT scores just below the threshold were about 2.7, while the same number among those students with scores just above the threshold significantly dropped to 2.0. Their empirical evidence implied that the threshold provided strong incentives to students to encourage them to retake the ACT. Similar retaking patterns influenced by state financial aid programs were also supported by empirical evidence from Florida (Zhang et al., 2016) and Tennessee

---

[3] In very rare circumstances, there may still exist manipulations among birthdays. See Huang et al. (2020) for more information.

[4] According to a survey of the current landscape of college admissions, institutions may prefer a "most recent", "single highest", or "combined highest" approach in defining what ACT Composite score it considers (see the ACT website for more information: https://www.act.org/content/act/en/products-and-services/the-act-postsecondary-professionals/scores/multi-scores.html). Though some institutions may treat multiple scores differently, since many students can increase their latest test scores through retesting (Harmston & Crouse, 2016), all three approaches seem to be friendly to retesting and encourage students to retake the ACT.

(Bruce & Carruthers, 2014; Welch, 2014).

Without specific notification, in this paper, the highest test score refers to the "single highest" approach since it is more frequently used in empirical papers about financial aid evaluation. It is necessary for researchers to confirm which score is used to determine the eligibility of the state-funded financial aid and to make sure whether the score can be manipulated precisely. In some scenarios, the regression discontinuity estimators are still consistent if there is an "optimization error" that prevents agents from precisely controlling the running variable (D. S. Lee & Lemieux, 2010). If the "single highest" approach is used, students will be aware of their aid eligibility and can precisely change it through retaking the test.

Hence, when using the reported highest test scores ("single highest") as the running variable, the self-selection issue will jeopardize the causality and likely overestimate the treatment effect. For example, to reduce the cost of higher education, students who want to enter colleges will retake the ACT to be eligible for financial aid, particularly when their scores are just below the threshold. If the highest test score is used as the running variable, many students with college enrollment plans will be assigned to the treatment group. Hence, the difference in the percentage of college enrollment around the cut-off can also be explained by self-motivation in addition to receiving financial aid. The difference will finally overestimate the impact of financial aid programs on college enrollment since it is hard to control the self-motivation bias.

Practice in the Literature

Table 3.1 summarizes eight recent empirical papers using regression discontinuity designs to evaluate five different state-funded merit-based scholarships (Florida, Iowa, Missouri, Tennessee, and West Virginia). These merit aids usually have requirements on the ACT score and the cut-offs have been utilized to implement either sharp regression discontinuity or fuzzy regression discontinuity designs according to the real administrative datasets. However, though most of these papers have discussed the threat on internal validity from retaking the ACT, the procedures to control the selection bias from the manipulation are various by authors.

To identify the existence of manipulation, the McCrary density test (McCrary, 2008) (including the visual examination of the histogram of the running variable) has been widely suggested and adopted in most articles (Bruce & Carruthers, 2014; Harrington et al., 2016; Leeds & DesJardins, 2015; Scott-Clayton, 2011; Scott-Clayton & Zafar, 2019; Welch, 2014; Zhang et al., 2016). Only studies in Tennessee (Bruce & Carruthers, 2014; Welch, 2014) reported the statistics of the McCrary test. Other studies just provided figures of visual examinations probably because the density functions in those papers were discontinuous around the threshold (Harrington et al., 2016; Zhang et al., 2016).

Unfortunately, many papers have indicated the manipulation of the ACT score, requiring more discussion about endogeneity. In Missouri and Tennessee, researchers have compared the different results of the McCrary test on the first ACT score and the highest ACT score. They found that the density function of the highest ACT score was not continuous around the threshold of aid eligibility, which strongly indicated the manipulation of the running variable (Bruce & Carruthers, 2014; Harrington et al., 2016; Welch, 2014). Based on the first-time ACT, the average times of retaking were also significantly different between students above and below the threshold, which implied students who were just below the threshold at the first attempt would retake more times to manipulate their scores (Harrington et al., 2016). Due to the absence of historical test records[5], studies in West Virginia did not report the comparison of histograms between the first ACT score and the highest ACT score (Scott-Clayton, 2011; Scott-Clayton & Zafar, 2019). Instead, they utilized the data before and after the aid implementation. They also found that students might react to the financial

---

[5] The administrative dataset in West Virginia only had at most one ACT or SAT score per student. Researchers presumed that students reported their highest score at the time of the application (Scott-Clayton & Zafar, 2019). The assumption is also correspondent to related policies in other states, such as Missouri and Tennessee.

Table 3.1

Manipulation Issues in Recent Evaluations of Merit-based Financial Aid Programs

Program	Eligibility	Paper	RDD	Running variable	Manipulation
Tennessee Education Lottery Scholarship	ACT>=21 or HS GPA>=3.0	Bruce and Carruthers (2014)	Fuzzy RDD	First ACT (unrounded)	Manipulation is excluded by fuzzy RDD using 2SLS; McCrary tests.
Tennessee Education Lottery Scholarship	ACT>=21 or HS GPA>=3.0	Welch (2014)	Fuzzy RDD	First ACT (unrounded)	Manipulation is excluded by fuzzy RDD using 2SLS; McCrary tests.
Missouri Bright Flight Program	ACT>=30	Harrington et al. (2016)	Sharp RDD; Fuzzy RDD	Highest ACT; first ACT (rounded)	Manipulation is excluded by fuzzy RDD using 2SLS; McCrary tests.
Iowa National Scholars Award	Index mixed with ACT and HS GPA	Leeds and DesJardins (2015)	Fuzzy RDD	Admissions Index Score/ Regent Admission Index (rounded ACT contained)	Authors claim the index is hard to be manipulated; McCrary tests.
West Virginia PROMISE Scholarship	ACT>=21 or HS GPA>=3.0	Scott-Clayton and Zafar (2019)	Fuzzy RDD	ACT (rounded, first/highest not specified)	Using a "donut-hole" specification that excludes observations closest to the cutoff.
West Virginia PROMISE Scholarship	ACT>=21 or HS GPA>=3.0	Scott-Clayton (2011)	Fuzzy RDD	ACT (rounded, first/highest not specified)	Using previous year data to control the bias from manipulation.
West Virginia PROMISE Scholarship	ACT>=21 or HS GPA>=3.0	Leguizamon and Hammond (2015)	Sharp RDD	ACT (rounded, first/highest not specified)	Manipulation is discussed but remained.
Florida Bright Futures Program	ACT>=20 or ACT>=28	Zhang et al. (2016)	Sharp RDD	Highest ACT (rounded)	Manipulation is excluded in the SAT sample; Lack of analysis of the ACT sample.

Note: RDD=Regression Discontinuity Design; 2SLS=Two Stage Least Squares.

aid and retake the test, especially for students just below the threshold of aid eligibility. Scholars using Florida administrative data also addressed the manipulations of the SAT score around two thresholds. They primarily analyze the sample of the SAT takers in Florida, but the paper lacked discussions about the existence of the manipulation in the ACT. According to the patterns in the SAT, it is possible to infer that the same situation happens in the ACT sample. Leeds and DesJardins (2015) found their running variable in the Iowa dataset passed the McCrary test, but their running variable was an index associated with the ACT instead of the ACT along. Though the manipulation of the ACT may exist, the other parts in the index, including high school GPA, may reduce the impact from retaking the ACT, making the index less manipulatable.

To solve the selection bias from manipulation, due to data availability, there are generally two main procedures in the literature, one is using the first ACT to replace the highest ACT and the other is using the "donut-hole" specification.

Replacing the Running Variable

After verifying the manipulation of the highest ACT, some researchers have recommended replacing the highest ACT with the first ACT as the primary running variable (Bruce & Carruthers, 2014; Harrington et al., 2016; Welch, 2014). They utilized the standardized nature of the ACT and the assumption that students were trying to score as high on the ACT as possible to claim that the first ACT score would not be manipulated. Figures in those papers also indicated that the first ACT scores had smooth density functions around the cut-offs and passed the McCrary density tests (Bruce & Carruthers, 2014; Welch, 2014). Here is a comparison between two regression discontinuity applications with these two different running variables separately.

To simply the comparison, I assume that the treatment variable is exclusively determined by the highest ACT score. So the framework of sharp regression discontinuity is introduced when the running variable is the highest ACT score. In the following equation, the coefficient β can be interpreted as the effect of the aid on related education outcomes. ACT_dist_i indicates the distance of the highest ACT score to the cut-off. $Above_i$ or $Below_i$ represents whether the score is above or below the threshold. And p indicates the order in the polynomial regression.

$$\begin{aligned} Y_i = \beta Above_i &+ \sum_{j=1}^{p} \gamma_j^{-} (ACT\_dist_i)^j * Below_i \\ &+ \sum_{j=1}^{p} \gamma_j^{+} (ACT\_dist_i)^j * Above_i + \epsilon_i \end{aligned} \tag{3.1}$$

The second model is under the framework of fuzzy regression discontinuity and the first ACT score is used as the running variable. Since the treatment variable is not exclusively determined by the first score, as a framework of treatment-on-treated, the estimate can be interpreted as the effect of the aid for those who became aid eligible on their first ACT attempt. But using the first score as the running variable will bring biased estimates due to the noncompliance (D. S. Lee & Lemieux, 2010). Fuzzy regression discontinuity design can remove the noncompliance bias through a two-stage procedure (2SLS). Using a similar setting of polynomial regression in the sharp regression discontinuity design above, the following equations show how the 2SLS procedure works:

In the first stage, being aid eligible, $Aid\_eligible_i$, is estimated based on whether the subject is aid eligible on the first attempt, $Above_i$, the distance of the first ACT score to the cut-off, ACT_dist_i.

$$\begin{aligned} Aid\_eligible_i = \alpha Above_i &+ \sum_{j=1}^{p} \delta_j^{-} (ACT\_dist_i)^j * Below_i \\ &+ \sum_{j=1}^{p} \delta_j^{+} (ACT\_dist_i)^j * Above_i + \varepsilon_i \end{aligned} \tag{3.2}$$

In the second stage, the predicted probability of receiving the aid, $\widehat{Aid}_i$, is substituted for being aid eligible to estimate the effect of the aid on dependent variables. With the exogenous $Aid\_eligible_i$ as the instrument variable, the causality can be captured by the fuzzy regression discontinuity design with the first ACT as the running variable.

$$\begin{aligned} Y_i = \beta \widehat{Aid}_i &+ \sum_{j=1}^{p} \gamma_j^{-} (ACT\_dist_i)^j * Below_i \\ &+ \sum_{j=1}^{p} \gamma_j^{+} (ACT\_dist_i)^j * Above_i + \epsilon_i \end{aligned} \tag{3.3}$$

This method has both pros and cons. Ideally, using the first ACT score is a perfect solution to secure causality. The first ACT score is normalized with a continuous density function. Though in very rare situations, the policy cut-off can motivate some students to take their first test more seriously and may slightly change the density function of the first ACT score when no financial aid exists. Based on thousands of observations, it is acceptable to assume that its density function is still smooth around the cut-off and the first score is irrelevant to the policy cut-off. However, researchers should be aware of the disadvantages when applying this strategy. The fuzzy regression discontinuity design mentioned above is equivalent to the application of an instrument variable. The fuzzy regression discontinuity estimate is the local treatment effect of the students near the threshold and the identification highly relies on the variation of outcome variables (e.g., college enrollment/choice) among compliers. The instrument can only capture the variation among a group of compliers who are less likely to retake the ACT based on their first ACT score and probably from low-income families. If the coefficient of the instrument in the first stage is small, which means the percentage of the compliers around the cut-off is low, the statistic power can be impacted due to fewer compliers, leading to inaccurate estimates with largely increased standard errors. Moreover, the compliers may also be sensitive to financial incentives and the fuzzy estimates may be amplified and the treatment effect can be over addressed to the audience if they are not very clear about the meaning of the local treatment effect in practice.

“Donut-hole” Specification

What if some administrative datasets do not collect historical test records? Without the unmanipulated first ACT score, researchers have tried to discuss how severe the selection bias is or whether the bias can be significantly controlled. Leeds and DesJardins (2015) argued that though their running variables were indexes that partially included the ACT scores, they believed that retaking the ACT would not largely affect their indexes. The McCrary test also supported that their running variables were too hard to be manipulated through retaking the test. If the selection bias cannot be ignored, a “donut-hole” specification that excludes observations closest to the cut-off is recommended. Observations near the cut-off are assumed to be non-compliers with a higher probability. These subjects are supposed to stay in the control group, but they actually receive the treatment through retaking the test. By removing these observations, researchers can exclude the selection bias and use the rest compliers to identify the treatment effect.

In practice, the “donut-hole” method is more widely utilized in health care settings when the running variables have manipulation issues, such as health care expenditure (Bajari et al., 2011); or when the running variables have data heaping issues, such as minute of birth (Almond & Doyle, 2011) and weight of babies (Barreca et al., 2011). Similarly, in educational settings, manipulation and heaping also exist in many running variables. GPA is a common running variable with heaping issues due to the averaging process of the raw letter grades. Since students may enroll in different numbers of courses, the divisor, which is the total number of credits, are not identical among all the students. With the unbalanced averaging process from different divisors, the density function of the GPA may have jumped at certain values. For example, students are easier to have a GPA of 3.00 instead of 2.90 because the latter GPA requires students to achieve ten graded credits (or the multiple of 10) so that the GPA may have a chance to be 2.90. “donut-hole” method is applied with GPA as the running variable (Goodman et al., 2019; Scott-Clayton & Schudde, 2019). When standardized test scores become the running variable, as mentioned above, manipulation issues from retaking are more concerned in regression discontinuity designs. “Donut-hole” specification is also adopted in the literature. (Cohodes & Goodman, 2014) run “donut-hole” regressions because of the small amount of bunching observed in the running variable. The running variable was derived from the raw math score in the ELA test and the histogram showed that the density function might not be continuous around the threshold. As a robustness check, estimates from “donut-hole” regressions can provide more solid empirical results. Scott-Clayton and Zafar (2019) also reported the “donut-hole” specification in addition to other model specifications with different bandwidths. They noticed the different density functions of the ACT score between before- and after- policy implementation, which might indicate the existence of manipulation.

If the records of the first ACT score are not available, “donut-hole” specification can be a solution to exclude selection bias from manipulation. However, researchers should be aware of the weakness in this method. The “donut-hole” may not be big enough to exclude the majority of students who retake the ACT. For example, in Missouri, the state merit-based financial aid eligibility was 30 or above before 2008. Conditional on the first ACT, about 25% of the students whose first score was 28 would achieve eligibility later, probably through retaking the test (Harrington et al., 2016). The empirical evidence indicated that a large number of students could increase their final scores in at least 1 or 2 points. When adopting the highest ACT as the running variable, how many data points should be removed near the cut-off to secure causality? Since the ACT score includes whole values that range from 1 to 36, the treatment group can only contain 7 data points (30-36). Considering that only a few students can achieve the ACT of 35 or 36, due to the smaller sample size, those two data points may not be very reliable if included in the estimation. In this situation, there is less freedom for researchers to dig the "donut-hole" in the ACT score.

Further, even the cut-off has enough data points on each side (e.g., the ACT score of 21 in several financial aid programs), the "donut-hole" specification may still have some issues. Researchers are facing the trade-off between precision and causality. Due to the large interval between the whole values in the ACT, when the "donut-hole" specification is applied, the causality may require a larger bandwidth with data points far away from the cut-off. As a result, the specification can also increase the inaccuracy of its estimates.

In conclusion, the running variables in heaping datasets may be more suitable for the "donut-hole" specification. Though most of their data points cannot be manipulated, due to the measurement errors, including the data points close to the threshold in regression models may be problematic. Excluding those observations can improve precision and benefit the estimation. The highest ACT score does not have similar features in heaping data since it has more manipulated data points, not limited to the ones very close to the threshold.

Recommendations When Faced with Potential Manipulation

The manipulation is always an unignored threat in regression discontinuity design and researchers should address this issue very carefully. I recommend that researchers should report the McCrary density test (McCrary, 2008) to verify the existence of the manipulation, especially when they want to exclude the selection bias from manipulation. If the test does not significantly imply the manipulation, researchers are also welcome to add the "donut-hole" specification into the section of robustness check to strengthen their results (Cohodes & Goodman, 2014).

If the manipulation exists, researchers should be aware of the trade-offs for both two methods mentioned above. The causal estimates are very important for policy evaluation, but they usually sacrifice a large proportion of the whole sample. To sum up, I strongly encourage researchers to use the first ACT as the running variable. Using the first ACT score is a great solution for causality since the exogeneity of the first test score has been widely recognized in the literature (Bruce & Carruthers, 2014; Harrington et al., 2016; Welch, 2014; Zhang et al., 2016). The main weakness is that it provides fuzzy regression discontinuity estimates and the result is interpreted as a local treatment effect. The fuzzy regression discontinuity estimation may lack external validity with less precision, particularly when the coefficient in the first-stage is not big enough.

If the dataset is not applicable for historical test records, I think the "donut-hole" specification is also a necessary supplement to improve the robustness, although it still has some issues in claiming causality. Researchers should be careful to make a causal argument when only using this specification to exclude selection bias. As a suggestion, this strategy may have a better performance when the running variable has more data points with smaller intervals, such as indexes adopted in Iowa National Scholars Award (Leeds & DesJardins, 2015); or only a few data points near the threshold are facing a threat from potential manipulation (or measurement errors), such as time of birth and weight of babies in health care settings (Almond & Doyle, 2011; Barreca et al., 2011).

Rounding Errors in the Running Variable

Dong (2015) framed a setting that the discrete running variable had a continuous latent running variable and only the rounded version was recorded. This measurement error, called rounding error, leads to inconsistent estimates of treatment effects, even when the real functional form of the outcome and the running variable is known and correctly specified. Unfortunately, the ACT score is a running variable with such features mentioned above. Specifically, the ACT score is reported as the average value of the four subject scores (English, mathematics, reading, science), rounded to the nearest whole number[6]. The rounded ACT score has a minimal interval of 1 while the unrounded score has a minimal interval of 0.25. Hence, in some circumstances, the rounding errors from the rounded ACT score may create biased estimates in some regression discontinuity designs.

Revising the Bias

More generally, Dong (2015) raised a model to revise the coefficient to remove the bias. Following her procedure, the regression model in a sharp regression discontinuity design[7] can be written like this (using a polynomial specification and the highest order is 4):

$$Y_i = d_0 + d_1 X_i + d_2 X_i^2 + d_3 X_i^3 + d_4 X_i^4 + (c_0 + c_1 X_i + c_2 X_i^2 + c_3 X_i^3 + c_4 X_i^4) T_i^* + \varepsilon_i \quad (3.4)$$

Y_i is the student outcome. X_i is the running variable, the ACT composite score, and rounded to the nearest whole value. Its minimal interval is 1. The running variable is also centered to the threshold so 0 refers to the cut-off point. T_i^* is a dummy indicator whether the running variable is above the threshold so that $T_i^* = 1(X_i > 0)$.

The naive discrete data treatment effect τ' will just be c_0 in equation 3.4. With the assumption that the running variable within a whole value is uniformly distributed, the true treatment effect τ is approximately equal to[8]:

$$\tau = c_0 - \frac{1}{2}c_1 + \frac{1}{6}c_2 - \frac{1}{30}c_4 \quad (3.5)$$

---

[6] For more information about the ACT subject scores, please see: https://www.act.org/content/act/en/products-and-services/the-act/scores/understanding-your-scores.html.

[7] To simplify my analysis, the fuzzy regression discontinuity design is not included. If interested, find more discussions about the fuzzy regression discontinuity design in Dong (2015)

[8] For more technical details about how to get this equation, please see Dong (2015).

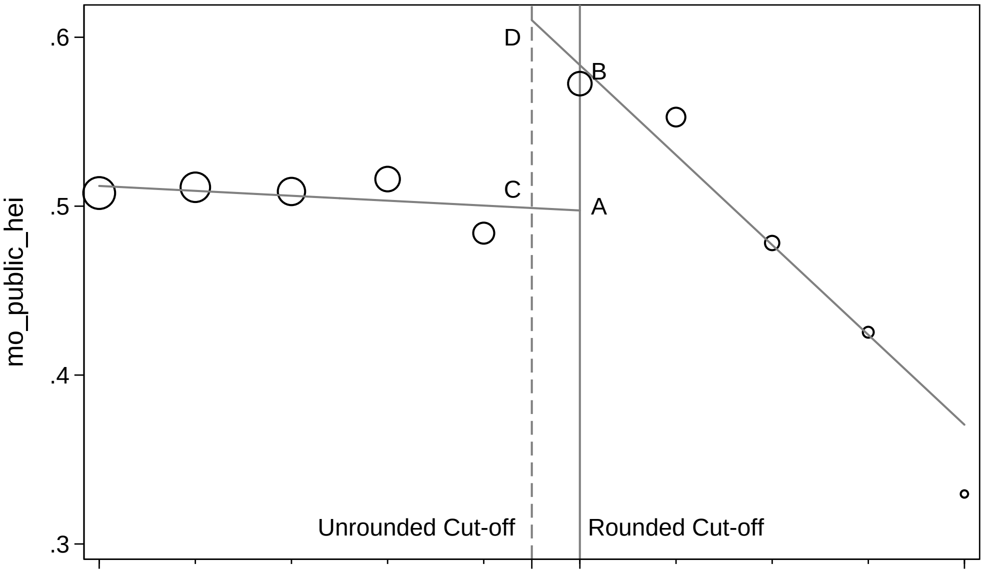
Hence, the bias from the rounding errors is:

$$\tau' - \tau = \frac{1}{2}c_1 - \frac{1}{6}c_2 + \frac{1}{30}c_4 \tag{3.6}$$

Figure 3.3 is a simple example that explains how the rounding errors may contribute to the biased point estimation. It describes a rough measurement of the treatment effect of a merit-based financial aid in Missouri on an interested student outcome (a dummy variable), choosing a Missouri public institution among all Missouri ACT takers. The highest ACT is the running variable and students who pass the threshold (the composite ACT score is 30 or above) will be eligible for the aid.

Figure 3.3

Bias from the Rounding Errors in Regression Discontinuity Design



Note: The figure describes the effect of the merit-based financial aid on choosing Missouri public higher education institutions. The running variable is the highest ACT score. Students who achieve an ACT composite score of 30 or above will be eligible for this award. The model setting may have issues since the running variable can be manipulated. So this figure is used to describe the bias directly and has no policy implication.

When the running variable is rounded to the whole values, the cut-off point will be 30 and the length of AB, $\overline{AB}$, measures the effect on the student outcome under a linear specification. However, if the unrounded ACT with quarter points is used, the new cut-off could be 29.5 and the length of CD, $\overline{CD}$, will be the new measurement of the treatment effect[9]. According to Figure 3.3, it has:

$$\overline{AB} - (-\frac{1}{2})Slope_{AC} = \overline{CD} - (-\frac{1}{2})Slope_{BD} \tag{3.7}$$

So the bias is:

$$\tau' - \tau = \overline{AB} - \overline{CD} = \frac{1}{2}(Slope_{BD} - Slope_{AC}) = \frac{1}{2}c_1 \tag{3.8}$$

It is consistent with Equation 6 when the model has a linear specification.

The bias from the rounding errors can also be regarded as the impact of relocating the threshold. The rounding process moves the real threshold (e.g., 29.5) to the rounded whole value (e.g., 30). Without detailed information between the real threshold and the new one due to the usage of a rounded running variable, the extrapolation may lead to a bias in estimation. In Figure 3.3, the bias is generated from using the extrapolated trend of $\overline{AC}$ to replace the real trend of $\overline{BD}$.

Since the high-order polynomials are not recommended in regression discontinuity designs (Gelman & Imbens, 2019), most of the empirical papers only adopt linear or quadratic specification when using the parametric method, I will only leave linear and quadratic terms to simplify the Equation 3.6.

For the linear specification, the bias is:

$$\tau' - \tau = \frac{1}{2}c_1 \tag{3.9}$$

---

[9] To simplify my comparison, I assume the distribution of scores within a whole value is uniform. Dong (2015) also provided a more general equation to revise the bias without the assumption of uniform distribution.

Furthermore, for the quadratic specification, the bias is:

$$\tau' - \tau = \frac{1}{2}c_1 - \frac{1}{6}c_2 \tag{3.10}$$

Based on the equations above, it is clear that the key determinant of the bias from the rounding errors is the difference of coefficients below and above the threshold, c_1 and c_2. In other words, if the slopes are restricted on both sides, there is no bias from the rounding errors. In the following section, I will review the practice in the empirical papers related to the rounding errors and try to use the equations above to evaluate how severe the bias from the rounding errors is in several papers.

Practice in the Literature

Table 3.2 summarizes the same recent eight empirical papers in financial aid evaluation using regression discontinuity designs. Many of them (6 of 8) use the rounded ACT score as the running variable (except studies of programs in Tennessee, which may indicate the bias from the rounding errors. Though facing issues in rounding errors, some papers still applied no slope restriction to their basic regression discontinuity regression models and did not test the hypothesis whether the slopes below and above the threshold are statistically different or not (Leguizamon & Hammond, 2015; Scott-Clayton, 2011; Scott-Clayton & Zafar, 2019). Instead, these studies provided figures with fitted lines and others can compare the difference of slopes visually. Only Harrington et al. (2016) reported estimates of both linear and quadratic specifications. Then in each functional form, the coefficients of the running variable were also provided, including results with/without slope restriction. Finally, after comparing all the functional forms, they picked the quadratic specification with slope restriction so that the bias from the rounding errors did not impact their primary estimates.

Considering the detail information about the regression discontinuity estimates

Table 3.2

Rounding Errors Issues in Recent Evaluations of Merit-based Financial Aid Programs

Program	Paper	RDD	Running variable	Specification	Running variable coefficient
Tennessee Education Lottery Scholarship	Bruce and Carruthers (2014)	Fuzzy RDD	First ACT (unrounded)	Linear; w/o slope restriction	Figures reported
Tennessee Education Lottery Scholarship	Welch (2014)	Fuzzy RDD	First ACT (unrounded)	Linear; w/o slope restriction	Figures reported
Missouri Bright Flight Program	Harrington et al. (2016)	Sharp RDD; Fuzzy RDD	Highest ACT; first ACT (rounded)	Liner, quadratic; w/ and w/o slope restriction	Coefficient reported
Iowa National Scholars Award	Leeds and DesJardins (2015)	Fuzzy RDD	Admissions Index Score/ Regent Admission Index (rounded ACT contained)	Linear, quadratic; w/ slope restriction	None
West Virginia PROMISE Scholarship	Scott-Clayton and Zafar (2019)	Fuzzy RDD	ACT (rounded, first/highest not specified)	Linear, quadratic; w/o slope restriction	Figures reported
West Virginia PROMISE Scholarship	Scott-Clayton (2011)	Fuzzy RDD	ACT (rounded, first/highest not specified)	Linear, quadratic; w/o slope restriction	Figures reported
West Virginia PROMISE Scholarship	Leguizamon and Hammond (2015)	Sharp RDD	ACT (rounded, first/highest not specified)	Linear, quadratic; w/o slope restriction	Figures reported
Florida Bright Futures Program	Zhang et al. (2016)	Sharp RDD	Highest ACT (rounded)	Linear, quadratic; w/o slope restriction	None for ACT

Note: RDD=Regression Discontinuity Design.

Table 3.3

Revising the Bias from the Rounding Errors

Bandwidth	ACT24-35	ACT25-34	ACT26-33	ACT27-32
Linear w/o slope restriction				
Aid eligibility	0.0220	0.0237	0.0267	0.0332
c1	-0.0286	-0.0266	-0.0247	-0.0163
Aid eligibility revised	0.0363	0.0370	0.0391	0.0414
Difference	-39.4%	-35.9%	-31.6%	-19.7%
Quadratic w/o slope restriction				
Aid eligibility	0.0334	0.0387	0.0465	0.0414
c1	-0.0146	-0.0128	-0.0048	-0.0404
c2	-0.0004	0.0014	0.0032	0.0179
Aid eligibility revised	0.0406	0.0453	0.0494	0.0646
Difference	-17.8%	-14.6%	-5.9%	-35.9%
Aid eligibility preferred	0.0454	0.0431	0.0425	0.0420

Note: The data is derived from Table 3 in Harrington et al. (2016). c_1 (linear), c_2 (quadratic) are the difference between slopes below and above the threshold. Difference= (Aid eligibility- Aid eligibility revised)/Aid eligibility revised. Column 2 is the authors' preferred bandwidth. The preferred estimate of aid eligibility in the last row is the estimate from the quadratic specification with slope restriction.

in Harrington et al. (2016) while other papers did not provide coefficients of their running variables, I plan to borrow the coefficients from Harrington et al. (2016) and use Equation 3.9 and Equation 3.10 to revise the coefficient to remove the bias from the rounding errors. I want to explore how severe impacts it has if the bias is ignored. The data for examination is in Table 3.3 from Harrington et al. (2016). Only functional forms without slope restriction are examined.

Table 3.3 presents the revised coefficients according to the relevant regression discontinuity estimates in Harrington et al. (2016). In the linear specification without slope restriction, due to the significant difference between two slopes below and above the threshold, the estimate in the paper probably underestimated about 35% of the true treatment effect. As to the quadratic specification, where the effect was underestimated by approximately 15% due to the rounding errors. Though the bias from the rounding errors was significant, the final functional form that Harrington et al. (2016, p.436) preferred actually provided a more accurate estimation and was closer to those revised estimates based on other specifications. Their point estimations of the treatment effects were still considered to be convincing even they did not use the unrounded ACT score as the running variable or adjust the coefficients for the rounding errors.

Recommendations When Faced with Rounded Running Variable

It is no doubt that the unrounded ACT score is a better option for the running variable in regression discontinuity designs. Though the unrounded score is still not continuous, the score reduces the interval from 1 to 0.25 and removes the bias from the round error. Besides, conditional on the same bandwidth, the unrounded score can provide more data points so that the estimation will be more precise.

I encourage researchers to use the unrounded ACT score if the four subject scores are available. The unrounded score can be calculated from four subject scores manually. If the rounded score is the only option, researchers should be very careful to choose the proper functional form to fit the regression discontinuity regression model. Further, if the proper functional form does not restrict the slopes below and above the threshold, researchers should be aware that these specifications may result in the bias from the rounding errors when the slopes are significantly different. Therefore, to evaluate how severe the bias could be, I also suggest researchers report the coefficients of the running variables, which are the slopes below and above the threshold, or the slope difference (equivalent to the interactive terms between the running variable and the indicator of below or above the threshold[10]). If the slopes on two sides are very similar and not significantly different, then it is possible to conclude that the issue is not serious and the rounding errors may not affect the regression discontinuity

---

[10] See Table 3 about different functional forms in Harrington et al. (2016, p.436).

estimates significantly.

Misleading Statistical Inference with the Discrete Running Variable

This section discusses the practice in dealing with a similar bias from extrapolation. When using a discrete running variable, researchers always have to extrapolate from the largest value of the score just below the cut-off value to the cut-off value. This extrapolation due to a discrete running variable may overlap with the extrapolation due to the rounding process. To clarify these two different types of extrapolation, I still use Figure 3.3 as an explanation. As mentioned above, the extrapolation due to the rounding process is from 29.50 to 30.00, which is the line $\overline{AC}$. However, since the unrounded ACT score is not perfectly continuous and still has a minimal interval of 0.25, researchers have to extrapolate the fitted line from 29.25 to 29.50 since there is no data point between 29.25 and 29.50. If researchers do not realize the rounding process of the ACT score and directly use the rounded one as the running variable, they may treat the two types of extrapolation together (e.g., extrapolating the line from 29.00 to 30.00 in Figure 3.3).

The extrapolation due to a discrete running variable may also bring concerns in empirical settings. Moreover, the bias cannot be revised because there is no latent continuous variable to replace the unrounded ACT score. In other words, the composite ACT score only has a more "continuous" (but discrete in fact) latent variable, and the revision method provided by Dong (2015) can only partially solve the problems unless there exists a real continuous ACT score (or the minimum interval is smaller enough compared to the preferred bandwidth).

Researchers have begun to use robust standard errors to cover this bias and make the statistical inference with discrete running variables. Several solutions have been frequently discussed in econometric papers. Traditionally, D. S. Lee and Card (2008) suggested using the standard errors clustered by the running variable and the imperfect fit of the parametric function away from the discontinuity point would be covered by the robust standard errors. Oppositely, in both theoretical and empirical settings, Kolesár and Rothe (2018) concluded the poor coverage properties when using standard errors clustered by the discrete running variables suggested by D. S. Lee and Card (2008). The clustered standard errors were smaller than the conventional Eicker-Huber-White heteroskedasticity-robust standard errors[11] robust standard errors and might lead to misleading statistical significance. They recommended not to distinguish the case of a discrete or a continuous running variable. Particularly, they argued that if the discrete running variable had rich support, researchers could make the bandwidth smaller to reduce the bias of the treatment effect estimate, and use heteroskedasticity-robust standard errors for inference. Considering the theoretical debates in the literature, in the following section, I will focus on examining the practice in recent empirical papers and try to conclude a better procedure to help researchers to choose the most appropriate standard error for inference.

Practice in the Literature

Besides the theoretical debates in econometric papers, what is the common practice of the statistical inference in the empirical world? Table 3.4 summarizes the statistical inference in the same eight recent empirical papers about financial aid evaluation. Some studies have followed recommendations from D. S. Lee and Card (2008) and used standard errors clustered by the ACT score to make the statistical inference (Bruce & Carruthers, 2014; Harrington et al., 2016; Leguizamon & Hammond, 2015; Welch, 2014). Conversely, Scott-Clayton (2011) claimed that the procedure suggested by D. S. Lee and Card (2008) was not clearly an improvement in her study and universally reduced the estimated standard errors. Instead, she reported robust errors without clusters. Similar reasons were also mentioned by Scott-Clayton and

---

[11] Without specific notification, the robust standard errors in this paper refer to the conventional Eicker-Huber-White heteroskedasticity-robust standard errors.

Zafar (2019).

Similar to the theoretical discussion, the empirical practices are not consistent among all these papers. Unfortunately, many of these papers also do not report robustness checks about statistical inference with different standard errors. Without further information, it is difficult to argue that which standard errors could be better in their specific empirical settings and how the inappropriate standard errors may affect their statistical inference. The conclusions of studies with strongly significant or insignificant estimates may not be impacted even if different standard errors are applied. Table 3.4 also reports the significance of the estimates in each paper. Results in some papers may need further discussions. Although many questions are left, several common points can still be summarized as recommendations for researchers to pick the proper standard error.

Recommendations for Choosing the Appropriate Standard Error

The improper inference may provide misleading statistical significance. What is the best procedure to make the statistical inference with the ACT score as the running variable in regression discontinuity designs? Previous studies seem to provide inconsistent practices. Considering the data points in the preferred bandwidth (see Table 3.4), I notice that most of the studies were in a scenario that there were only a few data points in their regression discontinuity designs (except Leeds and DesJardins (2015)). It is because the ACT score only has a smaller scale from 1 to 36 (compared to other parts in the index), particularly when the score is rounded to the nearest whole value. Because of the nature of the test score, the very few clusters in these papers (maximum number is about 40 in unrounded scores and 10 in rounded scores) can over-reject the hypothesis (Cameron et al., 2008) and it is usually recommended that there should have at least 42 clusters (Angrist & Pischke, 2008). Due to the unbalanced size and the small number of clusters, it is possible that why some researchers claimed the failure of the procedure suggested by D. S. Lee and Card (2008)

Table 3.4

Statistical Inference in Recent Evaluations of Merit-based Financial Aid Programs

Program	Paper	Running variable	Bandwidth (data points)	Standard Error	Inference
Tennessee Education Lottery Scholarship	Bruce and Carruthers (2014)	First ACT (unrounded)	±20	Standard error clustered by first ACT	No significance in general
Tennessee Education Lottery Scholarship	Welch (2014)	First ACT (unrounded)	±20	Standard error clustered by first ACT	No significance in general
Missouri Bright Flight Program	Harrington et al. (2016)	Highest ACT; first ACT (rounded)	±5	Standard error clustered by highest/first ACT	Strong significance (highest); Weak significance (first)
Iowa National Scholars Award	Leeds and DesJardins (2015)	Admissions Index Score/ Regent Admission Index (rounded ACT contained)	±40-50	Standard error	Strong significance in general
West Virginia PROMISE Scholarship	Scott-Clayton and Zafar (2019)	ACT (rounded, first/highest not specified)	±5	Robust unclustered errors	Significance in some outcomes
West Virginia PROMISE Scholarship	Scott-Clayton (2011)	ACT (rounded, first/highest not specified)	±5	Robust unclustered errors	Significance in many outcomes
West Virginia PROMISE Scholarship	Leguizamon and Hammond (2015)	ACT (rounded, first/highest not specified)	±5	Standard error clustered by ACT	Significance in some outcomes
Florida Bright Futures Program	Zhang et al. (2016)	Highest ACT (rounded)	±5	Standard errors	Significance in some outcomes

Note: RDD=Regression Discontinuity Design.

and used robust standard errors instead (Scott-Clayton, 2011; Scott-Clayton & Zafar, 2019). Hence, as suggested by the literature (Kolesár & Rothe, 2018), since bandwidths with the ACT score as the running variable are smaller and contains fewer clusters, it is better to use the robust standard error to make the proper statistical inference. Or if allowed, researchers can also compare both standard errors and try to use the larger one (when those errors are significantly different) to make a more robust inference, just as the procedure in Scott-Clayton (2011).

Conclusion

With increasing popularity in educational policy evaluation, regression discontinuity design is widely applied in many empirical papers. However, considering several specialties of educational settings, applying the regression discontinuity design should be even more careful. Researchers must be aware of the pros and cons of the design and interpret the empirical results to the public audience properly.

In this paper, I discuss the regression discontinuity application with a nationally accepted standardized test, the ACT. I primarily focus on three important issues when using the ACT scores as the running variables: retaking/manipulation, rounding errors, and statistical inference. Many state merit-based financial aid programs use the ACT composite score as the determinant of the aid eligibility. These policies provide great opportunities for researchers to utilize the cut-off to evaluate financial aid and many student outcomes. These outcomes include college choice and enrollment (Bruce & Carruthers, 2014; Leeds & DesJardins, 2015; Zhang et al., 2016), persistence and graduation (Scott-Clayton, 2011; Welch, 2014), and post-graduation performance (Harrington et al., 2016; Leguizamon & Hammond, 2015; Scott-Clayton & Zafar, 2019; Welch, 2014). The cut-offs in test scores provide great local-randomized samples for researchers which help them to claim the causality in their studies.

The ACT score is not a perfect running variable. In many scenarios, the score can be manipulated through retaking. As an important advantage of the regression discontinuity design, the causality will be jeopardized by students' retaking behaviors. What is even worse, due to the discontinuous nature of the score, the extrapolation from the largest value of the score just below the cut-off value to the cut-off value, brings the potential bias from the rounding errors and issues of statistical inference.

Due to data availability, all three issues discussed in this paper do not have perfect solutions at this time. More practically, the following options should receive more attention and the recommendations may not just be limited to the ACT[12] These options seem to be easily implemented, though they have not been widely adopted in some empirical papers discussed above.

First, many state-funded merit-based financial aid has been verified that it can create a strong incentive for retaking (Bruce & Carruthers, 2014; Harrington et al., 2016; Welch, 2014; Zhang et al., 2016). The manipulation in financial aid research is inevitable. Researchers cannot stop students from retaking the standardized tests. Instead, researchers can carefully maintain all the historical test scores of each student and take advantage of these records in regression discontinuity designs (e.g., using the first-time ACT score as the running variable). Second, due to the nature of the ACT, the running variable could not be perfectly continuous. However, researchers can recalculate a better ACT composite score with a smaller interval using four subject scores if the data is accessible. Last, researchers can compare different types of standard errors and choose the most robust ones if it is not confident to argue which standard error can solve the extrapolation and cover the bias properly.

---

[12] The SAT, the other nationally recognized standard tests, also has discrete values and faces manipulation. The SAT even has higher retaking rate than the ACT (Goodman et al., 2018; Harmston & Crouse, 2016).

CHAPTER 4

Retaking Policy Matters: Understanding Equity Gaps in Merit-based Financial Aid

In recent decades, merit-based financial aid programs have been widely adopted by many state governments (Dynarski, 2002; Frisvold & Pitts, 2018; Zhang & Ness, 2010). Most state merit-based financial aid programs set requirements for aid eligibility based upon quantitative measures of the applicants' academic ability such as standardized test scores, high school GPA, high school rank, etc. The rapid growth of state merit-based financial aid programs also raised much scholarly attention, including studies of the potential impacts on college enrollment (Cornwell, Mustard, & Sridhar, 2006; Dynarski, 2000, 2003; Singell et al., 2006), college choice (Bruce & Carruthers, 2014; Cohodes & Goodman, 2014; Dynarski, 2002; Zhang et al., 2016), persistence and graduation (Cohodes & Goodman, 2014; Scott-Clayton, 2011; Welch, 2014), interstate migration (Fitzpatrick & Jones, 2012; Harrington et al., 2016; Leguizamon & Hammond, 2015; Zhang & Ness, 2010), and post-graduation earning (Scott-Clayton, 2011; Scott-Clayton & Zafar, 2019; Welch, 2014). The examination of the impact of merit-based aid on these outcomes brings valuable empirical evidence to support policymakers to improve their programs and adjust their policy goals.

As another key outcome in education, equity, has been raised as an important concern when evaluating state merit-based financial aid programs. Researchers have argued about whether it is equitable and efficient to use state funds to provide discount tuition for relatively affluent students (Heller & Marin, 2004). For example, the merit aid program in Georgia (the HOPE Scholarship) was shown to have large inequities in access to financial aid funding across race, as Black students were less likely to receive the aid while the Asian students had a higher chance to get it (Cornwell & Mustard, 2004). Though the racial difference could be explained by the high school quality measures, the HOPE Scholarship may still reflect equity issues in choosing high schools or preparing the standardized tests since minority students may have difficulties in choosing better high schools and be well-prepared for standardized tests. The merit aid program in New Mexico (the New Mexico Lottery Scholarship) was also shown to be a less cost-effective way to attract minority and low-income students to the University of New Mexico since half of the recipients were non-minority students and 70% of them were from higher-income families (Binder & Ganderton, 2004). Besides, even though some merit aid programs have realized the equity issues and added means-tested eligibility requirements, these programs were still less effective than those merit aid programs only with relatively simple merit-based eligibility requirements (Domina, 2014).

Hence, merit-based financial aid is inequitable overall. Compared to need-based scholarships, merit aid programs are usually not friendly to certain groups of students, particularly minority students and students from low-income families. These underrepresented students are often facing challenges in achieving higher test scores, reducing their probability of being eligible for merit-based financial aid (Goodman et al., 2018; Harrington et al., 2016).

A less discussed cause of inequities in many merit aid programs is that they do not have specific restrictions on whether students can retake the standardized tests to increase their chances of eligibility. The two nationally recognized tests, the SAT and the ACT, allow students to retake their tests multiple times. Therefore, applicants can submit their highest test scores to state agencies for financial aid applications. It is no doubt that the best test score provides a higher probability of being awarded to many students. Since many test retakers are White and Asian students and are less likely from low-income families (Goodman et al., 2018), allowing retaking may make the equity problems in merit aid even more serious.

In this paper, I plan to carefully examine these ignored equity issues in merit-based financial aid under the conditions of different retaking policies. This study will look at how retaking contributes to or dampens equity in merit aid recipients. First, I will review the current retaking policies and student retaking behaviors. Second, I will briefly introduce the Bright Flight Scholarship, the merit-based financial aid in Missouri, and its policy implementation. Last, using the administrative dataset from the Missouri Department of Higher Education, I answer the following research question: Does the acceptance of retaking enlarge the equity gap of the aid eligibility in different racial groups or applicants with various levels of family incomes?

Retaking in Financial Aid Programs

Retaking rates for collegiate standardized tests (e.g., ACT and SAT) have steadily increased in recent years. For example, in 2009, the percentage of students who took multiple ACT tests before graduating from high school was 41%. By 2015, the same percentage had increased to 45% (Harmston & Crouse, 2016). In the SAT, the other major collegiate standardized test, the retaking rate is even higher. Based on over 10 million SAT takers from the high school classes of 2006-2014, Goodman et al. (2018) concluded that 54% of SAT takers retake the SAT at least once. The increasing general retaking rates also raise two concerns in merit-based financial aid. First, what is the "retaking policy" in most financial aid programs? In other words, more practically, how state governments treat scores from different attempts in financial aid applications[1]? Second, what is student retaking behaviors related to merit-based financial aid? Do their behaviors contribute to the final inequity?

When retaking is allowed, there exist multiple ways to define the best test

---

[1] State governments can not stop students from retaking the test, but they may have the authority to determine which score should be accepted. Retaking policy sometimes may be more related to documents in test companies. For example, ACT only allows students to retake the test up to 12, which is the retaking policy in the ACT inc.

score recognized by higher education institutions or state governments. According to a survey of the current landscape of college admissions, institutions may prefer a "most recent", "single highest", or "combined highest" approach in defining what ACT Composite score it considers[2]. Besides, in the SAT, nearly 75% of four-year colleges that accept SAT scores publicly claim to consider only a student's maximum score, which is the "single highest" approach (Goodman et al., 2018; The College Board, 2015). Though some institutions may treat multiple scores differently, since many students can increase their latest test scores through retesting (Harmston & Crouse, 2016), all three approaches seem friendly to retesting and encourage students to retake the test.

However, state governments may not define the best test scores based on different approaches. Focusing on merit aid programs with test score requirements, most of them do not even distinguish the first-time score or the highest score in their program descriptions (Frisvold & Pitts, 2018), which may imply that many of these programs may only use the highest score ("single highest" approach) till the deadline in the system to determine the aid eligibility. With a further investigation about the discontinuity in the density functions of both first-time score and highest score in several empirical papers, I can conclude that the highest score is accepted at least in Florida (Zhang et al., 2016), Missouri (Harrington et al., 2016), and Tennessee (Bruce & Carruthers, 2014; Welch, 2014). Students just below the financial aid threshold are more likely to retake the test, making the density function of the highest score is not continuous around the aid threshold. These discontinuous density functions of the highest score indicate that students from multiple states are responding to these financial aid incentives by retaking the test.

Given the monetary and time costs of retaking these tests, equity concerns

---

[2] See the ACT website for more information about multiple scores: https://www.act.org/content/act/en/products-and-services/the-act-postsecondary-professionals/scores/multi-scores.html.

arise given the disproportionate retaking rates among students of different ethnic groups, family incomes, and parental education levels (Harmston & Crouse, 2016; Mattern & Radunzel, 2019). For example, the percentage of white students who were single-testers was 8% lower than the percentage who were repeat-testers. In contrast, the percentage of Hispanic students who were single-testers was 6% higher than the percentages who were repeat-testers (Harmston & Crouse, 2016). Hence, the disproportionate retaking rates may finally reflect as the different probability of getting financial aid and create increased barriers for certain groups of students to access affordable postsecondary education.

Though the equity gaps have already been examined in general retaking behaviors, there is a lack of empirical discussions about the specific retaking policy and its potential unbalanced impact on financial aid recipients from different demographic groups. In other words, it is not clear whether students from different demographic groups react differently to the incentives from financial aid. Instead of a policy perspective, most previous studies have focused on the discussions related to the overall retaking behaviors. For example, Goodman et al. (2018) found that retaking the SAT could improve students' scores and benefit their college admissions in four-year institutions, particularly for low-income and underrepresented minority students. Using the preference derived from the literature that observed people focusing disproportionately on the leftmost digits of numbers ("left-digit" bias), Pope and Simonsohn (2011) first concluded this "left-digit" bias preference in the SAT-taking context and noticed that students scoring just below multiples of 100 might have a higher retaking rate than those at or above such round number thresholds.

Other studies have concentrated on the determinants or the incentives for making retaking decisions. Vigdor and Clotfelter (2003) used data on applicants of three selective universities. They found that the most common test score ranking policy, using the highest submitted scores, provided the large incentives to retake the SAT.

Zyphur et al. (2007) explored the correlations between retaking and some psychologic factors, the “Big Five” conceptualization of personality. Their results revealed that neuroticism, which is characterized by anxiety and rumination, predicted the number of times an individual took the SAT before attending college. Merit-based financial aid also becomes a significant incentive for retaking the test. Though many studies did not directly study the SAT retaking behaviors, their data analysis have indicated that students just below a financial aid cut-off were more likely to retake the SAT or the ACT than their peers who were just above the cut-off (Bruce & Carruthers, 2014; Harrington et al., 2016; Pantal, 2006; Welch, 2014; Zhang et al., 2016).

This study investigates the retaking behaviors from a policy perspective. By comparing financial aid allocations under different retaking policies, I plan to examine whether the retaking policy results in more serious equity problems. In practice, policymakers cannot stop students from retaking the test since some students also need higher test scores to apply for more selective institutions. However, they may have the authority to decide which score should be accepted so that their policy goals can be satisfied with proper adjustments. This paper fills the gap in the literature regarding the retaking policy in merit-based financial aid, and I hope it can help policymakers to revisit their current retaking policies and consider the potential equity issues that they may never realize.

The Missouri Bright Flight Program

To examine the equity gaps with different retaking policies, I will use the Bright Flight Scholarship Program, the state-funded merit-based financial aid in Missouri, as an example. The Missouri Bright Flight Scholarship Program was established by the Missouri Legislature in 1986. It is a merit-based financial aid program that encourages top-ranked high school graduates to enroll full-time at a participating[3] Missouri

---

[3] For a current list of participating institutions, please see here: https://dhewd.mo.gov/ppc/grants/documents/participatingschools.pdf.

institution, including 13 public 4-year universities, 14 public 2-year institutions, and over 25 private institutions. The eligibility of this financial aid award was intended to be the top 3% of all Missouri ACT takers with an initial amount of approximately $2,000 per academic year. Though the aid has a percentile threshold, the actual threshold was a round number and kept unchanged for decades. Before the 2008-2009 academic year[4], this program required initial students to achieve a composite ACT score of 30 or higher. The minimum ACT score was increased to 31 after the 2008-2009 academic year to fit the top 3% requirement (Kumar, 2008).

The initial amount of the scholarship was $2,000 per academic year and was not adjusted for inflation until the 2007-2008 academic year. In 2007, the authorizing legislation was revised to add the top 4th and 5th percentiles as eligible for a lesser award, becoming a two-tier award structure. The first tier (top 3%) recipients can get $3,000 while the second tier (4%-5%) recipients are awarded only $1,000. Nevertheless, that change did not become effective until the 2010-2011 academic year. It is important to note that the second-tier award has never received funding through the program and the first-tier award was never fully funded (at the maximum amount of $3,000) until the 2015-2016 academic year. From the 2008-2009 academic year to the 2014-2015 academic year, recipients only received approximately $2,500 per year.

To be eligible for this financial aid for the first time, initial applicants must achieve the qualifying score by the June test date immediately following their graduation from high school. Students can retake the ACT multiple times in high school, and only their highest ACT scores are considered. Besides the qualified ACT score, the eligibility also requires applicants to enroll full-time at a participating Missouri university and not pursue a degree or certificate in theology or divinity. Students can renew this financial aid for up to 10 semesters or until completed a bachelor's degree (whichever occurs first) if they keep enrolled full-time and maintain a cumulative

---

[4] The policy change announced in August 2008. Hence, it only affected students who gradated after the summer of 2009 (cohort 2009 and cohort 2010).

GPA over 2.5 and otherwise maintain satisfactory academic progress as defined by enrolled institutions.

Data

The analytic dataset is derived from administrative datasets maintained by the Missouri Department of Higher Education (MDHE). The dataset has every Missouri ACT taker's historical records, including each subject's composite score and four subject scores, graduation date, test time, and other demographic information. The sample includes all Missouri high school students who graduated between 1996 and 2008[5].

Definitions of the Primary Variables

It is noticeable that the ACT composite score is the average value of the four subject scores, rounded to the nearest whole value. Hence, without the specific notifications, the ACT score used in the following empirical analysis is unrounded. The threshold of the financial aid is changed to 29.5, and it will be rounded to 30 in the system of the state government. And the aid eligibility means that students have an unrounded ACT composite score of 29.5 or above.

As to the two equity issues examined in this paper, I recategorize the racial and family income variables following the procedures in Goodman et al. (2018). First, students are separated into three racial groups, underrepresented minority (URM), Asian, and White. Meanwhile, URM students refer to those who identify as Black, Hispanic, Native American, and others (e.g., two or more races)[6]. In my sample,

---

[5] In this paper, cohort 1996 refers the sample with students who graduated in the summer of year 1996. Similar definitions are also assigned to the following cohorts.

[6] The definition of URM students is far from being clear. However, the URM relatively consistent among schools. URM can be defined as a group of students whose percentage of the population in a given group (e.g., all the students in a specific university or college) is lower than their percentage of the population in the whole country. In many universities, URM students usually refer to Black, Hispanic, Native American, Native Hawaiian, and two or more races. See additional explanations at Pennsylvania State University (https://agsci.psu.edu/diversity/awareness/definitions) and California Institute of Technology (https://diversity.caltech.edu/reports-data/urm-definition).

students who are not Asian and White are coded as URM (except missing values).

Second, also suggested by Goodman et al. (2018), I recategorize the family annual income levels into three levels: low-, middle-, and high-incomes. Low-income students are those who report an annual income below \$50,000[7]. These students may be eligible for the free reduced lunch and can receive a test fee waiver. Middle-income students are from families with annual income between \$50,000 and \$100,000. Families with an annual income of more than \$100,000 are coded as high income.

Data Construction

I apply the following criteria to clean the raw data. Specifically, if a student has a test record that does not fit the following criteria, all of his or her records will be excluded:

1. Based on ACT's policy, every student can only take the test up to 12 times. Observations with more than 12 attempts will be excluded.

2. The test date is assumed no later than the graduation date. The latest test date should be before July of the gradation year, which corresponds to the requirement from the state financial aid program[8]. Thus, students are also assumed to take the test as high school students. Hence, the time between the test date and the graduation date should be between 0 and 48 months. Students with test records beyond that period will be excluded. These test takers may probably be untraditional students[9].

3. The demographic information (e.g., race and gender) may not be consistent

---

[7] According to National Center for Education Statistics (NCES), students from families with less than about \$50,000 annual income would usually be eligible for financial aid, addressing the financial needs for receiving higher education (Choy & Bobbitt, 2000, p.19). Hence, I think the annual family income below \$50,000 will be a good indicator to define low-income students.

[8] Missouri Bright Flight Scholarship Program, the state-funded merit aid, addresses that the qualifying ACT score must be achieved by the June test date immediately following the graduation from high school.

[9] For example, students who have enrolled in community colleges may take the ACT again after high school graduation if they want to transfer to other four-year institutions.

among all students' test records. Only the information in the first record will be used. If missing values exist, then the values in the last attempt will replace the missing one. If both records are missing, then the variable will be coded as missing.

4. Students' family income levels are determined by the values in their first record since family income levels may be changed in different years. Observations with missing values in family income levels are also kept.

Analysis

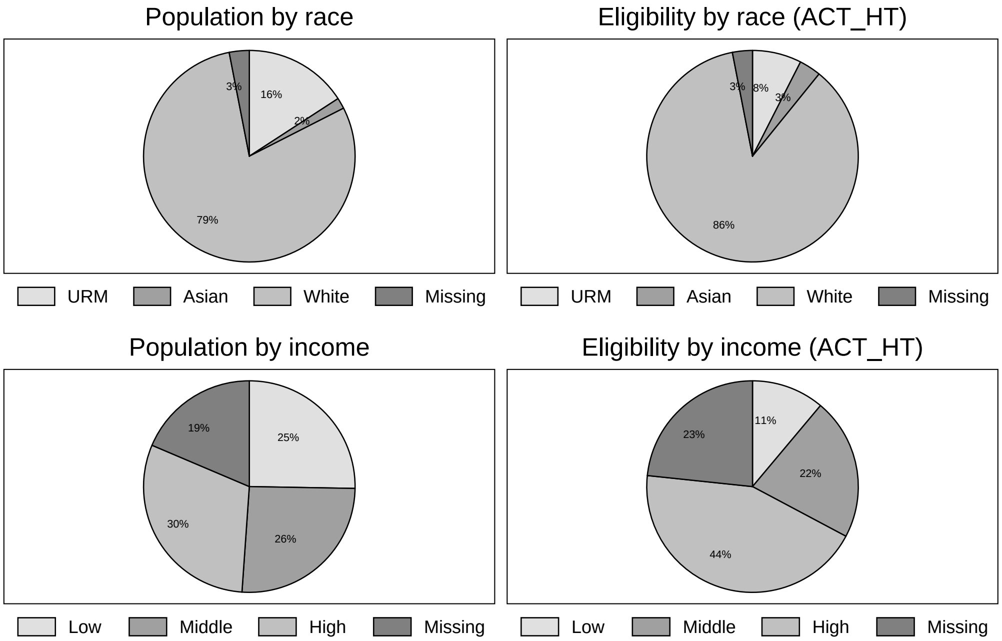
Overall Inequity of the Bright Flight Scholarship

Consistent with other state-funded merit-based financial aid programs (e.g., Georgia and New Mexico), the overall financial aid eligibility of the Missouri Bright Flight Scholarship is also inequitable. Figure 4.1 shows how the Bright Flight eligible students are distributed by race and family income levels based on their highest ACT scores, compared with the compositions of race and family income in the whole sample. The pie charts show that 16% of the whole sample are URM students, but URM students only occupy 8% of all the Bright Flight eligible students. A significant difference is also found in students from low-income families. About a quarter of the whole sample are low-income students, but only 11% of all the Bright Flight eligible students are from low-income families.

The results indicate that these underrepresented students face lots of challenges to meet the eligibility requirement of the Bright Flight Scholarship, which means they may not have enough test scores in their financial aid applications. Figure 4.2 shows how the ACT composite scores are distributed by race and family income levels. In general, the density functions in each sub-figure imply that underrepresented students are probably to get a lower ACT score than their peers in both the first-time and the highest test scores. The difference in the ACT scores will

Figure 4.1

Overall Bright Flight Eligibility by Race and Family Income



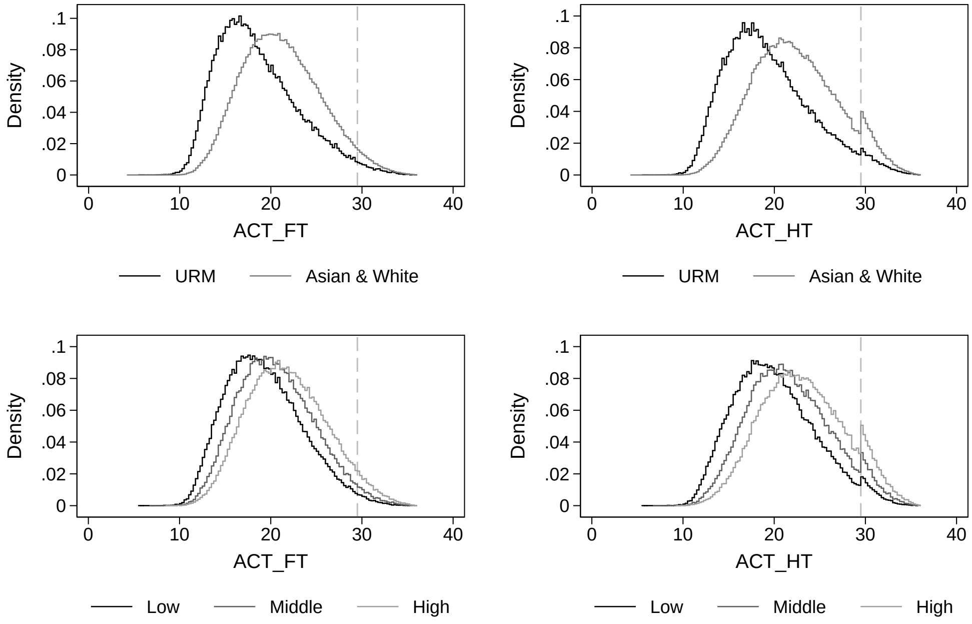
Note: The figure contains four sub-figures. Compositions are based on the whole sample and the restricted sample of the Bright Flight eligibility students. "ACT_FT" or "ACT_HT" means that the eligibility is determined by the first-time ACT or the highest ACT.

directly contribute to the lower percentages of the Bright Flight eligible students in URM and low-income students.

When comparing the density functions of the highest ACT and the first-time ACT, I find that that retaking matters in different demographic groups. The density functions of the first-time ACT scores in different demographic groups are almost continuous, indicating no test score manipulations. Without the disturbance from the retaking behaviors, the density functions of first-time ACT scores may directly show the equity gaps in students' academic performance. URM students and students from low-income families are more likely to achieve lower scores compared to their peers. As a result, they also have difficulties in receiving financial aid. However, the density function of the highest ACT score tells a different story. The significant

Figure 4.2

Histogram of the First-time and the Highest ACT Composite Score, Separated by Race and Family Income



Note: The figure contains four sub-figures. Asian and White students are combined in the two sub-figures in the first row. Groups of "race missing" and "income missing" are not included in those sub-figures. The threshold is 29.5, as the gray dash vertical line in each sub-figure. "ACT_FT" or "ACT_HT" indicates the histogram of the first-time ACT or the highest ACT.

jumps around the financial aid threshold in groups of Asian and White, middle- and high-income indicate that more students in those overrepresented groups benefit from retaking with a much higher percentage of the Bright Flight eligibility. Allowing retaking may enlarge the equity gaps.

Retaking Behaviors in Underrepresented Students

To further investigate how retaking behaviors affect underrepresented students, Table 4.1 displays the summary statistics of retaking rate, average attempts, average ACT scores, and the population of Bright Flight eligibility by different groups of students. Students are compared by different demographic groups, including race and family income levels.

Summary statistics in Table 4.1 show that Asians and students from high-income families have higher retaking rates than other peer groups. The higher probability of retaking also leads to more total attempts. Moreover, retaking behaviors also increase test scores substantially. For example, about half of the low-income students retake the ACT while the same number in the high-income group is nearly three quarters. Consequently, high-income students make substantially more attempts than low-income students, and these additional attempts can finally improve their highest ACT scores by up to 1.35 points. However, low-income students can only increase their scores by up to 0.78 points. It is no doubt that many high-income students will have more advantaged test scores in financial aid application, reflected as the larger population of the Bright Flight eligibility.

Measurement of the Equity Gaps

It is clear that underrepresented students do not take advantage of the retaking policies and may lead to larger equity gaps in being eligible for merit aid. This section focuses on two equity gaps: race and family income level, and I plan to quantify the equity gaps and to compare different retaking policies. In an ideal world without equity issues, the percentage of being aid eligible in each demographic group should be very consistent. Students will have similar possibilities to become eligible for financial aid. The difference in the eligibility rate can be used to describe the equity gap. Hence, I use the following equation to measure equity gaps. A similar method is also adopted in Blom and Monarrez (2020):

$$\begin{aligned}\text{Equity Gap} = {} & \frac{\text{No. of Aid Eligibility in Subgroup A}}{\text{Population of Subgroup A}} \\ & - \frac{\text{No. of Aid Eligibility in Subgroup B}}{\text{Population of Subgroup B}}\end{aligned}$$

Table 4.1

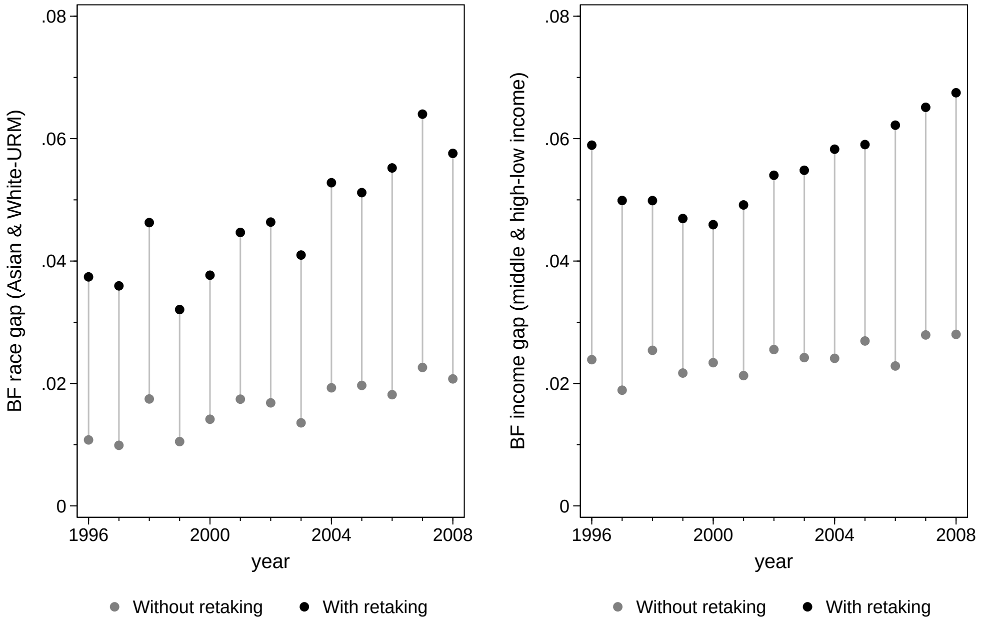
Retaking, Average ACT Scores, and the Bright Flight Eligibility

Variable	Retake rate	Attempts	ACT FT	ACT HT	BF eligible FT	BF eligible HT	Subgroup N
URM	54.66%	1.88	18.45	19.27	1,460	2,936	80,712
Asian	69.41%	2.42	21.55	22.91	713	1,331	8,685
White	65.37%	2.19	21.02	22.20	13,744	33,432	407,869
Race missing	30.50%	1.47	20.02	20.56	839	1,205	16,365
Low income	50.38%	1.82	19.17	19.95	1,723	4,365	130,364
Middle income	64.08%	2.16	20.51	21.62	3,239	8,393	131,714
High income	73.05%	2.37	21.72	23.07	7,464	17,064	155,586
Income missing	60.46%	2.10	20.84	21.96	4,330	9,082	95,967

Note: “Retaking rate” is based on the first-time score, indicating the percentage of students who make additional attempts after their first attempt. “ACT FT” or “ACT HT” is the average score of the first-time or the highest ACT composite score among particular students. “BF eligible FT” or “BF eligible HT” shows the percentage of the students who are aid eligible based on their first-time or their highest ACT composite score. Students who have a raw ACT composite score of 29.5 or above (which will be rounded to 30 later) will be considered as aid eligible. “Subgroup N” is the population in each demographic group.

Figure 4.3

Equity Gaps by Different Retaking Policies



Note: Two sub-figures describe the equity gaps in race and family income levels. To simply the analysis, Asian and White students are combined in the first sub-figure. Students from middle- and high-income families are also combined in the second sub-figure.

For example, if only 8% of the Black students are aid eligible while the same number in the White or Asian students is 10%, the racial equity gap, as the difference in eligibility rate, could be 10%-8%=2%. Figure 4.3 displays this kind of gap by both race and family income levels. By restricting the sample on the first-time and the highest ACT scores, the equity issue exists in both situations. The percentages of URM and low-income students who are eligible based on the first attempt are both lower than their peers, separately. What is worse, allowing retaking and using the highest scores make the gaps even bigger. Compared to the ones based on the first attempt, the gaps will increase by about 3-5% when using the highest ACT. Retaking let more URM and low-income students become aid eligible, but it benefits even more students in other overrepresented groups.

Figure 4.3 also describes the time trend of the gaps. Particularly, the equity gaps have increased in recent years. It may because due to the longer operation time of the financial aid programs, students are more familiar with the aid application process, requirement, and etc. Students from high-income families can collect more information about the financial aid and are more motivated to retake the test after a careful benefit-cost analysis of the retest.

Retaking decisions can be affected by different factors. Students with higher scores may still retake the test when they need a more competitive score to apply for the most selective institutions, such as Ivy League. Students with lower scores may retake the test for multiple reasons, such as placement tests and institutional merit scholarships. Focusing on selected samples more impacted by the financial aid through retaking could be a good supplement for policymakers. Table 4.2 particularly describes the behaviors of selected students who are not aid eligible after their first attempt. The first panel in Table 4.2 includes all students who are not aid eligible based on their first attempt. The second panel narrows the sample with the first-time score of the whole values between 25 and 29 so that other factors that may affect the retaking behaviors may be excluded. Both panels show the existence of all the disadvantages that URM and low-income students are still facing: the lower probability of retaking, fewer attempts, and less increased test scores. Retaking allows more students to be aid eligible. Moreover, in the first panel, only 6.7% of the additional eligible students are URM, while the whole URM population is about 16%.

Similarly, though students from low-income families represent 26% of the whole population, they only occupy about 12% of the additional aid eligible students. The additional aid eligible students are not equally distributed by race and family income levels. When the sample is narrowed down in the second panel, the equity gaps still exist but become smaller. It is because that the restricted sample has more white and

Table 4.2

Retaking, Average ACT Scores, and Additional Bright Flight Eligible Students in Selected Samples

Variable	Retake rate	Attempts	Score increased	BF eligible HT%	Subgroup%	Subgroup N
ACT_FT<30						
URM	54.79%	1.89	0.83	6.66%	15.95%	79,252
Asian	70.63%	2.47	1.42	2.79%	1.60%	7,972
White	65.77%	2.21	1.20	88.89%	79.32%	394,125
Race missing	30.71%	1.48	0.55	1.65%	3.12%	15,526
Low income	50.44%	1.82	0.79	11.93%	25.89%	128,641
Middle income	64.35%	2.17	1.13	23.27%	25.86%	128,475
High income	74.00%	2.40	1.39	43.34%	29.81%	148,122
Income missing	60.98%	2.12	1.14	21.46%	18.44%	91,637
24<ACT_FT<30						
URM	70.21%	2.26	1.10	6.65%	8.66%	7,549
Asian	83.33%	2.84	1.72	2.74%	2.02%	1,764
White	78.57%	2.58	1.41	88.96%	86.50%	75,422
Race missing	43.69%	1.70	0.72	1.64%	2.82%	2,456
Low income	67.09%	2.26	1.09	11.95%	15.83%	13,800
Middle income	77.95%	2.57	1.36	23.42%	23.99%	20,918
High income	82.05%	2.66	1.49	43.54%	39.52%	34,455
Income missing	73.62%	2.46	1.36	21.10%	20.66%	18,018

Note: "Score increased" is the difference between the average value of the highest ACT and the first-time ACT. "BF eligible HT%" shows the percentage of the additional aid eligible students in a certain demographic group among all additional aid eligible students. "Subgroup%" represents the percentage of the specific group of students in the whole population. "Subgroup N" is the population in each demographic group. Panel 1 includes the sample with the raw first-time ACT below 29.5. Panel 2 includes the sample with the raw first-time ACT between 24.5 and 29.5.

high-income students when the first-time ACT score is conditioned on a comparatively higher range.

Conclusion

State-funded merit-based financial aid programs have experienced rapid growth in recent decades. These programs bring many positive impacts on college student outcomes, including college enrollment, degree attainment, and post-graduation earnings Cornwell, Mustard, and Sridhar (2006), Dynarski (2000, 2003), and Scott-

Clayton (2011). However, these merit aid programs also result in many concerns, such as funding source and education equity (Heller & Marin, 2004). It is clear that merit-based financial aid is not designed to secure educational equity. Moreover, non-minority students from high-income families probably benefit more from these programs (Binder & Ganderton, 2004; Cornwell & Mustard, 2004). It is fair to assume that the merit aid's initial inequity from different test scores is inevitable and can be regarded as the trade-off when pursuing other policy goals, such as increasing college enrollment and the prevention of the brain drain. However, policymakers are facing more and more concerns about equity issues in these programs since taking care of underrepresented students weighs much more in current education systems.

Inspired by the gaps of retaking rate in different demographic groups, I notice that the acceptance of the retest may make equity problems in merit aid programs even more serious. The previous analysis has clearly shown that students from different demographic groups are not evenly benefited from the retaking. Students from underrepresented groups have difficulties in retaking the test for many reasons, including additional financial and time costs from the retest (Hyman, 2017). From a policy perspective, it is not easy for policymakers to revise these programs, and the complexity of these programs may also lead to a less effective impact on students (Domina, 2014).

Allowing retaking brings dilemmas to policymakers. On the one hand, retaking gives additional opportunities to underrepresented students to get financial aid. On the other hand, retaking brings more equity issues since underrepresented students are not as competitive as their peers to get financial aid through retaking. Besides, more financial aid recipients also create a heavy financial burden on state governments, and limited funding source has already been a problem for many state merit aid programs (Heller & Marin, 2004). For example, the Missouri Department of Higher Education increased the minimum requirement of the Bright Flight Scholarship after the summer in 2008, when the program had been operated for more than 20 years. Students have to get a composite score of 31 instead of 30. The increased threshold of the aid eligibility relief the financial burden from the high expenditure of the Bright Flight program but makes underrepresented students even more difficult to receive the award.

Here are some suggestions for policymakers to reconsider their retaking policies. First, reducing the cost of retaking for underrepresented students, such as introducing the mandatory college entrance exams[10]. Students, particularly underrepresented ones, can receive reimbursement for the test fee to reduce the cost of retaking. It can encourage them to make more test attempts. Besides, empirical evidence has also indicated that this policy could be a more cost-effective way than traditional student aid to boost postsecondary attainment (Hyman, 2017). Second, reducing the benefit of retaking for less underrepresented students. It may be possible to apply different retaking policies to different groups of students. For example, the financial aid eligibility can be determined by the highest score for underrepresented students and only by the first-time score for other students. With these possible adjustments on retaking policies, the equity problems of the retest can be well controlled. Underrepresented students can also receive more from this financial aid program, which finally may benefit the whole public higher education system.

---

[10] Missouri adopted this policy before. However, it only lasted from 2015 to 2017 (Taketa, 2017).

CHAPTER 5

Conclusions

This dissertation consists of three chapters that examine three different parts of merit-based financial aid research: outcome, method, and sample. Chapter 2 evaluates the impact of merit-based financial aid on students' major choice and degree completion. Chapter 3 revisits the frequently used quasi-experimental research design, regression discontinuity, and provides suggestions for researchers. Chapter 4 concentrates on the sample of test retakers and explores the impact of allowing test retaking on the educational equity of merit aid programs.

These topics are not fully examined in previous studies. Compared to STEM production, researchers are more interested in student outcomes such as college enrollment (Cornwell, Mustard, & Sridhar, 2006; Dynarski, 2000, 2003; Singell et al., 2006), college choice (Bruce & Carruthers, 2014; Cohodes & Goodman, 2014; Dynarski, 2002; Zhang et al., 2016), and persistence and graduation (Cohodes & Goodman, 2014; Scott-Clayton, 2011; Welch, 2014). Besides, it also lacks the application of regression discontinuity in merit-based financial aid research. Many studies focus on the difference between pre-aid and post-aid periods, using the difference-in-differences framework to identify the treatment effect (Cornwell, Lee, et al., 2006; Singell et al., 2006; Sjoquist & Winters, 2015a, 2015b; Zhang, 2011). Further, some technical issues in regression discontinuity applications do not raise enough attention, particularly the manipulation of the running variable. The highest test score is still adopted as the running variable in regression discontinuity design due to data availability. In contrast, the test score manipulation has been widely confirmed in many states, including the SAT in Florida (Zhang et al., 2016), the ACT in Missouri (Harrington et al., 2016) and Tennessee (Bruce & Carruthers, 2014; Welch, 2014). Last, though the overall inequity of merit aid programs has been fully addressed (Binder & Ganderton, 2004; Cornwell & Mustard, 2004), the consequences of different retaking policies on educational equity do not create many discussions. Policymakers need more empirical evidence to support them on policy review and revision.

In conclusion, I believe this dissertation will make contributions to both researchers and policymakers. Chapter 3 focuses on research methodology, which may benefit researchers and guide their studies in the future. Previous studies related to merit-based financial aid and the ACT have indicated the inconsistent procedures of regression discontinuity application. Due to reasons such as data availability, researchers might choose different ACT scores as running variables together with different standard errors for statistical inference. The inconsistent procedures can make the empirical results less convincing and scholars need to pay more attention about those issues when applying a regression discontinuity design in merit-based financial aid research. Chapter 2 and Chapter 4 are empirical analysis and may be more informative for policymakers. Chapter 2 emphasizes a recent important policy goal in the United States: producing more STEM graduates. Due to the shortage of the STEM workforce supply and to be competitive on global market, some states have created initiatives and strategies to assist state citizens in STEM fields. Hence, examining the causal impact of a current long-run policy, such as a merit aid program, on STEM-related outcomes can be very beneficial for policymakers. The conclusions can contribute to revising the current policies, which may bring more motivations to students in choosing STEM programs. Chapter 4 examines a less discussed issue: the impact of allowing retaking on the educational equity of merit aid programs. As a crucial part in the whole financial aid application package, the test scores are very critical to many applicants. Retaking the test can significantly increase the final test score and contribute to a higher chance of being awarded. To secure equity, policymakers should be aware of the disproportionate retaking rates among students from different demographic groups. It is also necessary for them to understand the retaking behaviors so that they can revise relevant financial aid policies to benefit underrepresented students with more effectiveness.

APPENDIX A

CIP Codes and Field of Study

The Classification of Instructional Programs (CIP) codes were developed by the U.S. Department of Education as the national taxonomic standard of academic program titles. It is widely used for federal surveys and state reporting. All CIP codes are searchable on the official website of the National Center for Education Statistics (NCES)[1]. The up-to-date version is published in 2020. However, in this dissertation, I use the 2010 version since the data collection was finished much earlier before 2020.

STEM and Engineering CIP codes

Researchers may have slightly different definitions of STEM programs. I borrow the categories in Darolia et al. (2020) and the detail STEM CIP codes are presented in Table A.1. Engineering, computer science, and technology are treated as the core engineering majors. Besides those core engineering majors, STEM also include majors in agricultural and animal sciences, natural science, biological science, mathematics, military/security science, physical science, psychology, business, social science, and health science, etc.

Non-STEM CIP Codes

The rest CIP codes are defined as non-STEM majors. However, to explore the different major choice behaviors between male and female students, I examine the effect of the financial aid on choosing some specific non-STEM majors. Based on the CIP codes' first two digits, I only choose three big categories with the most enrolled students. They are journalism (CIP2 09), interdisciplinary studies (CIP2

---

[1] See here: https://nces.ed.gov/ipeds/cipcode/.

30), and business (CIP2 52). Each of them has about 1,000 enrolled students in my preferred sample for regression discontinuity analysis. Though some specific majors in journalism and business are recorded as STEM, they only have a few students (<0.5%), which will probably not disturb my analysis of non-STEM major choice.

Table A.1

STEM and Engineering CIP Codes

STEM	Engineering	CIP
Agricultural & animal science		010308; 010901; 010902; 010903; 010904; 010905; 010906; 010907; 010999; 011001; 011002; 011099; 011101; 011102; 011103; 011104; 011105; 011106; 011199; 011201; 011202; 011203; 011299;
Natural resource		030101; 030103; 030104; 030199; 030205; 030502; 030508; 030509;
Computer science	X	CIP begins with 11
Engineering	X	CIP begins with 14
Technology	X	CIP begins with 15
Biological science		CIP begins with 26
Mathematics		CIP begins with 27
Military and security science		280501; 280502; 280505; 290201; 290202; 290203; 290204; 290205; 290206; 290207; 290299; 290301; 290302; 290303; 290304; 290305; 290306; 290307; 290399; 290401; 290402; 290403; 290404; 290405; 290406; 290407; 290408; 290409; 290499; 299999; 430106; 430116;
Physical science		CIP begins with 40
Psychology		422701; 422702; 422703; 422704; 422705; 422706; 422707; 422708; 422709; 422799;
Business and social science		450301; 450603; 450702; 521301; 521302; 521304; 521399;
Health science		511002; 511005; 511401; 512003; 512004; 512005; 512006; 512007; 512009; 512010; 512202; 512205; 512502; 512503; 512504; 512505; 512506; 512510; 512511; 512706;
Other STEM		040902; 090702; 100304; 130501; 130601; 130603; 300101; 300601; 300801; 301001; 301701; 301801; 301901; 302501; 302701; 303001; 303101; 303201; 303301; 410000; 410101; 410204; 410205; 410299; 410301; 410303; 410399; 419999; 490101;

BIBLIOGRAPHY

Ahn, T., Arcidiacono, P., Hopson, A., & Thomas, J. R. (2019). *Equilibrium grade inflation with implications for female interest in stem majors* (NBER Working Paper Series No. 26556). National Bureau of Economic Research. https://doi.org/10.3386/w26556

Almond, D., & Doyle, J. J. (2011). After midnight: A regression discontinuity design in length of postpartum hospital stays. *American Economic Journal: Economic Policy*, *3*(3), 1–34. https://doi.org/10.1257/pol.3.3.1

American Institutes for Research. (2013). *How much does it cost institutions to produce STEM degrees?* (Data Brief). American Institutes for Research. Ithaca, NY. https://deltacostproject.org/sites/default/files/products/Cost%20to%20Institutions%20of%20STEM%20Degrees.pdf

Andrews, R., & Stange, K. (2016). *Price regulation, price discrimination, and equality of opportunity in higher education: Evidence from Texas* (NBER Working Paper Series No. 22901). National Bureau of Economic Research. https://doi.org/10.3386/w22901

Angrist, J. D., & Pischke, J. (2008). *Mostly harmless econometrics: An empiricist's companion.* Princeton, NJ: Princeton University Press.

Arcidiacono, P., Aucejo, E. M., & Spenner, K. (2012). What happens after enrollment? An analysis of the time path of racial differences in GPA and major choice. *IZA Journal of Labor Economics*, *1*(1). https://doi.org/10.1186/2193-8997-1-5

Bajari, P., Hong, H., Park, M., & Town, R. (2011). *Regression discontinuity designs with an endogenous forcing variable and an application to contracting in health care* (NBER Working Paper Series No. 17643). National Bureau of Economic Research. https://doi.org/10.3386/w17643

Barreca, A. I., Guldi, M., Lindo, J. M., & Waddell, G. R. (2011). Saving babies? Revisiting the effect of very low birth weight classification. *Quarterly Journal of Economics*, *126*(4), 2117–2123. https://doi.org/10.1093/qje/qjr042

Binder, M., & Ganderton, P. T. (2004). The New Mexico lottery scholarship: Does it help minority and low-income students (D. E. Heller & P. Marin, Eds.). *State merit scholarship programs and racial inequality*, 101–122. https://files.eric.ed.gov/fulltext/ED489183.pdf

Blom, E., & Monarrez, T. (2020). *Understanding equity gaps in college graduation*. Urban Institute. https://www.urban.org/sites/default/files/publication/101638/understanding_equity_gaps_in_college_graduation.pdf

Bruce, D. J., & Carruthers, C. K. (2014). Jackpot? The impact of lottery scholarships on enrollment in Tennessee. *Journal of Urban Economics*, *81*, 30–44. https://doi.org/10.1016/j.jue.2014.01.006

Calonico, S., Cattaneo, M., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, *82*(6), 2295–2326. https://doi.org/10.3982/ECTA11757

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, *90*(3), 414–427. https://doi.org/10.1162/rest.90.3.414

Castleman, B. L., Long, B. T., & Mabel, Z. (2018). Can financial aid help to address the growing need for STEM education? The effects of need-based grants on the completion of science, technology, engineering, and math courses and degrees. *Journal of Policy Analysis and Management*, *37*(1), 136–166. https://doi.org/10.1002/pam.22039

Chen, X., & Weko, T. (2009). *Students who study science, technology, engineering, and mathematics (STEM) in postsecondary education* (NCES Stats in Brief

No. 2009-161). National Center for Education Statistics. https://nces.ed.gov/pubs2009/2009161.pdf

Choy, S. P., & Bobbitt, L. (2000). *Low-income students: Who they are and how they pay for their education* (NCES Statistical Analysis Report No. 2000-169). National Center for Education Statistics. https://nces.ed.gov/pubs2000/2000169.pdf

Cohodes, S., & Goodman, J. (2014). Merit aid, college quality, and college completion: Massachusetts' Adams scholarship as an in-kind subsidy. *American Economic Journal: Applied Economics*, *6*(4), 251–285. https://doi.org/10.1257/app.6.4.251

Cornwell, C., Lee, K. H., & Mustard, D. B. (2006). *The effects of state-sponsored merit scholarships on course selection and major choice in college* (IZA Discussion Papers No. 1953). Institute of Labor Economics. https://www.econstor.eu/bitstream/10419/33322/1/507466136.pdf

Cornwell, C., & Mustard, D. B. (2004). Georgia's HOPE scholarship and minority and lowincome students: Program effects and proposed reforms (D. E. Heller & P. Marin, Eds.). *State merit scholarship programs and racial inequality*, 77–100. https://files.eric.ed.gov/fulltext/ED489183.pdf

Cornwell, C., Mustard, D. B., & Sridhar, D. J. (2006). The enrollment effects of merit-based financial aid: Evidence from Georgia's HOPE Program. *Journal of Labor Economics*, *24*(4), 761–786. https://doi.org/10.1086/506485

Darolia, R., Koedel, C., Main, J. B., Ndashimye, J. F., & Yan, J. (2020). High school course access and postsecondary STEM enrollment and attainment. *Educational Evaluation and Policy Analysis*, *42*(1), 22–45. https://doi.org/10.3102/0162373719876923

Denning, J. T., & Turley, P. (2017). Was that SMART? Institutional financial incentives and field of study. *Journal of Human Resources*, *52*(1), 152–186. https://doi.org/10.3368/jhr.52.1.0414-6340R1

Domina, T. (2014). Does merit aid program design matter? A cross-cohort analysis. *Research in Higher Education*, *55*(1), 1–26. https://doi.org/10.1007/s11162-013-9302-y

Dong, Y. (2015). Regression discontinuity applications with rounding errors in the running variable. *Journal of Applied Econometrics*, *30*(3), 422–446. https://doi.org/10.1002/jae.2369

Dynarski, S. (2000). Hope for whom? Financial aid for the middle class and its impact on college attendance. *National Tax Journal*, *53*(3), 629–661. https://doi.org/10.17310/ntj.2000.3s.02

Dynarski, S. (2002). *The consequences of merit aid* (NBER Working Paper Series No. 9400). National Bureau of Economic Research. https://doi.org/10.3386/w9400

Dynarski, S. (2003). Does aid matter? Measuring the effect of student aid on college attendance and completion. *American Economic Review*, *93*(1), 279–288. https://doi.org/10.1257/000282803321455287

Ehrenberg, R. G. (2011). *2011 survey of differential tuition at public higher education institutions*. Cornell University. https://archive.ilr.cornell.edu/download/8606

Evans, B. J. (2017). SMART money: Do financial incentives encourage college students to study science? *Education Finance and Policy*, *12*(3), 342–368. https://doi.org/https://doi.org/10.1162/EDFP_a_00199

Fayer, S., Lacey, A., & Watson, A. (2017). *STEM occupations: Past, present, and future*. U.S. Bureau of Labor Statistics. https://www.bls.gov/spotlight/2017/science-technology-engineering-and-mathematics-stem-occupations-past-present-and-future/pdf/science-technology-engineering-and-mathematics-stem-occupations-past-present-and-future.pdf

Fiegener, M. K. (2015). *Science and Engineering Degrees: 1966-2012*. The National Science Foundation. http://www.nsf.gov/statistics/2015/nsf15326/

Fitzpatrick, M. D., & Jones, D. (2012). *High education, merit-based scholarships and post-baccalaureat migration* (NBER Working Paper Series No. 18530). National Bureau of Economic Research. https://doi.org/10.3386/w18530

Freeman, J. A., & Hirsch, B. T. (2008). College majors and the knowledge content of jobs. *Economics of Education Review*, *27*(5), 517–535. https://doi.org/10.1016/j.econedurev.2007.07.004

Frisvold, D. E., & Pitts, M. (2018). *State merit aid programs and youth labor market attachment* (NBER Working Paper Series No. 24662). National Bureau of Economic Research. https://doi.org/10.3386/w24662

Gelman, A., & Imbens, G. W. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business and Economic Statistics*, *37*(3), 447–456. https://doi.org/10.1080/07350015.2017.1366909

Goodman, J., Gurantz, O., & Smith, J. (2018). *Take Two! SAT retaking and college enrollment gaps* (NBER Working Paper Series No. 24945). National Bureau of Economic Research. https://doi.org/10.3386/w24945

Goodman, J., Melkers, J., & Pallais, A. (2019). Can online delivery increase access to education? *Journal of Labor Economics*, *37*(1), 1–34. https://doi.org/10.1086/698895

Griffith, A. L. (2010). Persistence of women and minorities in STEM field majors: Is it the school that matters? *Economics of Education Review*, *29*(6), 911–922. https://doi.org/10.1016/j.econedurev.2010.06.010

Harmston, M., & Crouse, J. (2016). *Multiple testers: What do we know about them?* (ACT Research & Policy). ACT. https://www.act.org/content/dam/act/unsecured/documents/5195-Multiple-Testers.pdf

Harrington, J. R., Muñoz, J., Curs, B. R., & Ehlert, M. (2016). Examining the impact of a highly targeted state administered merit aid program on brain drain: Evidence from a regression discontinuity analysis of Missouri's Bright Flight program. *Research in Higher Education*, *57*, 423–447. https://doi.org/10.1007/s11162-015-9392-9

Heller, D. E., & Marin, P. (Eds.). (2004). *State merit scholarship programs and racial inequality*. The Civil Rights Project at Harvard University. https://files.eric.ed.gov/fulltext/ED489183.pdf

Huang, C., Zhang, S., & Zhao, Q. (2020). The early bird catches the worm? School entry cutoff and the timing of births. *Journal of Development Economics*, *143*, Article 102386. https://doi.org/10.1016/j.jdeveco.2019.102386

Hyman, J. (2017). Act for all: The effect of mandatory college entrance exams on postsecondary attainment and choice. *Education Finance and Policy*, *12*(3), 281–311. https://doi.org/10.1162/EDFP_a_00206

Imbens, G. W., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, *79*(3), 933–959. https://doi.org/10.1017/CBO9781107415324.004

Kolesár, M., & Rothe, C. (2018). Inference in regression discontinuity designs with a discrete running variable. *American Economic Review*, *108*(8), 2277–2304. https://doi.org/10.1257/aer.20160945

Kumar, K. (2008). Bright Flight scholarships will require 31 on ACT. *St. Louis Post-Dispatch*. https://www.stltoday.com/news/local/education/bright-flight-scholarships-will-require-31-on-act/article_f5f24778-a07e-5cac-bffd-5c0ee43dc103.html

Lee, D. S., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, *142*(2), 655–674. https://doi.org/10.1016/j.jeconom.2007.05.003

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(2), 281–355. https://doi.org/10.1257/jel.48.2.281

Leeds, D. M., & DesJardins, S. L. (2015). The effect of merit aid on enrollment: A regression discontinuity analysis of Iowa's National Scholars Award. *Research in Higher Education*, *56*(5), 471–495. https://doi.org/10.1007/s11162-014-9359-2

Leguizamon, J. S., & Hammond, G. W. (2015). Merit-based college tuition assistance and the conditional probability of in-state work. *Papers in Regional Science*, *94*(1), 197–218. https://doi.org/10.1111/pirs.12053

Malcom, S., & Feder, M. (2016). *Barriers and opportunities for 2-year and 4-year STEM degrees: Systemic change to support students' diverse pathways*. National Academies Press. https://doi.org/10.17226/21739

Mattern, K., & Radunzel, J. (2019). *Does superscoring increase subgroup differences ?* (ACT Research & Policy). ACT. https://www.act.org/content/dam/act/unsecured/documents/R1774-superscoring-subgroup-2019-07.pdf

McCall, B. P., & Bielby, R. M. (2012). Regression discontinuity design: Recent developments and a guide to practice for researchers in higher education. In J. C. Smart & M. B. Paulsen (Eds.), *Higher education: Handbook of theory and research* (pp. 249–290). Springer.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*(2), 698–714. https://doi.org/10.1016/j.jeconom.2007.05.005

Missouri Department of Higher Education. (2010). *DHE financial assistance and outreach*. Missouri Department of Higher Education. https://dhe.mo.gov/files/research/statsum/table23_25_0910.pdf

Montmarquette, C., Cannings, K., & Mahseredjian, S. (2002). How do young people choose college majors? *Economics of Education Review*, *21*(6), 543–556. https://doi.org/10.1016/S0272-7757(01)00054-1

National Science Board. (2018). *Science & Engineering Indicators 2018*. The National Science Foundation. https://www.nsf.gov/statistics/2018/nsb20181/assets/901/science-and-engineering-labor-force.pdf

Nguyen, T. D., Kramer, J. W., & Evans, B. J. (2019). The effects of grant aid on student persistence and degree attainment: A systematic review and meta-analysis of the causal evidence. *Review of Educational Research*, *89*(6), 831–874. https://doi.org/10.3102/0034654319877156

Pantal, M. A. (2006). *Retaking ACT: The effect of a state merit-based scholarship on test-taking, enrollment, retention, and employment* (Doctoral dissertation). University of Missouri-Columbia.

Pope, D., & Simonsohn, U. (2011). Round numbers as goals: Evidence from baseball, SAT takers, and the lab. *Psychological Science*, *22*(1), 71–79. https://doi.org/10.1177/0956797610391098

Price, J. (2010). The effect of instructor race and gender on student persistence in STEM fields. *Economics of Education Review*, *29*(6), 901–910. https://doi.org/10.1016/J.ECONEDUREV.2010.07.009

Rothstein, J., & Rouse, C. E. (2011). Constrained after college: Student loans and early-career occupational choices. *Journal of Public Economics*, *95*(1-2), 149–163. https://doi.org/10.1016/j.jpubeco.2010.09.015

Schochet, P., Cook, T., Deke, J., Imbens, G. W., Lockwood, J., Porter, J., & Smith, J. (2010). *Standards for regression discontinuity designs*. https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_rd.pdf

Scott-Clayton, J. (2011). On money and motivation: A quasi-experimental analysis of financial incentives for college Achievement. *Journal of Human Resources*, *46*(3), 614–646. https://doi.org/10.3368/jhr.46.3.614

Scott-Clayton, J., & Schudde, L. (2019). The consequences of performance standards in need based aid: Evidence from community colleges. *Journal of Human Resources*, (March), 0717–8961r2. https://doi.org/10.3368/jhr.55.4.0717-8961r2

Scott-Clayton, J., & Zafar, B. (2019). Financial aid, debt management, and socioeconomic outcomes: Post-college effects of merit-based aid. *Journal of Public Economics*, *170*, 68–82. https://doi.org/10.1016/J.JPUBECO.2019.01.006

Singell, L. D., Waddell, G. R., & Curs, B. R. (2006). HOPE for the Pell? Institutional effects in the intersection of merit-based and need-based aid. *Southern Economic Journal*, *73*(1), 79–99. https://doi.org/10.2307/20111875

Sjoquist, D. L., & Winters, J. V. (2015a). State merit aid programs and college major: A focus on STEM. *Journal of Labor Economics*, *33*(4), 973–1006. https://doi.org/10.1086/681108

Sjoquist, D. L., & Winters, J. V. (2015b). The effect of Georgia's HOPE scholarship on college major: a focus on STEM. *IZA Journal of Labor Economics*, *4*(1), 15. https://doi.org/10.1186/s40172-015-0032-6

Stange, K. (2015). Differential pricing in undergraduate education: Effects on degree production by field. *Journal of Policy Analysis and Management*, *34*(1), 107–135. https://doi.org/10.1002/pam.21803

Stater, M. (2011). Financial aid, student background, and the choice of first-year college major. *Eastern Economic Journal*, *37*(3), 321–343. https://doi.org/10.1057/eej.2009.41

Stinebrickner, R., & Stinebrickner, T. R. (2004). Time-use and college outcomes. *Journal of Econometrics*, *121*(1-2), 243–269. https://doi.org/10.1016/j.jeconom.2003.10.013

Taketa, K. (2017). Missouri will no longer offer the ACT for free to juniors. *St. Louis Post-Dispatch*. https://www.stltoday.com/news/local/education/missouri-will-no-longer-offer-the-act-for-free-to-juniors/article_78c14bd8-5d45-5af1-8458-9b7b5597a034.html

The College Board. (2014). *Trends in College Pricing 2013*. College Board. https://research.collegeboard.org/pdf/trends-college-pricing-2013-full-report.pdf

The College Board. (2015). *SAT score-use practices by participating institution*. College Board. https://secure-media.collegeboard.org/digitalServices/pdf/professionals/sat-score-use-practices-participating-institutions.pdf

University of Missouri System. (2006). *University of Missouri System tuition rates (FY06-07)*. University of Missouri System. https://www.umsystem.edu/media/fa/budget/2007educfees.pdf

Vigdor, J. L., & Clotfelter, C. T. (2003). Retaking the SAT. *Journal of Human Resources*, *38*(1), 1–33. https://doi.org/10.3368/jhr.xxxviii.1.1

Wang, X. (2013). Why students choose STEM majors: Motivation, high school learning, and postsecondary context of support. *American Educational Research Journal*, *50*(5), 1081–1121. https://doi.org/10.3102/0002831213488622

Welch, J. G. (2014). HOPE for community college students: The impact of merit aid on persistence, graduation, and earnings. *Economics of Education Review*, *43*, 1–20. https://doi.org/10.1016/j.econedurev.2014.08.001

Wiswall, M., & Zafar, B. (2015). Determinants of college major choice: Identification using an information experiment. *Review of Economic Studies*, *82*(2), 791–824. https://doi.org/10.1093/restud/rdu044

Zafar, B. (2013). College major choice and the gender gap. *Journal of Human Resources*, *48*(3), 545–595. https://doi.org/10.1353/jhr.2013.0022

Zhang, L. (2011). Does merit-based aid affect degree production in STEM fields? Evidence from Georgia and Florida. *The Journal of Higher Education*, *82*(4), 389–415. https://doi.org/10.1353/jhe.2011.0024

Zhang, L., Hu, S., Sun, L., & Pu, S. (2016). The effect of Florida's Bright Futures program on college choice: A regression discontinuity approach. *The Journal of Higher Education*, *87*(1), 115–146. https://doi.org/10.1353/jhe.2016.0003

Zhang, L., & Ness, E. C. (2010). Does State Merit-Based Aid Stem Brain Drain? *Educational Evaluation and Policy Analysis*, *32*(2), 143–165. https://doi.org/10.3102/0162373709359683

Zyphur, M. J., Islam, G., & Landis, R. S. (2007). Testing 1, 2, 3, ...4? The personality of repeat SAT test takers and their testing outcomes. *Journal of Research in Personality*, *41*(3), 715–722. https://doi.org/10.1016/j.jrp.2006.06.005

VITA

Junpeng Yan was born in Jiangsu, China. He received a bachelor's degree in statistics and a master's degree in economics at Shanghai University of Finance and Economics. He worked as a graduate research assistant in the Department of Educational Leadership and Policy Analysis at the University of Missouri. His research focuses on using quantitative tools to analyze the effects of educational policy on college students. Upon the completion of his Ph.D. in higher education policy, Junpeng plans to work as a policy analyst in Shanghai, China.