# Essays on the Economics of Higher Education and Employment

Seung Eun Park

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

# ABSTRACT

Essays on the Economics of Higher Education and Employment

Seung Eun Park

This dissertation studies legal and institutional policies that help to reduce the barriers to educational attainment and employment. The first chapter examines the effect of availability of juvenile record laws on education attainment and employment using state statue revisions after the passage of the federal Second Chance Act. The second chapter examines enrollment patterns of students who drop out from community colleges and identify four typologies of college dropouts and important factors that contribute to college success. The third chapter estimates the impact of federal Pell Grant eligibility on financial aid packages, labor supply while in schools, and academic outcomes for community college students. The three chapters together shed light on how federal, state, and institutional policies can help reduce the academic and employment barriers for the marginalized population in the United States.

# Contents

# List of Figures

# List of Tables

# Acknowledgements

This work would not have been completed without the help and support from my dissertation committee members, colleagues and friends, and my family. First and foremost, I would like to express my deepest gratitude to my academic advisor, Judith Scott-Clayton. Judy gave me the opportunity to work with her as a research assistant at the Community College Research Center (CCRC) and co-authoring one of my dissertation chapters. She has been an amazing professor, academic mentor, and advisor. I have benefited tremendously from her guidance and unconditional support in every step of my Ph.D. journey.

I am deeply grateful for Peter Bergman, my second reader, who has motivated my research interests from the beginning of my dissertation. His insightful feedback and methodological advice gave me the strength to pursue deeper into my research. I am thankful for Jordan Matsudaira and Bryan Keller for their very helpful comments that significantly contributed to advancing my papers. I also benefited greatly from Dan O'Flaherty, who introduced me to the field in the economics of crime, which fueled my motivation, excitement, and enjoyment in the research topic.

I have benefited enormously from conversations from other faculty members and CCRC staffs. I am grateful to Thomas Bailey, Alex Bowers, Clive Belfield, Alex Eble, Joydeep Roy, Nikki Edgecombe, Young Sun Lee, Ilja Cornelisz, and Susana Loeb for sharing their comments. I would also like to thank my research assistants, Adnan Moussa, Casey Nguyen, and Omar Rodriguez-Esparza, who went through various legal documents and helped me in categorizing different laws in my job market paper.

My deep appreciation goes to Robert On, who is my former colleague, mentor, and best

friend. We have exchanged many conversations about research, life, and values. I can not imagine my Ph.D. life without him being involved. He gave me strength through my turmoils and supported me through technical conversations while developing my research papers.

I also thank my Ph.D. and CCRC colleagues Florence Ran, Veronica Minaya, Beth Kopko, Rachel Yang, Yilan Pan, Amy Brown, and Vivian Liu, who shared their support and encouragement. Many thanks to my former colleagues and friends who trusted in me, valued my life decisions, and encouraged persistence throughout. To name a few, Jihoon Hyun, Hannah Lee, Rachel Yoo, Dawhe Park, Albert No, Hyechang Lee, and Erica Chung.

I am very lucky to have met my friends at Milal, New York. They shared their happiness and love that encouraged me throughout my Ph.D. journey. I hope our society becomes a better world by appreciating and loving persons with intellectual disabilities.

Lastly, words can not describe the support and love of my parents, aunt, and uncle. I am indebted to their support and love emotionally, financially, and maturity.

*This dissertation is dedicated to my parents, family, and*

*Milal Mission in New York,*

*who shared their endless love, strength, and laugh through the hills and valleys of my Ph.D. journey*

# Preface

The United States continues to face achievement and employment gaps by race. In 2015-2016, the adjusted cohort graduation rate among white, public high school students was 88% while black, public high school students graduated at a rate of 76% (Digest of Education Statistics, 2017, table 219.46). At the college level, white students achieved six-year completion rates that were 1.5 times higher than those observed for black students (62 percent versus 37 percent) (Shapiro, et al., 2017). Similarly, despite historically low national unemployment levels (Bureau of Labor Statistics, 2018, Report 1076), the unemployment rate for blacks is about two times higher than the unemployment rate for whites (7.5 versus 3.8 percent). In my dissertation, I focus on three marginalized populations who are predominantly comprised of racial minorities, ex-juvenile offenders, community college dropouts, and low-income students. I examine legal and institutional policies that can help reduce the barriers for these populations. In particular, I provide empirical evidence that (1) information sharing with the justice system and schools can help improve educational attainment in the aggregate, (2) separating out by dropout typologies can help better identify specific factors that contribute to completion, and (3) a modest Pell grant can help reduce the need to work while enrolled in college and thereby increase enrollment intensity.

Chapter one examines the effect of availability of juvenile records on high school graduation, the probability of ever attending college, and employment. In this chapter I evaluate state revisions to juvenile record laws (39 states) after the passage of the federal Second Chance Act (Y2008-2015) (39 states), which were intended to help re-entry of ex-juvenile offenders. After categorizing the laws into four types - information sharing with schools, information sharing across interagency, limiting information sharing with the public, and expanding sealing and expungement eligibility - I utilize a difference-in-difference strategy to estimate the effects of each four types on educational outcomes and employment. I find significant and positive impacts of school notification laws on both educational outcomes. In addition, I find suggestive positive effects of expanding sealing and expungement eligibility on employment. By contrast, I find no robust impact from reforms that shared information across interagency and limited public access to juvenile records. This study contributes to the economic debate on information sharing and confidentiality in the justice system.

In chapter two, I use cluster analysis to identify four typologies of community college dropouts using enrollment patterns: college trials (dropouts after one semester of trials), high credit attempting medium dropouts, low credit attempting medium dropouts, and late dropouts. Matching students who complete to students who dropout with similar enrollment patterns, I identify different sets of variables that are important to predicting completion. Specifically, using gradient boosting algorithm, I find third-year enrollment variables (i.e., credits withdrawn, full-time status, and remedial credits attempted) largely contribute to predictions of completion among late dropouts. On the contrary, math and science credits earned in the first and second year are most predictive of completion among high credit attempted medium dropouts.

Chapter three, co-authored paper with Judith Scott-Clayton, examines the effects of receiving modest Pell Grant ($500) aid on financial aid packages, labor supply while in school, and academic outcomes for community college students. The federal Pell grant program is the nation's largest source of grant aid for students from lower income families. We compare community college students just above and below the expected family contribution (EFC) cutoff for receiving a Pell Grant and find that other financial aid adjusts in ways that vary by institution: students at schools that offer federal loans borrowed more if they just missed the Pell eligibility threshold, but at other schools, students who just missed the cutoff for Pell were compensated with higher state grants. Focusing on the loan-offering

schools where students face a discontinuity in total grant aid, we find suggestive evidence that receiving a modest Pell Grant instead of additional loans leads students to reduce labor supply and increase enrollment intensity. We also provide indirect evidence that students' initial enrollment choices are influenced by an offer of Pell Grants versus loans.

To summarize, together these three chapters provide empirical evidence for legal and institutional policies that can help reduce barriers in educational attainment and employment for marginalized populations. The findings of my dissertation recommend that (i) states and institutions collaboratively build an information sharing system between the justice and school systems to better aid students for successful re-entry, (ii) institutions create standard practices in identifying at-risk dropout types and employ specific interventions for completion, and (iii) institutions promote need-based grant aid over additional loans.

# Chapter 1

# The Effects of Revisions to Juvenile Record Laws on Education Outcomes and Employment after the Second Chance Act of 2007

Rina Seung Eun Park

# I Introduction

On April 9, 2008, the Second Chance Act of 2007 (H.R. 1593) was signed into law under the title “to reauthorize the grant program for reentry of offenders into the Omnibus Crime Control and Safe Streets Act of 1968, to improve reentry planning and implementation, and for other purposes.” The Second Chance Act authorized up to $165 million in federal grants to state and local government agencies under the stated goal of reducing recidivism, increasing public safety, and promoting successful re-integration (The Council of State Government Justice Center, 2018). Since then, the majority of states have made significant changes to their criminal laws, including those related to juvenile justice.

During the years between 2008 and 2015, 39 states made revisions to their juvenile records legislation.[1] These revisions include strengthening protections of juvenile records from the public, expanding eligibility and modifying procedures needed for sealing and expungement, expanding interagency[2] information sharing, and expanding information sharing to schools.[3] Protecting juvenile records from being accessed by the general public is one way to mitigate unnecessary collateral damage a juvenile might face after being released and transitioning to adulthood. Similarly, sealing and expungement procedures offer another layer of protection by sealing or erasing one’s youthful criminal history from the public (Shah, Fine, & Gullen, 2014). On the other hand, information sharing between agencies that serve and arrest juvenile offenders has been promoted as means to identify at-risk students, provide early

[1] Author’s calculation using the National Conference of State Legislature’s Juvenile Justice Bills Tracking Database, available at http://www.ncsl.org/research/civil-and-criminal-justice/ncsls-juvenile-justice-bill-tracking-database.aspx. The sample included 51 states, including the District of Columbia.

[2] *Interagency* refers to any individual or institution that can provide information in the juvenile justice process. For example, interagency information sharing includes interactions between social workers and probation officers, law enforcement and the Bureau of Immigration and Customs Enforcement, and Juvenile and Domestic Relations District Court Records and the Department of Motor Vehicles.

[3] During the time period of this study, six states (Arizona, Florida, Louisiana, Minnesota, New Hampshire, and Virginia) allowed law enforcement to collect personal information such as photographs and fingerprints. The majority of these laws were enacted in 2008. As there is no obvious mechanism between these laws and education and employment outcomes, this paper does not include these laws in the analysis.

supervision and treatment, avoid juvenile court involvement, and make informed decisions with more information (Griffin, 2000; Teske, 2011; Wachter, 2017).

A large body of research demonstrates that having a juvenile record and sharing that information can have negative consequences on high school graduation, college enrollment, and employment. For an example, conviction (incarceration) while in high school can double (quadruple) the odds of dropping out from high school (Aizer & Doyle Jr, 2015; Hjalmarsson, 2008; Sweeten, 2006). College and financial aid applications that ask about the applicant's criminal history can limit college access for those with a criminal record (Lovenheim & Owens, 2014; Pierce, Runyan, & Bangdiwala, 2014; Stewart & Uggen, 2018). And a handful of research finds that juvenile offenders have high unemployment rates, short job tenures, and low wages after confinement and incarceration (Beckett & Western, 2001; Gottfredson & Barton, 1993; Nagin & Waldfogel, 1995; Tanner, Davies, & O'Grady, 1999).

Prior research indicates that both labeling effects and behavioral effects are the underlying mechanisms behind these negative consequences. High school students with juvenile records are often stigmatized and re-classified as special education students, segregated into specialized programs, or referred to alternative schools under exclusionary polices (Aizer & Doyle Jr, 2015; Kirk & Sampson, 2013). At the college level, one study found that applications with felony convictions had a 12 percentage point higher rejection rate than matched pair applications without felony convictions at four-year colleges (Stewart & Uggen, 2018). Another study found that when the Free Application for Federal Student Aid (FAFSA) introduced a question about drug convictions in 2001, immediate college enrollments with drug convictions fell by 12 to 22 percentage points for high school graduates (Lovenheim & Owens, 2014). However, another strand of research finds stronger evidence for behavioral effects being a major barrier to progress in schooling and employment. For example, one study found that students with a felony conviction were more likely (62 percent) to leave college applications unfinished when asked about their criminal history compared with those without a conviction (21 percent) (Rosenthal, NaPier, Warth, & Weissman, 2015).

Another study finds that among first-time convicted young adults, incarceration reduced employment mainly through labor non-participation rather than through unemployment (5–12 percentage points of a total 9–15 percentage point decrease in employment was due to labor non-participation) (Apel & Sweeten, 2010).

In contrast, we know much less about how information sharing can affect education and employment outcomes in the context of juvenile offenses. Economic theory of information sharing indicates that information sharing from the justice system to schools can benefit by allowing schools to make better informed decisions and more efficiently allocate resources (e.g., school counselors, parent/guardian, parole/probation officer, and/or assigned school mentors) to those who need it the most (Bennett, 2010). There has long been a positive view of information sharing, when the federal government included building infrastructure for information sharing as one program of the Juvenile Accountability Incentive Block Grant (JAIGB) Act in 1997, a $250 million federal block grant to local governments. Information sharing started to receive more attention when the Second Chance Act included improving information sharing capabilities as one of its goals. In addition, at the K-12 level, multi-agency information sharing is viewed as an alternative strategy to a zero tolerance policy to promote campus safety, as zero tolerance policies increase referral and suspension rates (Marsh, 2014). Under the premise that information sharing will allow schools to improve their capacity to identify at-risk students at a early period, schools can provide early support and prevent at-risk students from engaging in serious criminal activities. As the literature has increasingly reflected evidence of the negative consequences of a zero tolerance policy, Teske recommended using a multi-disciplinary integrated system as an alternative to zero tolerance policy and demonstrated a 67.4% reduction in school referrals, a 20% increase in graduation rates, and a 51% reduction in felony rates (Teske, 2011; Teske, Huff, & Graves, 2013).

There is far less, if any, empirical research that provides causal estimates of information sharing in the juvenile justice context. A few early studies questioned the existence of negative labeling effect among juveniles after being arrested. For an example, one study found that boys did not experience any change in their relationship with family or teachers after being in contact with law enforcement agencies. Two recent studies report that arrest is so common that students find little to no stigmatizing effects from family, teachers, or friends (Adams, Robertson, Gray-Ray, & Ray, 2003; Hirschfield, 2008). Alternatively, a handful of re-entry literature finds some evidence of better allocating in-need support (e.g., health services) to juveniles whose information is shared between the justice system and health services (Bai, Wells, & Hillemeier,2009; Chuang & Wells, 2010). There are a few descriptive studies that identified a lack of collaboration, communication, and data sharing capabilities between schools and juvenile justice agencies as a major barrier to successful re-entry into schools (Feierman, Levick, & Mody, 2009; Leone & Weinberg, 2010; Richardson, DiPaola, & Gable, 2012; Seigle, Walsh, & Weber, 2014; Wojcik, Schmetterer, & Naar, 2008).

It is less obvious how policies can play an adequate role to reduce the barriers and promote reintegration for ex-offenders, in the aggregate. Past efforts have focused primarily on restricting access to criminal information; however, simply restricting access has had unintended social consequences. Litwok (2014)'s dissertation compared states that allowed automatic expungement with states that accept petitions for expungement and found that those who had been through the juvenile justice system and were living in states with automatic expungement had higher college attendance and graduation rates as well as higher average earnings. However, "ban-the-box" policies, which prevent job applications from asking about prior criminal involvement, caused an unintended increase in statistical discrimination against young, low-skilled black man by 3.4 percentage points (Doleac & Hansen, 2017).

In this paper, I implement a difference-in-differences design that examines the impact of the availability of juvenile records on educational attainment and employment through revisions in record laws since the Second Chance Act (2007). The period of this study is between 2008 and 2015; during this period, the public shifted away from a tough-on-crime position and focused on rehabilitating youth who were in contact with the justice system. This is referred to as the "Fourth Wave reform to juvenile justice." I focus on revisions to records legislation and code them into four categories: laws that (1) expand information shared with youth-serving systems (e.g., schools), (2) expand interagency information sharing, (3) limit public access, and (4) expand sealing and expungement eligibility.

As a preview of the results, using October Current Population Survey (CPS) data among ages 18 to 24, I find that states that expanded information sharing to schools has an average increase of 1.1 to 2.5 percentage points in their high school graduation rates and an increase of 2.0 to 3.5 percent points on the likelihood of ever attending college. In addition, I find suggestive evidence that states that expanded eligibility for sealing and expungement experienced an increase employment; however, this finding is not significant across specifications. Similarly, I find suggestive evidence for the negative effects of interagency information sharing (excluding schools) on ever attending college; however, this is not significant across specifications. In contrast, I find no robust impact on educational attainment and employment from states that enacted laws limiting public access to juvenile records.

This paper contributes to the literature in two ways. First, this paper is unique that it explores policy changes regarding the availability of juvenile records and how such changes affect education and employment outcomes in the aggregate.[4] Furthermore, this paper brings in new aspect on the benefits of some information sharing, an area that is very understudied. By examining both information sharing and protecting aspects of record policy, the findings of this study can provide clear guidance to practitioners and policy makers in identifying

[4] Much prior work has focused on identifying the negative consequences and underlying mechanisms between involvement in the justice system and educational attainment or employment rates. For example, a handful of studies convincingly demonstrated the negative consequences of having a criminal record on one's education and employment prospects. Arrest while in high school, adjudication, and confinement can increase the odds of dropping out of high school (Aizer & Doyle Jr, 2015; Hjalmarsson, 2008; Kirk & Sampson, 2013; Sweeten, 2006). Having a felony conviction increases one's probability of receiving a rejection from a college application by 12–13 percentage points (Stewart & Uggen, 2018), and 62% of individuals with prior felony conviction failed to complete the application process (Rosenthal et al., 2015). Introducing a question about drug convictions on the FAFSA in 2001 decreased college enrollment by 12–22 percentage points for high school graduates with drug convictions (Lovenheim & Owens, 2014). Other claims regarding collateral consequences included limited housing opportunities, driving privileges, and social welfare (Shah & Strout, 2016).

areas where information sharing or protecting can be beneficial or harmful.

The remainder of the paper proceeds as follows: Section 2 provides background on juvenile records legislation. Section 3 describes data and sample for the analysis and section 4 describes the difference-in-differences strategy. Section 6 presents the basic difference-in-differences results and section 7 presents the heterogeneous treatment effects. Section 8 provides sensitivity checks. Lastly, section 9 discusses policy implications and remaining questions.

# II Juvenile Record Laws

## A Background

In 2016, more than 850,000 people under age of 18 were arrested (8% of total arrests). Among juvenile arrests, 7% were arrested for offenses on the violent crime index, 21% for property crime, and 11.5% for drug abuse. The demographic characteristics for juvenile arrests (under age of 18) are predominately male (71%), late teens (78% are older than the age of 15), and white (62%) (Puzzanchera & Kang, 2017). Although national crime statistics for any given year are available, what is less clear is the population size of a cumulative arrest, which is the population affected by changes in records legislation. One recent study using the National Longitudinal Study of Youth 1997 estimates about 15.9–26.8% of youth have ever been arrested by age 18 and 25.3–41.4% by age 23 (Brame, Turner, Paternoster, & Bushway, 2012). A follow-up study finds that males are two times more likely than females and blacks are slightly more likely (30%) than whites (22%) to have ever been arrested by age 18 (Brame, Bushway, Paternoster, & Turner, 2014).

The US juvenile justice system was initially established to distinguish youth from adults and to focus on rehabilitation and child welfare (Sickmund & Puzzanchera, 2014). Consequently, the juvenile justice system was very different from the criminal justice system in handling cases and court hearings informally and keeping records confidential unless, for the purpose to help rehabilitation (Juvenile Law Center, 2014). However, during the 1980s and 1990s when serious juvenile crime rate reached a peak in 1994 (see Figure 1.1), public sentiment began to favor a stricter justice system. This led to significant changes in state legislation on juvenile crime. These changes include transferring provisions from the juvenile to the criminal justice system, expanding sentencing options, and modifying the confidentiality of juvenile records and proceedings. Between 1992 and 1997, 45 states made changes in transfer provisions, 31 states in sentencing authority, and 47 states in confidentiality laws (Sickmund & Puzzanchera, 2014). Related to record laws, the changes allowed more exceptions to record confidentiality (sealing, expungement, or deletion), expansion in the ability to collect personal information such as DNA or fingerprints, and more information sharing between justice agencies and youth-serving systems (Sickmund & Puzzanchera, 2014; Torbet, Gable, Montgomery, & Hurst, 1996; Willison, Mears, Shollenberger, Owens, & Butts, 2010). Ever since, the confidentiality level of juvenile records has varied across states. For an example, in 2009, 31 states had exceptions for sealing or expungement in subsequent offenses, 41 states had no age restrictions for taking fingerprints of an offender, and 46 states had school notification laws (Sickmund & Puzzanchera, 2014).

In contrast to the 1990s, the past fifteen years have been the lowest era for juvenile crimes. Juvenile arrest and confinement rates have continue to drop reaching the lowest since the 1980s. In 2016, about 850,000 youths under age 18 were arrested, which is about half of what it was in 1997 (OJJDP Statistical Briefing Book, 2018). Similarly, confinement rates in juvenile detention and correctional facilities fell by about half in the corresponding years with less than 50,000 youth arrests in 2015 (OJJDP Statistical Briefing Book, 2018).[5] The decline in juvenile arrests outweighed that of adults in nearly all offenses (Sickmund & Puzzanchera, 2014). As the juvenile crime rate has trended down and with Congress passing

[5] In 2010, the violent crime arrest rate reached the lowest level since the 1980s, with a decline to less than 250 arrests of juveniles under age of 18 per 100,0000 juveniles between ages of 10 and 17 (this is 55% fewer arrests than during the peak at 1994). Juvenile arrests are about 10% of all arrests and 20% of all property crime index offenses, liquor law violations, and disorderly conduct.

the Second Chance Act, most states have once again shifted their focus to rehabilitation in the juvenile justice system. Between 2008 and 2015, 39 state legislatures enacted statutes that revised legislation related to juvenile records, which is the main focus of this paper.

## B Mechanisms of Availability of Juvenile Record to Education and Labor Market

It is quite common for schools and employers to have access to juvenile criminal records in a slightly different way. At high schools, the most common mechanism is through direct conversation with either local law enforcement agencies or from the court under school notification law. Additionally, it is quite frequent to have offense take place in schools.

School notification laws are quite common across the United States. In 2008, 46 states had some form of school notification laws (Sickmund & Puzzanchera, 2014), though there is less guidance on protection and access policies on these information (Shah et al., 2014). In contrast, the most common way for colleges to find out about a student's criminal history is through a question on the college application that asks applicants to identify their previous criminal history (Shah & Strout, 2016). One national survey found that 60–80% of private colleges and 55% of public colleges include a question about the applicant's criminal history in their undergraduate applications (The Center for Community Change & National Employment Law Project, 2014; Pierce et al., 2014; Stewart, 2015). For employers, the most common mechanism is to ask consumer reporting agencies. Consumer Reporting Agencies (CRAs) have a large dataset gathered from internet searches, court storage centers, or purchased from subcontractors who gather arrest and conviction data from various resources (e.g., state systems or issue reports).[6]

Prior studies have identified two theoretical frameworks for understanding the negative consequences of a criminal record on educational attainment or employment: the stigma

[6] There are some accuracy concerns about these data and how frequently they are updated regarding sealing and expungement (The Center for Community Change & National Employment Law Project, 2014; Shah & Strout, 2016).

hypothesis and the behavioral hypothesis. The stigma hypothesis follows the labeling theory where institutions, teachers, peers, college administrators, or employers change their way of thinking about a student/applicant when they learn about the person's prior criminal involvement (Aizer & Doyle Jr, 2015; Kirk & Sampson, 2013; Lemert, 1951).. The behavioral hypothesis suggests that their experiences in the justice system can have a negative impact on youth's behavior or skills such as, lowering self-esteem, creating detachment to schools, or have less accumulation of human capital (Lambie & Randell, 2013; Nye, Short Jr, & Olson, 1958; Stewart & Uggen, 2018).

Alternatively, a far less studied mechanism is where information sharing can benefit in allocating the resources to the population who needs it most (Bennett, 2010). That is, information shared about a youth can help society efficiently allocate scarce resources (e.g., school counselors, parent advocate, parole/probation officer, and/or assigned school mentors) to those students who need it the most.

This paper explores state changes in juvenile record policy that alters these underlying mechanisms.

## C Four Categories of Juvenile Record Laws

A juvenile record can refer one of two types of records: a law enforcement record or a court record. The law enforcement record is created at the time of arrest and includes detailed information about the arrest, police reports, detention and charging documents, witness and victim statements, and if applicable, photographs, fingerprints, and DNA samples. These records are primarily stored in the law enforcement agency's record system (Shah & Strout, 2016). Second, the court record is created after law enforcement agencies decide to send an arrest to the juvenile system.[7] These court records include information such as family background, behavioral and health history, education, and prior interactions with law enforcement.

[7] In 2010, 68% of arrests were referred to juvenile court (Sickmund & Puzzanchera, 2014).

Between 2008 and 2015, 39 state legislatures enacted statutes revising laws related to juvenile records. I divide these laws into four categories:[8]

- Type 1: Expand information sharing to youth-serving systems - 11 states
- Type 2: Expand interagency information sharing (excluding youth-serving systems) - 16 states
- Type 3: Restrict public access to juvenile records - 9 states
- Type 4: Expand eligibility and modify procedures for sealing and expungement[9] - 23 states

Figures 1.2–1.5 show the years that states enacted a revision law, separately by category. The control states are those that did not make any changes to their juvenile record record laws during the time period of this study, 2008 to 2015. There are 12 control states in this study: Arizona, Arkansas, Georgia, Idaho, Massachusetts, Michigan, Missouri, Montana, New Jersey, New Mexico, Rhode Island, and Vermont. These four types of record laws are indicated separately into my difference-in-differences estimate to estimate individual impacts on education and employment outcomes.

[8] This categorization is adapted and modified from (Willison et al., 2010). Willison et al. (2010) makes four categorizations: information sharing with youth serving systems, interagency information sharing, limiting public access, and data collection on personal information (e.g. DNA, fingerprints, photographs, etc). As mentioned before, I exclude the fourth category and consider laws related to sealing and expungement instead.

[9] Juvenile records are generally confidential. The term *confidentiality* means protected during or right after juvenile proceedings. States have varying specifics of which records is confidential and not. For an example, these vary in who has access, what information is accessible, and what type of offense. Also, most states have exceptions to confidentiality by offense type, offender age, and repeatability of crime. Law enforcement and court personnel are generally exceptions from confidentiality. In formal definition, confidentiality refers to "preventing access to, dissemination or use of juvenile record outside of juvenile court unless it is intended for youth's case planning (Shah et al., 2014)." Confidentiality is different from sealing in that sealing refers to protecting records after the case is closed. Although some states uses the term sealing and expungement interchangeably, the two terms differ by whether it is access closure or physical destruction of the record. In formal definition, sealing refers to "a record that is unavailable to the public, but accessible to select individuals or agencies" and expungement refers to "actual physical destruction and erasure of juvenile record (Shah et al., 2014)." In this paper, I use sealing and expungement terminology interchangeably. Similar to confidentiality, sealing and expungement laws can vary in many dimensions across states. States can choose which record is eligible (law enforcement and/or court), which procedure is needed (automatic or petitioning), and how states can notify eligibility to recipients for sealing or expungement.

# III Data and Sample

The main dataset used in this study is the Current Population Survey (CPS) for October. CPS data are collected monthly with a particular focus on labor force statistics. I selected the October CPS in order to align with the academic year (starting in the fall and ending in the spring), as two of the outcomes of interest are education outcomes. In addition, the October CPS has a supplemental survey that allows me to separate out those who received GEDs from those who received high school diplomas. However, using CPS data for my main analysis has several limitations. One limitation is the sample coverage; CPS only collects data from non-institutionalized individuals. Since I am interested in youth in contact with the juvenile justice system, it would have been helpful to have the same information for institutionalized individuals as well. This raises the question of how much of the sample is missing as a result of this restriction. Fortunately, unlike the adult criminal system, the juvenile justice system typically has relatively short stay in juvenile residential placement facilities. The median stay among committed juvenile offenders ranges between 103 to 113 days. Less than 11% of committed and 2% of detained offenders remain in residential placement more than a year since admission (OJJDP Statistics Briefing Book).[10] The short length of stay results in only a small fraction of the sample missing from the data for any given year. Between 2003 and 2015, only 0.152 – 0.303 percent of the juvenile population aged 10 to the upper age of juvenile court jurisdiction were in residential placement (Sickmund, et al., 2017). The second limitation of using CPS is that it does not have a questionnaire asking about cumulative arrest. Exclusion of prior criminal involvement limits my analysis to only able to estimate a general effect of changes to record laws on the entire non-institutionalized population. In other words, I am not able to disentangle the direct effect on ex-offenders with a juvenile record from spill-over effects on non-offenders. In addition, without knowledge on how

[10] Data is retrieved from OJJDP Statistical Briefing Book. Online. `https://www.ojjdp.gov/ojstatbb/corrections/qa08401.asp?qaDate=2015`. Released on June 01, 2017.

large the affected population is by these law changes can raise concern on sample coverage. Fortunately, since a juvenile record is created as early as at the time of arrest and the policy change can affect anyone even with one arrest, coverage from the non-institutionalized sample should provide reasonable sample coverage. Using National Longitudinal Survey of Youth 97, Brame, et al. (2012) calculates approximately 15.9 - 26.8% of the national representative sample had more than one cumulative arrest by age of 18.

In order to gather record laws that were enacted during 2008–2015, I use the National Conference of State Legislatures (NCSL) juvenile justice bills tracking database. This database gathers all introduced or enacted juvenile justice statutes starting from 2008 to the current date. From this database, I extracted only enacted legislation between 2008 to 2015 that fell under the category of "Records and Information."

Additional state-specific time varying variables are gathered from different sources. The Bureau of Labor Statistics provides monthly, seasonally-adjusted unemployment rates for each state. I take the average unemployment rate from the previous October to September to align with CPS October calendar years. In addition, annual state specific juvenile arrest rates including the violent crime index, the property crime index, and drug abuse is collected from the Office of Juvenile Justice and Delinquency Prevention Statistical Briefing Book, which are calculated using the FBI's Uniform Crime Reporting (UCR) data.[11] Lastly, log juvenile population by state and year is gathered from the Easy Access to Juvenile Population database from the Office of Juvenile Justice and Delinquency Prevention.

I restrict the sample to ages 18–24, which is the population most likely to be affected by these legislative changes. Table 1.1 provides the descriptive characteristics of the sample using CPS data for October during pre-treatment periods between 2004 and 2007. Columns 2 through 12 group individuals who lives in treated states by type of legislation. For each category, three columns show sample mean, standard deviation, and p-values from a t-

[11] The annual juvenile arrest rate is calculated using the number of arrests for a particular offense among every 100,000 persons of ages 10–17.

test comparing average differences between control group (columns 13–14). The legislation categories are listed left to right from type 1 to type 4, as listed above. The two rightmost columns present the mean and standard deviation for individuals in the control states.

In terms of outcomes, Table 1.1 shows that there is no significant difference in outcome averages between the treated states and the control group. In contrast, the next rows presenting individual demographic data demonstrate a few noticeable demographic differences between treated states and control states. In particular, treated states are generally less white (except for states that limit public access to juvenile records), less black, have larger populations of other races, and are more metropolitan (except Type 3 states). Also, treated states generally have higher juvenile crime rates, lower unemployment rates, and a high juvenile population. Particularly, states that reformed information sharing with schools stands out with higher juvenile arrest rates than the control group.[12]The bottom panel of Table 1.1 shows that treated states and control states are spread out across regions throughout the country.

# IV Empirical Methodology

The main design of the study is a difference-in-differences (DD) strategy, where the model compares trends in outcomes with control states before and after juvenile record reforms. The simplest form of DD specification is:

$$y_{ist} = \beta_1 TreatPost_{st-1} + \gamma_s + \delta_t + X_{ist} + X_{st} + \varepsilon_{ist} \tag{1.1}$$

where, $y_{ist}$ is an outcome variable for person $i$, in state $s$, during year $t$. Outcome variables are whether or not one has received a high school diploma (excluding those with GEDs),

[12] Similarly, I find higher rate of juveniles placed in residential facility for states that revised laws on information sharing with schools. On average, 331 out of 100,000 juvenile population were in residential facilities between years 2003 and 2007 (Sickmund, et al., 2017). On the contrary, 226 out of 100,000 juvenile population were in residential facilities for control states during the same period.

ever attended college, and employment status. $TreatPost_{st-1}$ is an indicator with value 1 for treated states in post-treatment years. I have a lag of one year to capture the gap between enactment and implementation of the laws.[13] The coefficient of interest is $\beta_1$, an estimate of pre-post change in outcomes among individuals in states that revised a law type compared with those in control states. $\gamma_s$ is a state fixed effect and $\delta_t$ is a year fixed effect. $X_{ist}$ are individual controls and include indicators for living in a metropolitan area, gender, age, race, and marital status. $X_{st}$ are state-specific time-varying controls such as the unemployment rate, annual juvenile arrest rates (violent crime index, property crime index, drug abuse violations), and the log of the juvenile population (age $< 18$).

I enrich this basic model by first adding a state linear time trend to allow for differential cross-state trends and second by adding an interaction of regional dummies and year dummies to control for regional common shocks. I cluster standard error by state to address serial correlation (Bertrand, Duflo, & Mullainathan, 2004). Lastly, in order to allow for at least one year of pre-treatment period for laws enacted in 2009, I examine outcomes from 2008 to 2016.

## A Defining Treatment Dummies

Using the NCSL Juvenile Justice Bills Tracking database, a total of 165 records and information laws were enacted between 2008 and 2015. Eleven laws (from six states) related to allowing the collection of personal data (e.g., fingerprints or photographs) were excluded from the analysis. The remaining 154 laws were assigned year $t$ if the enacted date lies between October of year $t-1$ and September of year $t$. For each law type $j$, $treatedstates_j$ is defined as those states that ever enacted a type $j$ law during the study period. *controlstate* refers to states that did not experience any amendment activity during the study time period. This study uses a total of 12 states as control states: Arizona, Arkansas, Georgia, Idaho,

[13] the enacted law year $t$ was calculated using both the year and month of the enactment date. That is, to align with the data cycle of October, I assign laws that were enacted between the previous October to this year September as belonging to the current year.

Massachusetts, Michigan, Missouri, Montana, New Jersey, New Mexico, Rhode Island, and Vermont. The $TreatPost$ variable is coded as 0 prior to enactment and 1 after enactment only among treated states.[14] Within each law type, the majority of laws were enacted in a monotone direction.[15] However, if an enacted law is in the opposite direction to existing law, then that state will have a missing value for $TreatPost$ for all subsequent years.[16]

Figures 1.6–1.9 graph $TreatPost$ values by year (one-year lag) along the x-axis, separately by legislation type. Missing values indicate the year a contrasting law was enacted. For type 1 laws (Figure 1.6), there are no missing values, indicating that all enacted laws had a monotone direction. For type 2 laws (Figure 1.7), Maryland has missing values starting from 2014, meaning that a law was enacted in the opposite direction in 2013. Overall, Figure 1.8 confirms that the majority of laws were enacted in a monotone direction. In addition, Figure 1.9 shows that majority of revisions to records legislation occurred prior to 2013.

# V Results

## A Basic Difference-in-Differences Results

Table 1.2 provides results from the difference-in-differences regressions. Columns 1 to 5 contain the estimated impact of each law type on the likelihood of earning a high school diploma (excluding GED earners), columns 6 to 10 on the likelihood of ever attending college, and columns 11 to 15 on employment. For each outcome, the first column presents a simple regression coefficient with state and fixed effects, the second column adds individual controls, and the third column adds state-specific controls and are estimated using equation (1.1). The

[14] Because I add a one year lag, the $TreatPost_{st-1}$ indicate as 1 one year after the enactment year.

[15] Here, direction is defined as protection versus sharing. For example, for type 1 laws (information sharing with schools), laws that limit sharing with schools are regarded as being in the opposite direction of laws that expand sharing with schools.

[16] In cases where a state had multiple amendments within the same year $t$ that were all in the same direction, I treat the amendments as one. However, if multiple amendments within the same year had contrasting directions, I exclude that year from treatment and replace the missing value for the $TreatPost$ variable.

fourth column adds linear state trend and the fifth column adds region × year dummies in addition to equation (1.1).

States that revised laws on information sharing with schools have, on average, a 1.7 to 2.5 percentage point higher high-school graduation rate (this is about a 2.0 to 3.1 percent increase from the baseline of 80.5%) than the control states. Similarly, these states also have 2.0 to 3.5 percentage point higher rates on ever attending college (about a 4 to 7 percent increase from a baseline of 50%) than control states. The coefficients are significant and robust across all specifications. If approximately 16–27% of the general population had at least one cumulative arrest (Brame, et al., 2012) and if the information sharing with schools only effected directly on the population with a juvenile record, I would expect the treatment effect size to be larger (about four to six times larger) in magnitude to those with a juvenile record. On the other hand, Teske (2011) finds some evidence of positive spill-over effects from multi-integrated systems through reduction in the number of students who are detained on school offenses by 86%, reduction on school offense referral to the court by 43%, and reduction in the number of serious weapons on campus by 73%. This suggests that it is possible that information sharing with schools may have non-direct positive effects on non-offenders through the mechanism of increasing safety of schools.

I find a suggestive increase in employment by 0.9 to 2 percentage points (about 1.4 to 3.2 percent increase from a baseline of 80.5%) for states that expanded sealing and expungement eligibility compared with no-change states. The coefficients are positive across all specifications; however, two specification are not significant. Including state-specific or regional trends reduced the magnitude of the coefficients on employment, suggesting that treated states were already experiencing an upward trend in employment rates. In contrast, I do not find any robust effects for states that revised laws regarding interagency information sharing or public access to records.

Including linear state trends or region interacted year dummies generally doesn't change the direction of coefficients. However, the coefficients on ever attended college becomes slightly larger and less stable after including both specification. To take the conservative estimate, I choose the simpler model, as specified in Equation (1.1), with state and year fixed effects, individual controls, and time-varying state controls as preferred model for my sub sequent analyses.

## B Effects of all four laws simultaneously

Given that a mixture of law types were enacted during the study period, it may be worth exploring the joint effect of the four laws. In order to do so, I fit the following equation:

$$y_{ist} = \sum_{j \in lawtypes} \beta_j (Treat \times Post_{Lawtype_j})_{st-1} + \gamma_s + \delta_t + X_{ist} + X_{st} + \varepsilon_{ist} \tag{1.2}$$

Furthermore, it may be worth exploring whether it is necessary to separate the data according to type of legislation. To examine this issue, I create an indicator $anylaw_{st-1}$ that has a value of 1 for periods subsequent to when any of the four type laws were enacted and 0 otherwise.[17] This $anylaw_{st-1}$ is replaced from $TreatPost_{st}$ indicator from equation (1.1) above.

Table 1.3 presents regression coefficients from equation (1.2). The first four columns estimate the joint impact of the four laws or any law indicator on the likelihood of receiving a high school diploma (excluding GED), columns 5 to 8 on the likelihood of ever attending college, and columns 9 to 12 on employment. The first column of each outcome is the regression coefficient of equation (1.2), the second column adds a linear state trend, and the third column adds region × year dummies. The fourth column indicates the regression coefficient when replacing equation (1.1) with any law indicator, which equals 1 after any juvenile record law revisions were made in a given state and year and 0 otherwise.

The significant and robust positive coefficients of information sharing with schools on high school graduation rates and ever attending college is consistent with the findings in the previous section. The estimation magnitudes in columns 1 to 3, between 1.1 to 1.4 percentage

[17] If any laws were enacted in a conflicting direction within a same year and state, I replace that state and year as missing.

points, are slightly smaller than the separate estimations shown in Table 1.2, which indicates that separating out one type of legislation partially captures effects from other the types of legislative changes. However, no coefficients from other laws are significantly affecting high school diploma attainment rates. On the other hand, the estimation magnitudes for columns 5 to 7, between 2.2 to 3.3 percentage points, are very similar to estimations that do not include other record laws. This suggests that the educational benefits of information sharing with schools remains positive regardless of the other types of laws that were enacted in the same year.

I observe three patterns that deserve a mention. First, I find a suggestive negative impact by 0.7 to 2.2 percentage points from states that enacted legislation regarding interagency information sharing on ever attending college, which was not observable in previous separate estimations. The significance level is not achieved across all three model specifications (columns 5–7); however, the coefficients are consistently in a negative direction. It appears that those states that made revisions to laws regarding interagency information sharing may be correlated with another law with positive impact, which negated the negative effect on ever attending college from Table 1.2. This finding suggests that subsequent analyses should examining the joint effect of any enacted legislation. Second, the suggestive positive coefficient of expanded sealing and expungement eligibility laws on employment still holds and still, is marginally significant in only one specification after including other laws. Lastly, columns 4, 8, and 12 indeed suggest that separating out the individual effects of each types of juvenile record laws can provide better interpretations. Unlike the few significant estimations from separating into four types of laws, none of the coefficients on any law indicator is significant.

The results in Tables 1.2 and 1.3 suggest that information sharing with schools have robust and significant positive benefits on both attaining a high school diploma and ever attending college. The magnitude of the impact on earning a high school diploma is between 1.1 to 2.4 percentage points (a 1.3 to 2.9 percent increase from a baseline of 80.5%) and the magnitude impact on ever attending college is between 2.0 to 3.5 percentage points (a 4 to 7 percent increase from a baseline of 50%). In addition, I find a suggestive positive impact from legislation that expanded sealing and expungement eligibility upon employment; however, the findings are not robust across specifications. Using joint effect estimation, the coefficients on interagency information sharing suggest a negative impact on ever attending college; however, this is not significant across specifications. In order for difference-in-differences estimations to be valid, a critical assumption is the common trends assumption. The following section provides assumption and sensitivity checks on these basic results.

# VI Sensitivity Checks

## A Pre-trends

Difference-in-differences estimates rely on the common trends assumption. The common trends assumption requires that treated states and control states have comparable trends without treatment. One way to check for parallel trends is to visualize outcome changes by year in the pre-treatment period. Figures 1.10–1.12 show residualized outcome changes between years 2004 to 2016.[18] The residualized outcome is estimated using a regression of state and year fixed effects, individual controls, and state-specific controls on outcomes and plotting the residuals by each year. Year 2008 is excluded as a baseline. The red line divides pre-treatment years and post-treatment years. Because my treatment is multi-year law changes starting from 2008 to 2016, I am only able to plot the first year of the treatment period. In order for the common trends assumption to hold, treated and control states should have roughly similar trends prior to the first year of treatment, between 2004 and 2008. Note that the common trends assumption is valid even the lines are at different levels, as long as the pre-treatment slope changes are roughly the same. Reassuringly, Figure 1.10 indicates

[18] In Appendix Figures A1.1 – A1.3, I directly plot outcome changes (un-residualized) by year. Generally, I find similar trends during pre-treatment years.

that the common treatment assumption holds; treated and control states appear to have roughly similar pre-treatment trends. In addition, for significant regression estimators, one may be able to detect trend differences between treated states versus control states during the post-treatment period. Figures 1.10 and 1.11, which indicate the post-treatment trends for high school graduation rates and ever attending college, show visible differences in trends for control states from trends from states that made revisions on information sharing with schools.

## B Lags and Leads

An alternative way to check for the common trends assumption is to pool all years of data and augment the data with beads of treatment indicator. If coefficients on leads have a non-zero coefficient, the previous main results may be obscured from a reverse causality interpretation. In addition, adding lags of treatment indicator would provide the dynamics of treatment effects. To explore common trends assumption and the dynamics of treatment effect, I regress my outcomes with three years of leads and two year of lags as below:

$$Y_{ist} = \sum_{\tau=-m}^{q} \beta_j (D_{Lawtype_j})_{st-1+\tau} + \gamma_s + \delta_t + X_{ist} + X_{st} + \varepsilon_{ist} \tag{1.3}$$

where, $D_{Lawtype_t}$ refers to indicators for 1–3 years before the enacted legislation, 1 year after the enacted legislation, and 2 years or more years since the enacted legislation. I include leads and lags for all four law types simultaneously, because there was correlation among states that implement more than one type of record law as seen from the previous result. Because I have states that enacted the same type of record laws multiple times between years 2008 to 2016, I use all years between the earliest and latest legislation as the years of enactment. For example, if state $s$ enacted legislation expanding information sharing with schools in both 2009 and 2012, the indicator for $\tau = -1$ will have be 1 in 2008 for state $s$, 1 for $\tau = 0$ in all years between 2009 and 2012, and one for $\tau = 1$ in 2013.

If the common trends assumption holds, the coefficients on the leads should be no different from zero. The beta estimations on the lags will identify the dynamic nature of the treatment spell. Figures 1.13 and 1.14 present coefficient plots for $\beta_j$ for information sharing with schools on high school graduation and ever attending college, respectively. Figure 1.13 confirms that there is no evidence for reverse causality and that the positive impact appears immediate after one year of enactment. However, from Figure 1.14, I find negative coefficients for indicators 3 years prior to enacting legislation expanding information sharing with schools on ever attending college. That said, it is quite hard to make reverse causality interpretation; states made revisions to information sharing with schools laws as a result of low rates of ever attending college. Figure 1.14 may indicate instead that my common trends assumption only holds within a relatively short pre-treatment period.

## C Aggregation Results

So far, my regression estimates use individual level data. An alternative model specification is to aggregate individuals up to the state level and explore aggregated estimations on the effect of juvenile record law revisions. The coefficients on state-level aggregate difference-in-differences regressions should have similar point estimates with but different standard errors from the main results. In order to align with the models using individual data, outcome variables are residualized on individual covariates prior to aggregation in order to control for individual differences. The state-level aggregated regression is modeled as follows:

$$\widetilde{y_{st}} = \beta_1 TreatPost_{st-1} + \gamma_s + \delta_t + X_{st} + \varepsilon_{st} \tag{1.4}$$

where $\widetilde{y_{st}}$ is a residualized outcome from regressing individual controls. Table 1.4 presents the regression results of equation (1.4). Reassuringly, the magnitude of the coefficients across all specifications resembles that of the main results from Table 1.2. This confirms that my aggregated model is equivalent to individual models. Even with state-level aggregated model, general finding still holds: information sharing with schools on earning a high school diploma and on ever attending college is positive and significant. In addition, I still find significant positive effects (three out of four specifications) from legislation that expanded sealing and expungement eligibility upon employment. Thus the overall findings hold even with using state-level aggregated data.

## D Heterogeneous Treatment Effects

The findings in the previous sections consistently demonstrate the positive impact of expanding information sharing with schools on educational outcomes. However, the underlying mechanism of how information sharing with schools benefits educational attainment is not as clear. Prior research has suggested that in sharing information on students' involvement with the juvenile justice system, schools are able to identify at-risk students and allocate resources to taking preventative measures. This creates a more stable educational environment. To explore this hypothesis, I run a subgroup analysis on various demographic and regional characteristics using equation (1.2). As changes in juvenile record legislation only directly impact those who have ever been arrested, I would expect these legislative changes to have the largest effect on the subgroup that most closely resembles juvenile offender characteristics.

Table 1.5 presents regression coefficients on various subgroups. The coefficients only show impact estimations from the states that made revisions on information sharing with schools, although the actual regression includes all four types of record legislation simultaneously. Columns 1–3 show difference-in-differences estimations on earning a high school diploma (excluding GED earners), columns 4–6 on ever attending college, and columns 7–9 on employment, respectively. Column 1 reflects regression results from equation (1.2), column 2 adds linear state trends, and column 3 adds region and year dummies from equation (1.2). The table separates the estimates by gender (males or females only), by race (black, white, or other race), by age (traditional college-age, between 18 to 21, and older, between 22 to

24), and by state indexes for drug abuse, violent crime, and property crime.

Across all subgroups, information sharing with schools has positive effects on high school attainment and ever attending college. Given the larger magnitude on ever attending college, I find more coefficients reaching significance levels in columns 4–6. The male population group and black population group have the largest significant effect sizes on both education outcomes, which resembles the demographic characteristics of juvenile offenders. In particular, when looking at the male-only population, I find a 0.5 to 2.1 percentage point increase on high school attainment and a 2.6 to 3.5 percentage point increase on ever attending college, which is slightly larger in magnitude than the general findings from Table 1.2. When I restrict the estimate to the black population, the magnitude becomes even larger, between a 1.9–4.8 percentage point increase on high school attainment and 2.0–5.8 percentage point increase on ever attending college. The positive impacts of sharing information with schools on education attainment and particularly on ever attending college are spread out across gender and race. This suggest that there might be positive spill over effects from increasing campus safety.

In terms of age, I find increased information sharing to have larger effects on the older student population. One possible explanation is that students who have been in contact with the justice system are less likely to graduate high school on time. It is possible that a subgroup of young potential college students may still be in high school and that the benefits might be as readily apparent for the younger population. Alternatively, the benefits may take some time to reflect as an increase in outcome. For an example, additional resources from information sharing with schools (e.g., counselors or mentors) may have behavioral benefits (e.g., less detachment from school or less self-discouragement), which may appear in the later life span.

The last three panels show regression estimates after restricting to states that have high rates of juvenile crime and drug abuse, and high violent crime and property crime indexes. Table 1.5 shows that the benefits of sharing juvenile records with schools are significant in states with high violent crime and property crime indexes. The heterogeneous treatment effects by offense type reveals that records legislation may have a differential impact by offense type.

# VII Discussion

This paper looks at the impact of revisions to juvenile records legislation on education and employment outcomes after the Second Chance Act of 2007. I identify four types of legislation that address juvenile records: (1) expanding information sharing with youth-serving systems (e.g., schools), (2) expanding interagency information sharing, (3) restricting public access, and (4) expanding sealing and expungement eligibility. My difference-in-differences estimates find increases in high school attainment and probability of ever attending college for states that revised their statues to expand information sharing with schools. In particular, I find a 1.1 to 2.4 percentage points increase (1.3 to 2.9 percent increase from a baseline of 80.5%) in high school diploma attainment and a 2.0 to 3.5 percentage points increase (4 to 7 percent increase from baseline of 50%) on the rate of ever attending college. My heterogeneous estimates indicate that treatment effects are largest among the Black and male sub populations, which aligns with the demographic characteristics of juvenile offenders.

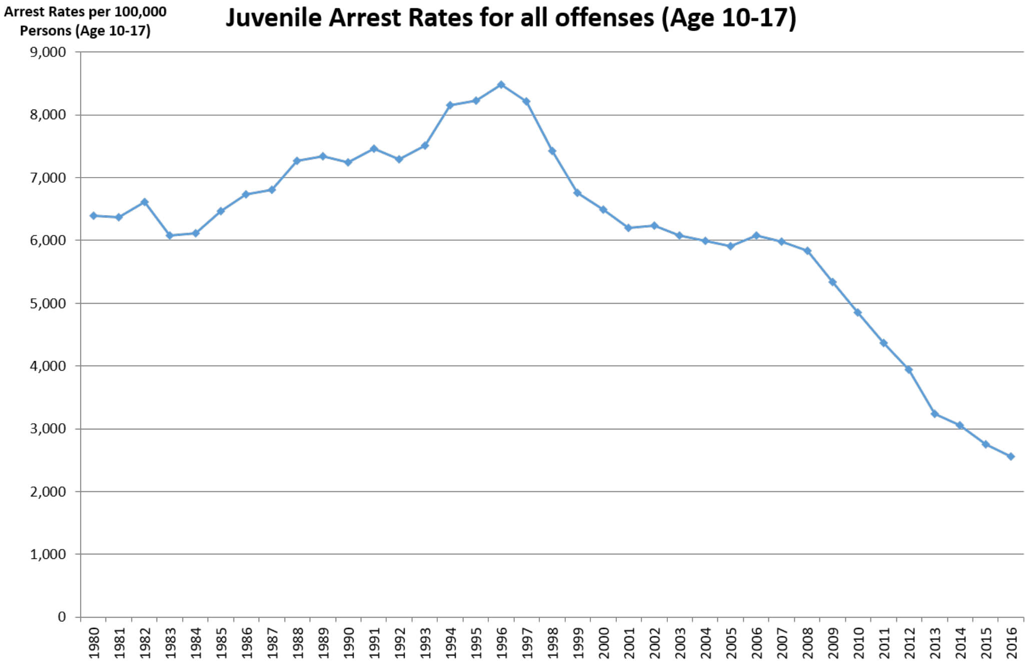
While the effects of juvenile records legislation have not been empirically studied, prior research found similar trends. One case study that implemented a multi-integrated protocol in Clayton County Juvenile experienced a 20% increase in graduation rates (from a baseline of 60%) after implementation. However, this study did not control for any secular trends or confounding factors, which makes it less compelling to compare with my causal estimations. From Litwok (2014)'s paper, the model that is comparable with my analysis is the interaction coefficient of juvenile arrest from living in automatic expungement states compared to non-automatic(application) states, which finds a non-significant 0.3 percentage point decrease on ever attended college and a non-significant 2.9 percentage point increase in log of average income. Although my suggestive positive outcome is on employment and not average income, Litwok (2014)'s finding of positive labor market outcomes aligns well with my study.

The findings of this paper have several policy implications. First, information sharing between schools and the justice system can help improve educational attainment rates in the aggregate. Information sharing theory sheds some guidance on the possible mechanisms behind this: improved detection of at-risk students and allocation of resources to take preventative steps before at-risk students engage in more serious activities. This is consistent with the findings in re-entry literature which shows that collaboration is key to successful re-entry (Leone & Weinberg, 2010; Richardson et al., 2012). Second, I find suggestive improvements from legislation that expanded sealing and expungement eligibility on employment. Lastly, heterogeneous treatment estimations suggest possible implications for spillover effects on the non-offender population. In addition, the differential effect sizes by regions with higher offense type suggests that juvenile record laws can have differential impacts by offense type.

As a result of this paper, a few critical areas are identified for further research. First, it is necessary to explore the underlying mechanism between information sharing with schools and the juvenile justice system. How do schools re-distribute resources for a student who is identified as at-risk? Does availability of schools resource matter? One possible scenario is through changes in expenditures for student services or increasing the number of guidance counselors. To examine this possibility, I used Common Core of Data (CCD) and estimated my difference-in-difference regression. However, I did not find any changes in the number of guidance counselors or total expenditures on student services among school districts that have secondary level schools and are in a state that revised information sharing with schools (author can provide these results upon request). This null finding is not surprising given that re-distribution of resources is not equivalent to additional resources that these regressions capture. Unfortunately, CCD does not contain information about distributional practices of existing resources. In addition, learning the connection between high school intervention as a result of information sharing to individual choices on college attendance can be critical to understand further in my findings of improvement on ever attending college. Second, differentiating out the effect of information sharing on ex-offender and non-offenders seems important. My analytic sample relies on the CPS dataset, which includes both ex-offenders along with youths and non-offenders. My heterogeneous estimations shed light on the possibility of spillover effects of information sharing with schools for non-offenders. Unfortunately, there is no direct way for me to identify whether an individual in CPS has a record or not. Also, it is worth mentioning that CPS includes only non-institutionalized individuals, which may result in underestimating the positive effects by excluding an institutionalized sample who may be influenced by this legislative change. Third, it would be interesting to see if any of the positive effects from information sharing with schools influence recidivism rates. If redistribution of resources deterred at-risk students from engaging in further criminal activities, we may be able to detect a decreased recidivism rate. Lastly, my differential impact estimates by offense type raise the possibility that juvenile record laws can have differential impact by offense type. It may be worth exploring the possibility of differential impacts by offense type, which can have critical policy recommendations.

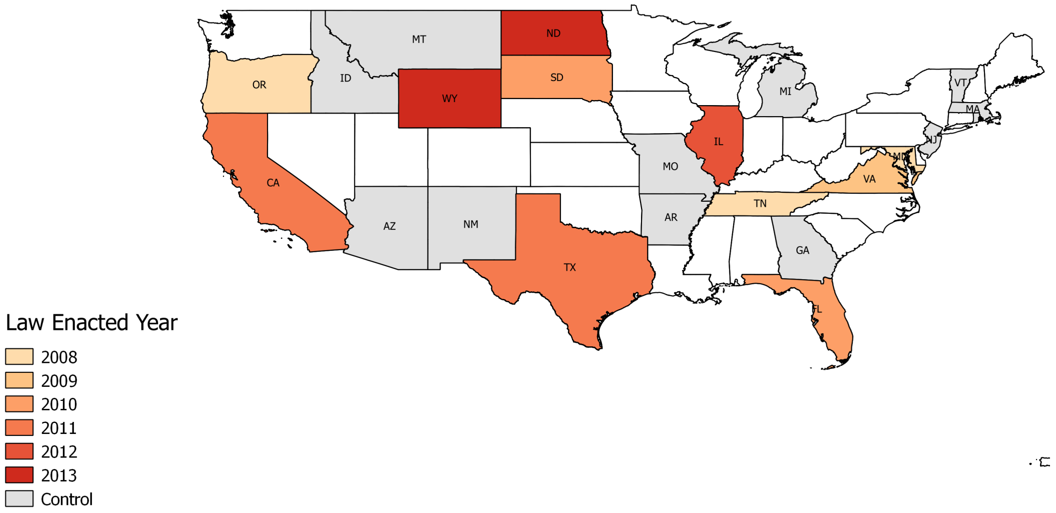
# VIII Figures

Figure 1.1: Historical juvenile arrest rates per 100,000 persons aged 10–17 (1980–2016)



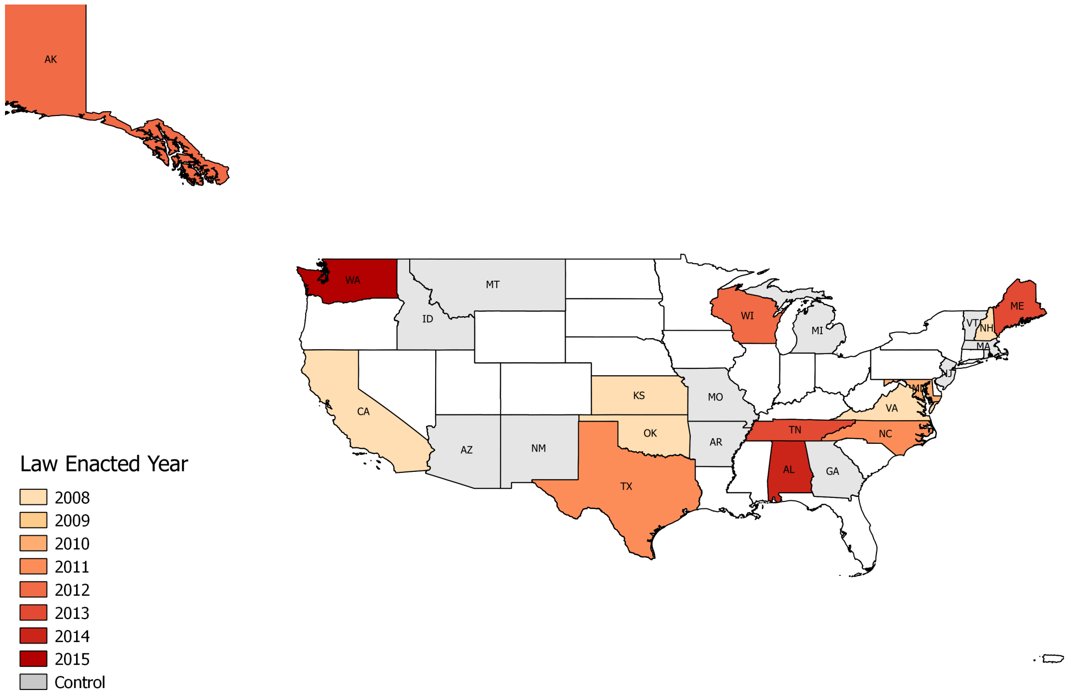
This figure show historical juvenile arrest rates for aged 10–17, between 1980 and 2016. The y-axis show arrested person per 100,000 juvenile population. Data source: Arrest estimates developed by the Bureau of Justice Statistics and disseminated through ”Arrest Data Analysis Tool.” Online. Available from the BJS website `https://www.ojjdp.gov/ojstatbb/crime/ucr_trend.asp?table_in=1`.

Figure 1.2: Enacted revisions to juvenile record laws: Expand information sharing to youth-serving systems, 2008–2015 (11 states)



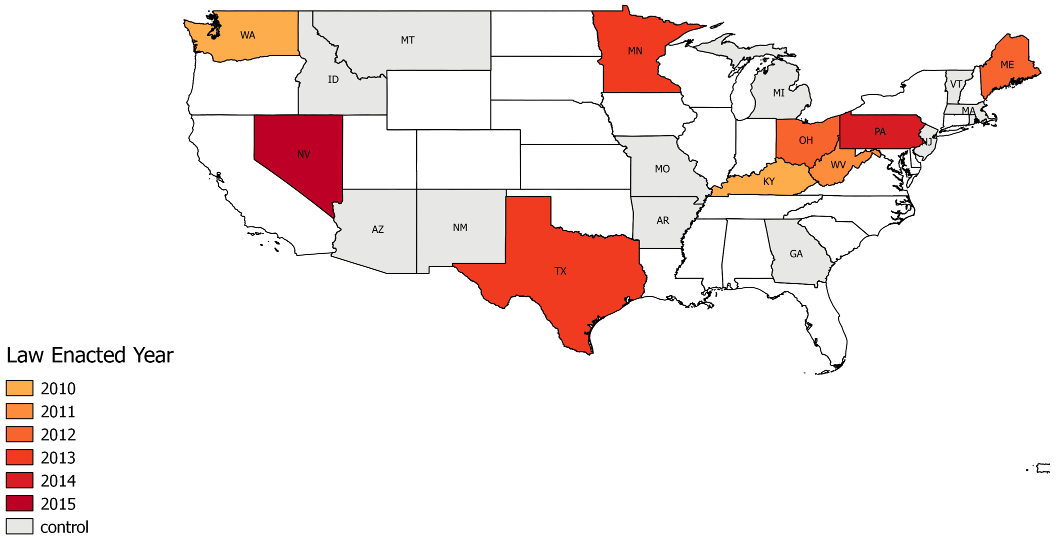
This figure present geographical representation of states that enacted revisions to juvenile record laws between 2008 and 2015. From yellow to dark red colored states indicate recent years of enactment. Gray colored states indicate control states.

Figure 1.3: Enacted revisions to juvenile record laws: Expand interagency information sharing, 2008–2015 (16 states)



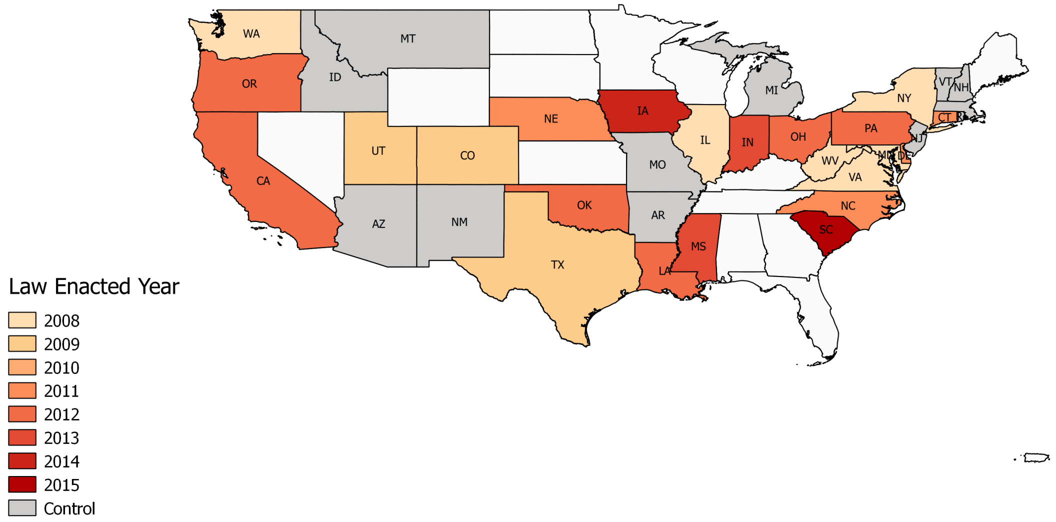
This figure presents geographical representation of states that enacted revisions to juvenile record laws between 2008 and 2015. From yellow to dark red colored states indicate recent years of enactment. Gray colored states indicate control states.

Figure 1.4: Enacted revisions to juvenile record laws: Limiting public access to records, 2008–2015 (9 states)



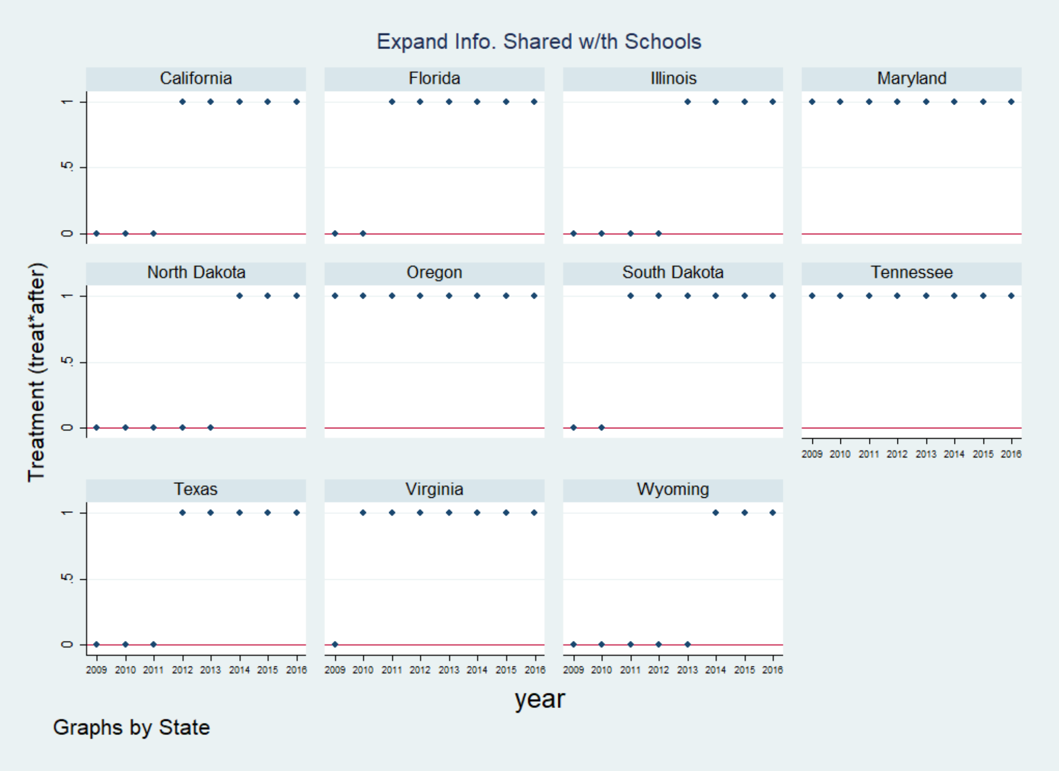
This figure presents geographical representation of states that enacted revisions to juvenile record laws between 2008 and 2015. From yellow to dark red colored states indicate recent years of enactment. Gray colored states indicate control states.

Figure 1.5: Enacted revisions to juvenile record laws: Expand eligibility for sealing and expungement, 2008–2015 (23 states)



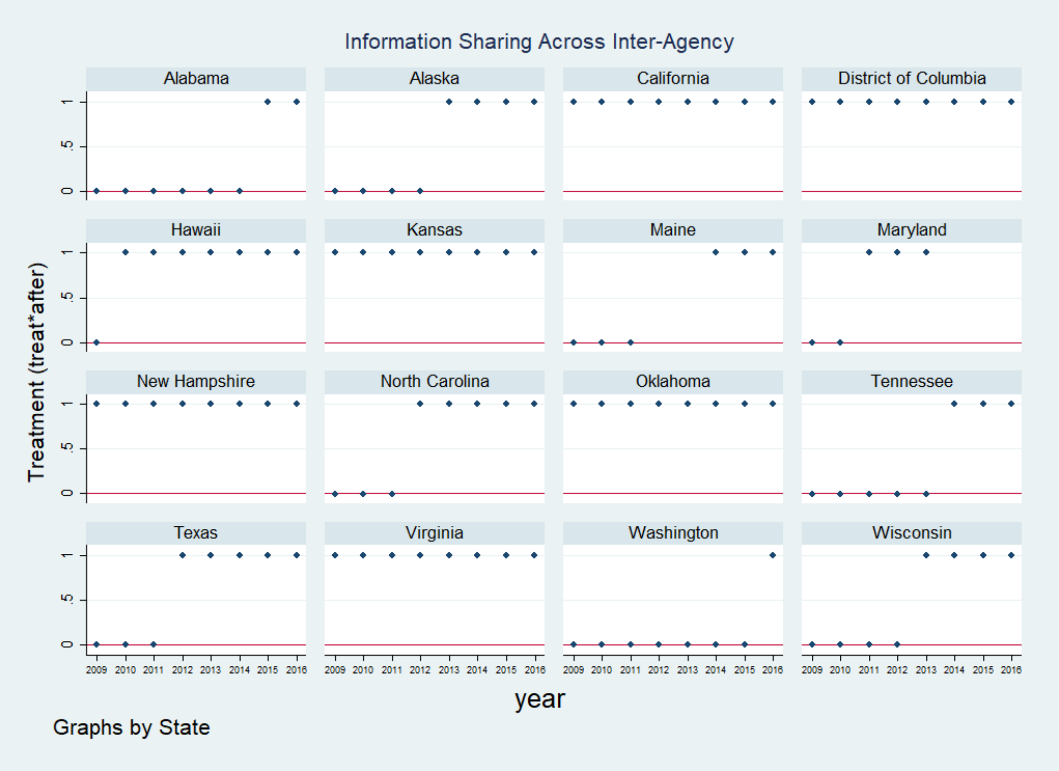
This figure presents geographical representation of states that enacted revisions to juvenile record laws between 2008 and 2015. From yellow to dark red colored states indicate recent years of enactment. Gray colored states indicate control states.

Figure 1.6: Treatment assignment: Expand information sharing with youth-serving systems



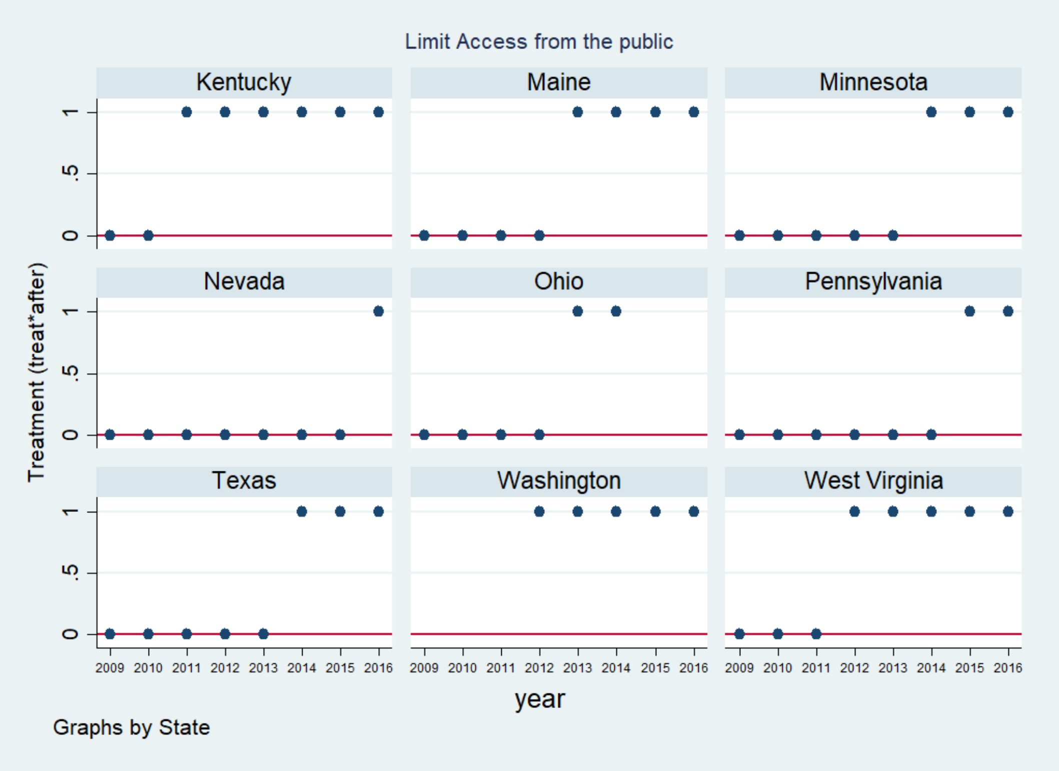
This figure presents *TreatAfter* indicator by state and year. States that are included are states that expand information sharing with schools. Treatment year is one-year lagged from actual law year. A state that enacted law in a contrasting direction are indicated with missing values in subsequent years.

Figure 1.7: Treatment assignment: Expand interagency information sharing



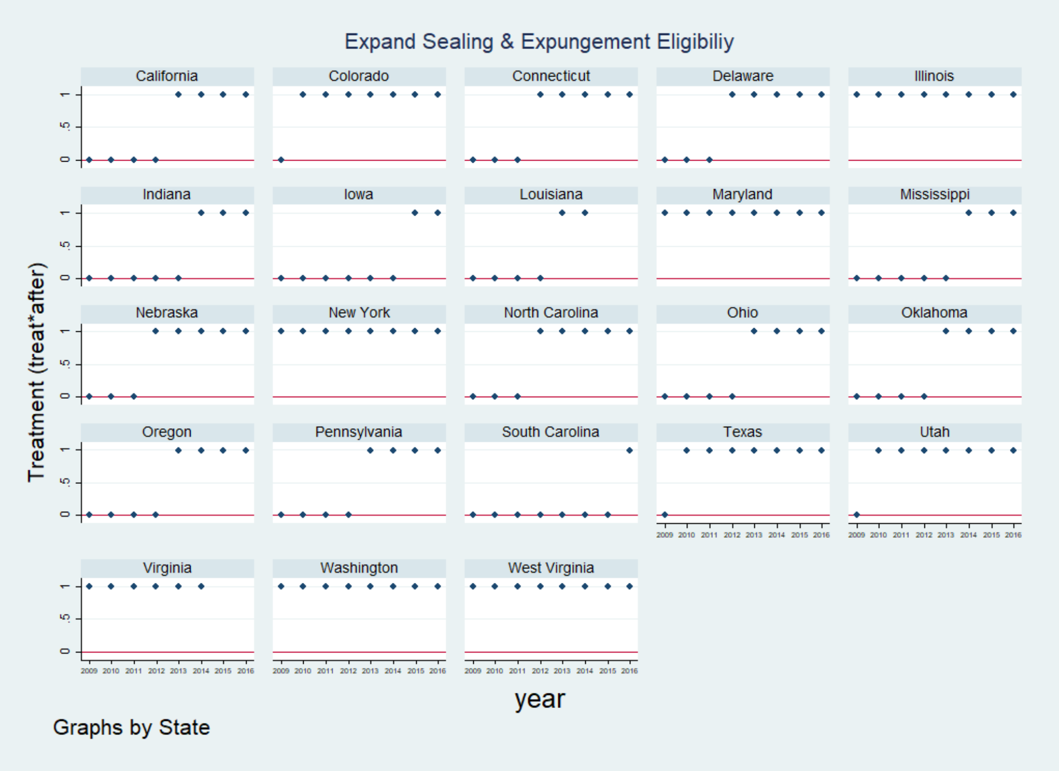
This figure presents *TreatAfter* indicator by state and year. States that are included are states that expand interagency information sharing. Treatment year is one-year lagged from actual law year. A state that enacted law in a contrasting direction are indicated with missing values in subsequent years.

Figure 1.8: Treatment assignment: Limit public access to records



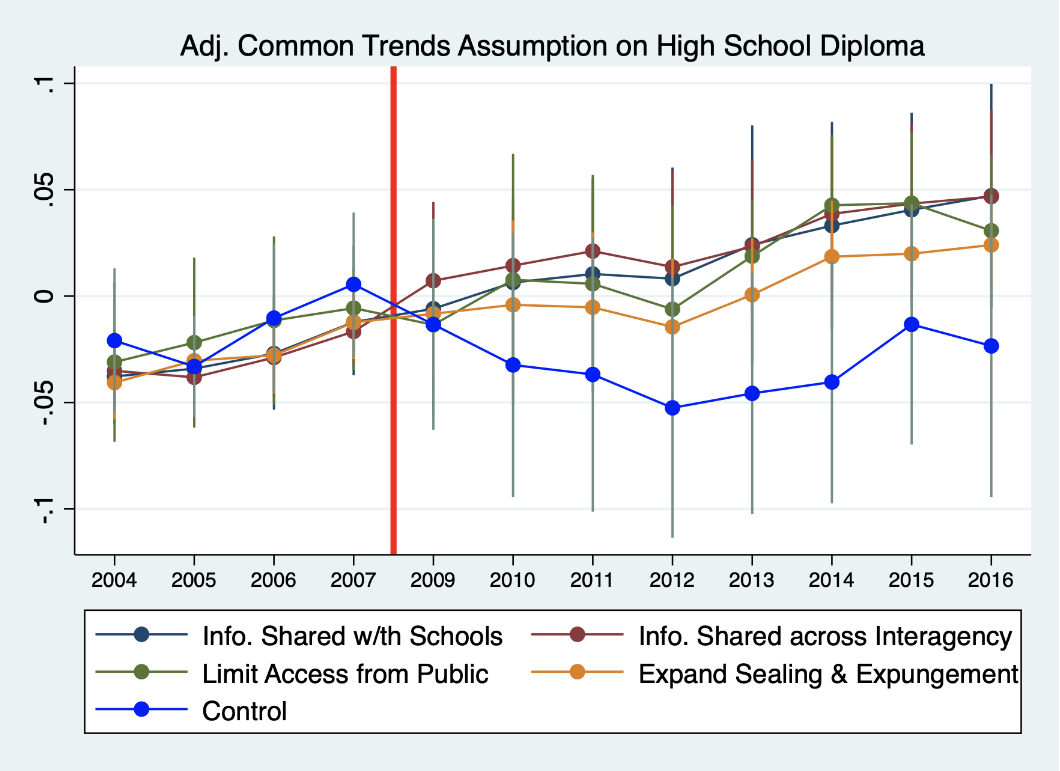
This figure presents $TreatAfter$ indicator by state and year. States that are included are states that limit public access to records. Treatment year is one-year lagged from actual law year. A state that enacted law in a contrasting direction are indicated with missing values in subsequent years.

Figure 1.9: Treatment assignment: Expand eligibility for sealing and expungement



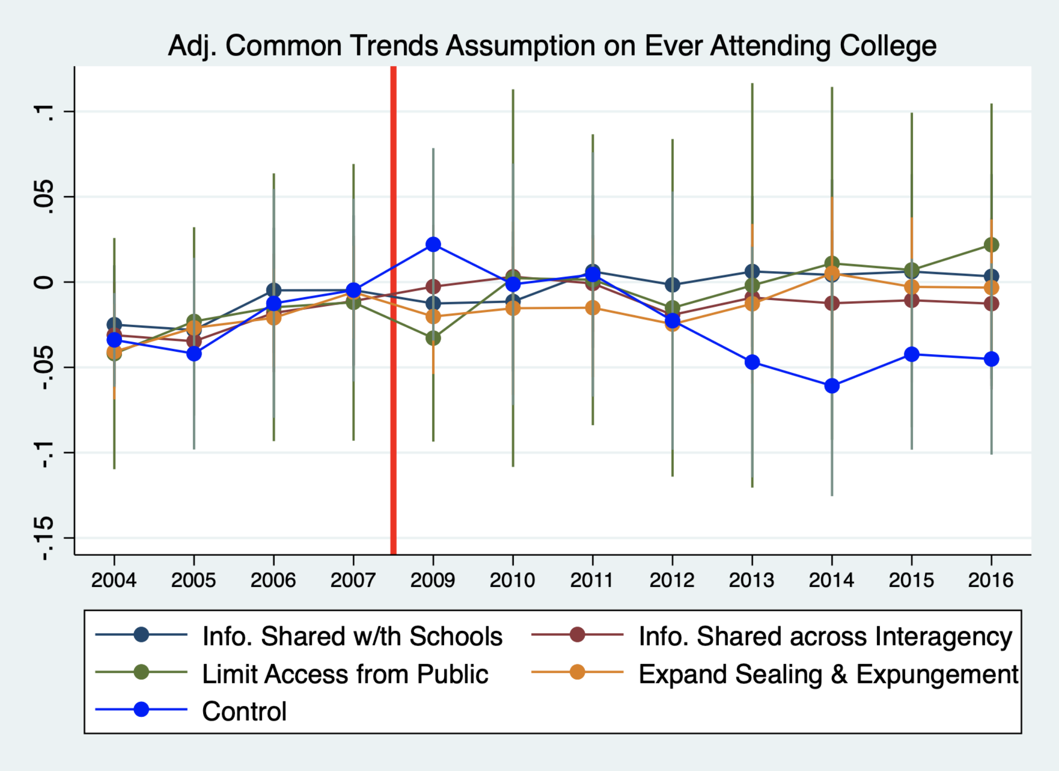
This figure presents $TreatAfter$ indicator by state and year. States that are included are states that expand eligibility for sealing and expungement. Treatment year is one-year lagged from actual law year. A state that enacted law in a contrasting direction are indicated with missing values in subsequent years.

Figure 1.10: Common trends assumption check: Probability of high school graduation



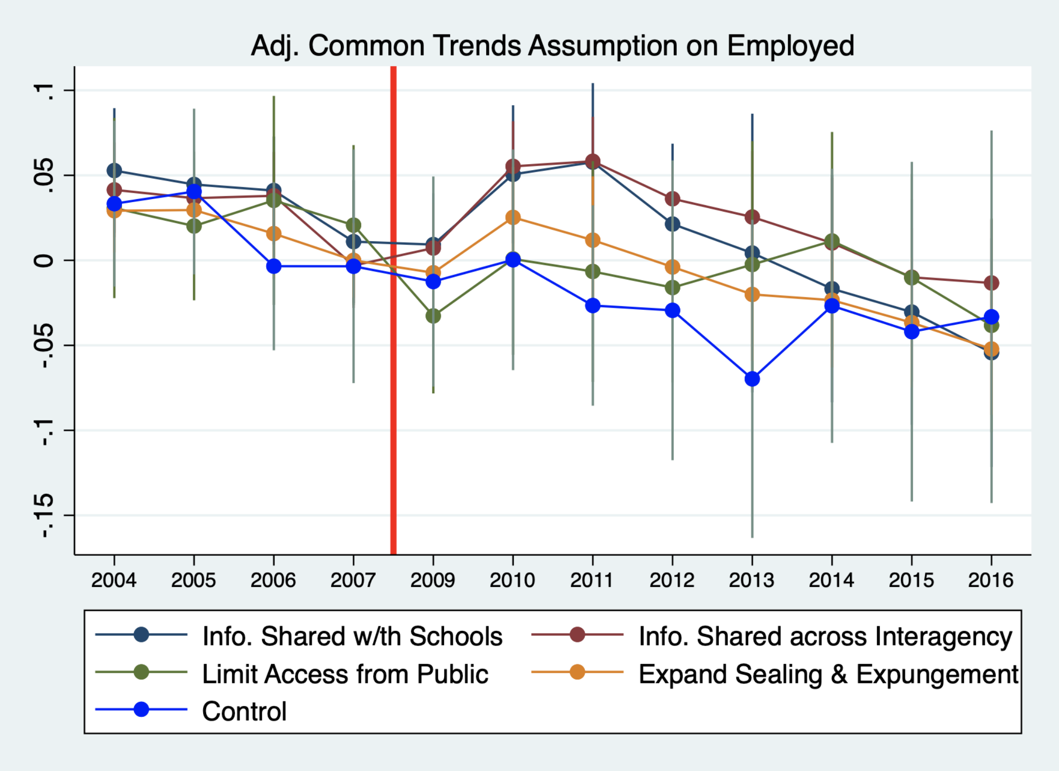
This figure shows adjusted common trends assumption check on outcome, probability of high school graduation. The adjusted (residualized) outcome is estimated using a regression of state and year fixed effects, individual controls, and state-specific controls on outcomes and plotting the residuals by each year. Year 2008 is excluded from the regression as a reference year. The red line indicates first year of law changes in treatment.

Figure 1.11: Common trends assumption check: Probability of ever attending college



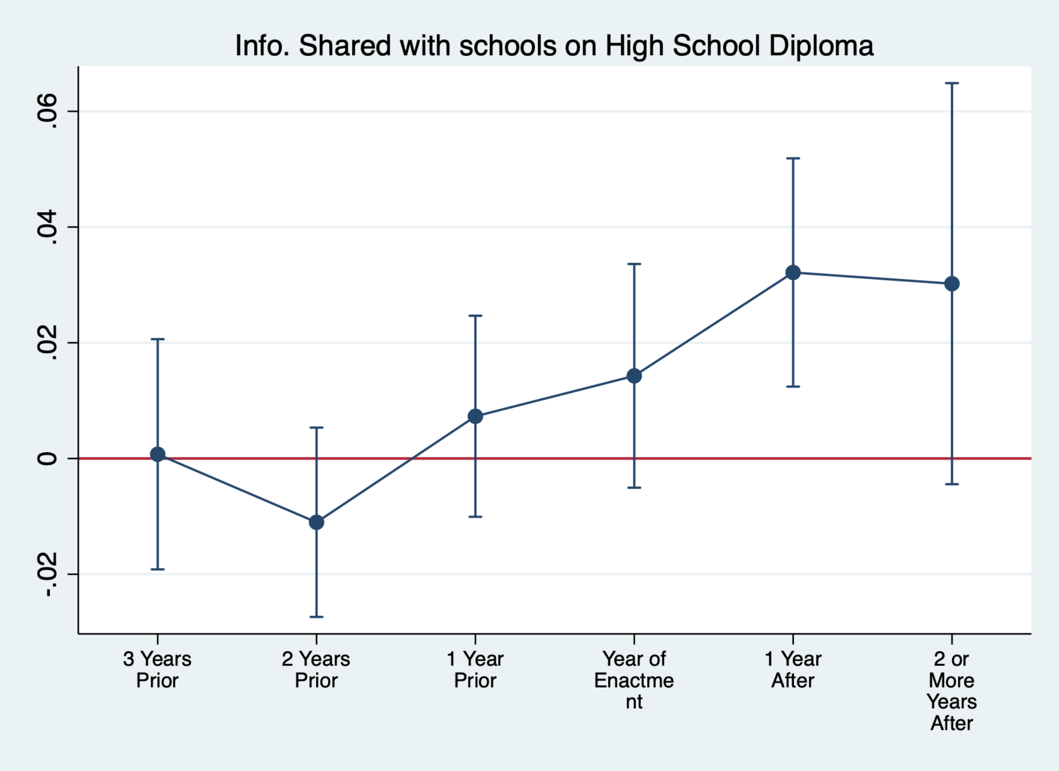
This figure shows adjusted common trends assumption check on outcome, ever attending college. The adjusted (residualized) outcome is estimated using a regression of state and year fixed effects, individual controls, and state-specific controls on outcomes and plotting the residuals by each year. Year 2008 is excluded from the regression as a reference year. The red line indicates first year of law changes in treatment.

Figure 1.12: Common trends assumption check: Employment



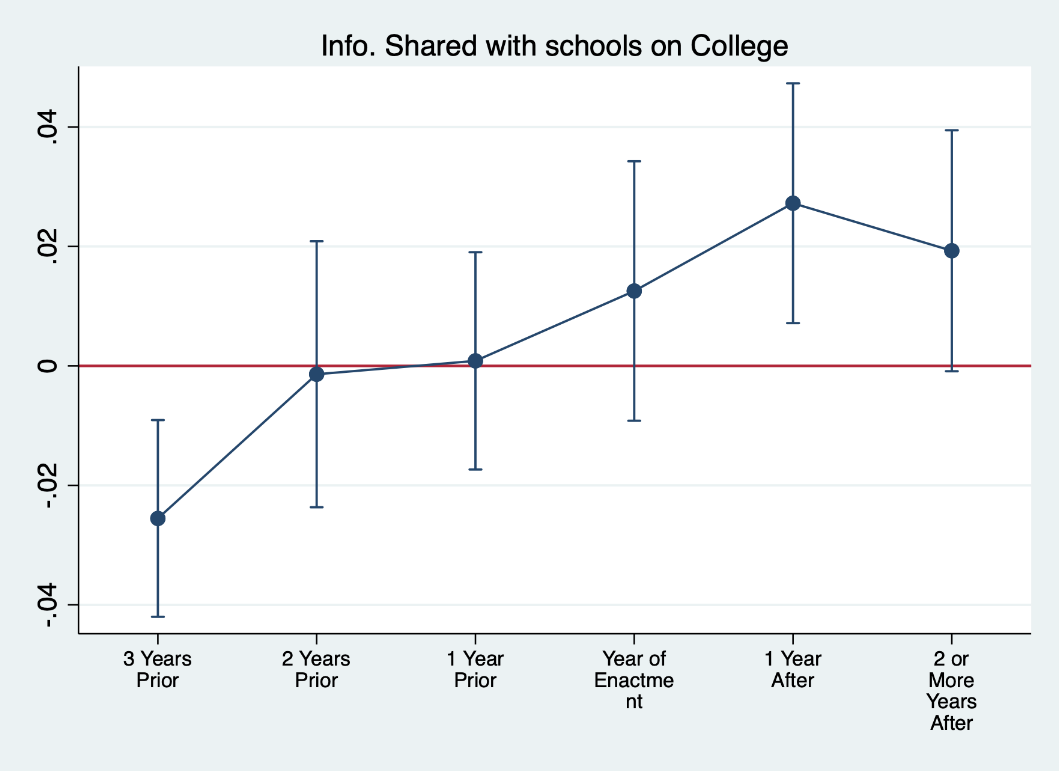
This figure shows adjusted common trends assumption check on outcome, employment. The adjusted (residualized) outcome is estimated using a regression of state and year fixed effects, individual controls, and state-specific controls on outcomes and plotting the residuals by each year. Year 2008 is excluded from the regression as a reference year. The red line indicates first year of law changes in treatment.

Figure 1.13: Leads and Lag of Information Sharing with Schools on High School Diploma



This figure shows present coefficient plots for $\beta_j$ for information sharing with schools on high school graduation and ever attending college, respectively. The x-axis indicate relative years since enactment from 3 years prior to 2 or more years after. All prior indicators except ”2 or more years after” have value one if relative year is equal to $\tau$. 2 or more years after indicator has value one for all years post 2 years of enactment. The bars indicate 95% confidence interval of each coefficient estimates. Samples are restricted to states that revised juvenile laws on information sharing with school and control states.

Figure 1.14: Leads and Lag of Information Sharing with Schools on Ever Attending College



This figure shows present coefficient plots for $\beta_j$ for information sharing with schools on high school graduation and ever attending college, respectively. The x-axis indicate relative years since enactment from 3 years prior to 2 or more years after. All prior indicators except "2 or more years after" have value one if relative year is equal to $\tau$. 2 or more years after indicator has value one for all years post 2 years of enactment. The bars indicate 95% confidence interval of each coefficient estimates. Samples are restricted to states that revised juvenile laws on information sharing with school and control states.

# IX Tables

Table 1.1: Descriptive Statistics: 2004–2007

| | Info. shared w/th schools Reform (11 states) | | | Info. shared across Interagency Reform (16 states) | | | Limit Access from Public Reform (9 States) | | | Expand Sealing & Expungment Reform (23 States) | | | Control Group No Change (12 states) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | p-val | Mean | SD | p-val | Mean | SD | p-val | Mean | SD | p-val | Mean | SD |
| *Outcomes* | | | | | | | | | | | | | | |
| High School Diploma | 79.9% | 0.004 | 0.3328 | 79.5% | 0.003 | 0.0795 | 80.6% | 0.004 | 0.89 | 80.4% | 0.003 | 0.8405 | 80.5% | 0.005 |
| Ever Attended College | 50.6% | 0.004 | 0.4231 | 50.4% | 0.004 | 0.5637 | 49.5% | 0.006 | 0.58 | 51.3% | 0.003 | 0.0654 | 50.0% | 0.006 |
| Employed | 62.8% | 0.004 | 0.3185 | 62.3% | 0.004 | 0.0968 | 63.9% | 0.005 | 0.6 | 61.6% | 0.003 | 0.0051 | 63.5% | 0.006 |
| *Individual Controls* | | | | | | | | | | | | | | |
| Female | 48.9% | 0.004 | 0.5051 | 49.0% | 0.004 | 0.6108 | 49.7% | 0.006 | 0.69 | 49.7% | 0.003 | 0.619 | 49.4% | 0.006 |
| Age | 21.093 | 0.018 | 0.0411 | 21.086 | 0.017 | 0.0664 | 21.055 | 0.022 | 0.47 | 21.062 | 0.013 | 0.2768 | 21.031 | 0.024 |
| white | 77.5% | 0.004 | 0.005 | 75.8% | 0.004 | 0 | 82.9% | 0.005 | 0 | 77.5% | 0.003 | 0.0018 | 79.3% | 0.005 |
| black | 13.4% | 0.003 | 0.0955 | 13.2% | 0.003 | 0.0409 | 10.6% | 0.004 | 0 | 14.3% | 0.003 | 0.9033 | 14.3% | 0.005 |
| Other Race | 9.1% | 0.003 | 0 | 11.0% | 0.003 | 0 | 6.5% | 0.003 | 0.76 | 8.3% | 0.002 | 0 | 6.3% | 0.003 |
| Married ever | 15.4% | 0.003 | 0.1467 | 15.9% | 0.003 | 0.016 | 15.3% | 0.004 | 0.24 | 14.0% | 0.002 | 0.2296 | 14.6% | 0.004 |
| Metro Area | 91.5% | 0.002 | 0 | 86.8% | 0.003 | 0.0029 | 82.1% | 0.004 | 0 | 86.7% | 0.002 | 0.002 | 85.3% | 0.004 |
| *State-Specific Controls* | | | | | | | | | | | | | | |
| Violent Crime Index | 400.741 | 2.178 | 0 | 275.88 | 0.875 | 0.0002 | 237.561 | 1.15 | 0 | 332.655 | 1.43 | 0 | 271.373 | 0.841 |
| Property Crime Index | 1332.04 | 3.164 | 0 | 1248.582 | 3.452 | 0 | 1264.246 | 3.411 | 0 | 1,279.93 | 2.088 | 0 | 1,160.64 | 4.652 |
| Drug abuse | 804.622 | 5.377 | 0 | 543.057 | 1.357 | 0.0011 | 544.866 | 1.668 | 0.02 | 668.811 | 3.419 | 0 | 551.147 | 2.081 |
| Unemployment (Oct.) | 4.976 | 0.008 | 0 | 5.057 | 0.007 | 0 | 5.191 | 0.007 | 0.48 | 5.164 | 0.005 | 0.0045 | 5.201 | 0.012 |
| log(juvenile population) | 15.31 | 0.006 | 0 | 15.014 | 0.008 | 0 | 14.818 | 0.008 | 0 | 14.924 | 0.006 | 0 | 14.245 | 0.006 |
| *Region* | | | | | | | | | | | | | | |
| New England | 0.0% | 0 | 0 | 2.3% | 0.001 | 0 | 2.0% | 0.001 | 0 | 1.7% | 0.001 | 0 | 14.6% | 0.004 |
| Middle Atlantic | 0.0% | 0 | 0 | 0.0% | 0 | 0 | 17.7% | 0.004 | 0 | 17.0% | 0.003 | 0 | 14.5% | 0.004 |
| East North Central | 11.4% | 0.003 | 0 | 4.9% | 0.002 | 0 | 17.0% | 0.004 | 0.08 | 16.4% | 0.003 | 0.0016 | 18.1% | 0.005 |
| West North Central | 1.5% | 0 | 0 | 2.6% | 0.001 | 0 | 8.5% | 0.002 | 0 | 2.8% | 0.001 | 0 | 11.2% | 0.004 |
| South Atlantic | 24.9% | 0.004 | 0 | 18.2% | 0.003 | 0.6192 | 2.2% | 0.001 | 0 | 14.3% | 0.002 | 0 | 17.9% | 0.005 |
| East South Central | 5.1% | 0.002 | 0 | 9.3% | 0.003 | 0 | 5.9% | 0.002 | 0 | 1.5% | 0.001 | 0 | 0.0% | 0 |
| West South Central | 20.9% | 0.004 | 0 | 24.0% | 0.004 | 0 | 34.7% | 0.006 | 0 | 17.3% | 0.003 | 0 | 5.1% | 0.002 |
| Mountain | 0.5% | 0 | 0 | 0.0% | 0 | 0 | 3.4% | 0.001 | 0 | 4.3% | 0.001 | 0 | 18.6% | 0.004 |
| Pacific | 35.6% | 0.004 | 0 | 38.7% | 0.004 | 0 | 8.6% | 0.003 | 0 | 24.7% | 0.003 | 0 | 0.0% | 0 |

*Note.*The table was generated using October CPS data for the pre-treatment period 2004–2007. Mean values are weighted. P-value is calculated using t-test between averages for the treated group versus the control group.

Table 1.2: Difference-in-Differences Estimates of Revisions to Record Laws on Educational Attainment and Employment, 2004–2016

|  | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) | (13) | (14) | (15) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| OUTCOMES | HS Diploma (excluding GED) (Age 18–24) |  |  |  |  | Ever Attended College (Age 18–24) |  |  |  |  | Employed (Age 18–24) |  |  |  |  |
| Info. shared with schools | 0.025** | 0.025** | 0.022** | 0.017*** | 0.021** | 0.025** | 0.023* | 0.020** | 0.035*** | 0.031*** | 0.002 | 0.001 | 0.001 | 0.010 | 0.018 |
| (11 states) | (0.009) | (0.009) | (0.008) | (0.006) | (0.010) | (0.012) | (0.012) | (0.009) | (0.011) | (0.010) | (0.008) | (0.008) | (0.006) | (0.010) | (0.012) |
| Observations | 76,033 |  |  |  |  | 76,033 |  |  |  |  | 76,033 |  |  |  |  |
| Interagency info. sharing | 0.010 | 0.009 | 0.008 | -0.005 | 0.029*** | -0.002 | -0.006 | -0.006 | -0.016 | 0.018** | 0.002 | -0.000 | 0.005 | 0.007 | 0.023*** |
| (16 states) | (0.008) | (0.008) | (0.008) | (0.008) | (0.005) | (0.009) | (0.009) | (0.009) | (0.011) | (0.008) | (0.015) | (0.014) | (0.008) | (0.008) | (0.007) |
| Observations | 81,758 |  |  |  |  | 81,758 |  |  |  |  | 81,758 |  |  |  |  |
| Limit Public Access | 0.009 | 0.008 | 0.005 | 0.017** | 0.009 | -0.011 | -0.014 | -0.011 | 0.001 | -0.026* | -0.010 | -0.011 | -0.012 | -0.031 | 0.010 |
| (9 states) | (0.008) | (0.007) | (0.007) | (0.008) | (0.010) | (0.016) | (0.015) | (0.013) | (0.017) | (0.015) | (0.014) | (0.014) | (0.014) | (0.019) | (0.016) |
| Observations | 58,817 |  |  |  |  | 58,817 |  |  |  |  | 81,758 |  |  |  |  |
| Expand Sealing & Expugement | 0.003 | 0.002 | 0.001 | -0.010 | 0.001 | -0.000 | -0.003 | -0.002 | -0.005 | -0.001 | 0.020** | 0.018** | 0.012 | 0.019* | 0.009 |
| (23 states) | (0.009) | (0.009) | (0.009) | (0.011) | (0.007) | (0.009) | (0.010) | (0.010) | (0.013) | (0.007) | (0.008) | (0.008) | (0.008) | (0.010) | (0.007) |
| Observations | 110,068 |  |  |  |  | 110,068 |  |  |  |  | 110,068 |  |  |  |  |
| State and Year FE | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X |
| Individual Controls |  | X | X | X | X |  | X | X | X | X |  | X | X | X | X |
| Time-Varying State Controls |  |  | X | X | X |  |  | X | X | X |  |  | X | X | X |
| Linear State Trends |  |  |  | X |  |  |  |  | X |  |  |  |  | X |  |
| Region × Year Dummies |  |  |  |  | X |  |  |  |  | X |  |  |  |  | X |

*Note.* Individual controls include: female, age, white, black, asian, married ever. State-time varying controls include metro area, violent crime, property crime, and drug abuse index, and seasonally adjusted unemployment rate. Huber-White robust standard errors are in parentheses and is clustered within each state. ***$p$ ¡ .01. **$p$ ¡ .05. *$p$ ¡ .1.

Table 1.3: Difference-in-differences Estimates of revisions to Record Laws on Education Attainment and Employment, Y2004-2016: Contrasting Any Record Laws versus Specific Record Laws

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| OUTCOMES | HS Diploma (excluding GED) (Age18-24) | | | | Ever Attended College (Age18-24) | | | | Employed (Age18-24) | | | |
| Info. shared w/th schools | 0.013* | 0.011* | 0.014* | | 0.022*** | 0.033*** | 0.031*** | | -0.012* | 0.004 | 0.006 | |
| (11 states) | (0.007) | (0.006) | (0.008) | | (0.005) | (0.010) | (0.008) | | (0.006) | (0.010) | (0.007) | |
| Interagency info. sharing | -0.001 | -0.012* | 0.008 | | -0.020*** | -0.022** | -0.007 | | 0.008 | 0.010 | 0.016* | |
| (16 states) | (0.007) | (0.007) | (0.007) | | (0.006) | (0.009) | (0.008) | | (0.007) | (0.007) | (0.008) | |
| Limit Public Access | -0.008 | 0.016* | -0.009 | | -0.019 | -0.006 | -0.025** | | -0.009 | -0.022* | -0.001 | |
| (9 states) | (0.006) | (0.008) | (0.006) | | (0.012) | (0.015) | (0.010) | | (0.011) | (0.012) | (0.008) | |
| Expand Sealing & Expugement | 0.003 | -0.006 | 0.001 | | 0.001 | -0.002 | -0.001 | | 0.013 | 0.017 | 0.013* | |
| (23 states) | (0.007) | (0.012) | (0.007) | | (0.007) | (0.013) | (0.006) | | (0.008) | (0.011) | (0.007) | |
| Any Laws | | | | 0.003 | | | | -0.004 | | | | 0.011 |
| (39 states) | | | | (0.007) | | | | (0.008) | | | | (0.007) |
| State and Year FE | X | X | X | X | X | X | X | X | X | X | X | X |
| Individual Controls | X | X | X | X | X | X | X | X | X | X | X | X |
| Time-Varying State Controls | X | X | X | X | X | X | X | X | X | X | X | X |
| Linear State Trends | | X | | | | X | | | | X | | |
| Region × Year Dummies | | | X | | | | X | | | | X | |
| Observations | 135,971 | | | | | | | | | | | |

*Note.* Individual controls include: female, age, white, black, asian, married ever. State-time varying controls include metro area, violent crime, property crime, and drug abuse index, and seasonally adjusted unemployment rate. Huber-White robust standard errors are in parentheses and is clustered within each state. ***$p$ ¡ .01. **$p$ ¡ .05. *$p$ ¡ .1.

Table 1.4: Difference-in-differences Aggregated States Estimates on Education Attainment and Employment, Y2004-2016s

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| OUTCOMES | HS Diploma (excluding GED) (Age18-24) | | | | Ever Attended College (Age18-24) | | | | Employed (Age18-24) | | | |
| Info. shared w/th schools<br>(11 states) | 0.024**<br>(0.008) | 0.021**<br>(0.007) | 0.017**<br>(0.006) | 0.021<br>(0.014) | 0.023*<br>(0.012) | 0.020*<br>(0.009) | 0.035**<br>(0.013) | 0.030*<br>(0.015) | 0.002<br>(0.010) | 0.002<br>(0.007) | 0.010<br>(0.015) | 0.018<br>(0.018) |
| Observations | 299 | | | | 299 | | | | 299 | | | |
| Interagency info. sharing<br>(16 states) | 0.008<br>(0.008) | 0.007<br>(0.008) | -0.005<br>(0.008) | 0.029***<br>(0.008) | -0.006<br>(0.008) | -0.006<br>(0.007) | -0.016<br>(0.011) | 0.018*<br>(0.009) | -0.000<br>(0.014) | 0.006<br>(0.007) | 0.008<br>(0.009) | 0.023<br>(0.014) |
| Observations | 359 | | | | 359 | | | | 359 | | | |
| Limit Public Access<br>(9 states) | 0.008<br>(0.009) | 0.005<br>(0.010) | 0.017<br>(0.012) | 0.009<br>(0.015) | -0.014<br>(0.014) | -0.011<br>(0.013) | 0.001<br>(0.019) | -0.026<br>(0.032) | -0.010<br>(0.017) | -0.012<br>(0.017) | -0.030<br>(0.028) | 0.010<br>(0.019) |
| Observations | 274 | | | | 274 | | | | 274 | | | |
| Expand Sealing & Expungement<br>(23 states) | 0.002<br>(0.009) | 0.001<br>(0.009) | -0.011<br>(0.010) | 0.001<br>(0.011) | -0.003<br>(0.009) | -0.002<br>(0.010) | -0.005<br>(0.013) | -0.001<br>(0.010) | 0.018**<br>(0.007) | 0.012*<br>(0.006) | 0.019**<br>(0.008) | 0.009<br>(0.009) |
| Observations | 465 | | | | 465 | | | | 465 | | | |
| State and Year FE | X | X | X | X | X | X | X | X | X | X | X | X |
| Individual Controls | X | X | X | X | X | X | X | X | X | X | X | X |
| Time-Varying State Controls | | X | X | X | | X | X | X | | X | X | X |
| Linear State Trends | | | X | | | | X | | | | X | |
| Region × Year Dummies | | | | X | | | | X | | | | X |

*Note.* Individual controls include: gender, age, white, black, asian, ever married. State-time controls include metro area, violent crime, property crime, and drug abuse index, and seasonally adjusted unemployment rate. Huber-White robust standard errors are in parentheses and is clustered within each state. ***$p$ ¡ .01. **$p$ ¡ .05. *$p$ ¡ .1.

Table 1.5: Difference-in-differences Heterogeneous Estimates Education Attainment and Employment, Y2004-2016

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
|---|---|---|---|---|---|---|---|---|---|---|
| Info. shared w/th schools (11 states)<br>OUTCOMES | HS Diploma (excluding GED) | | | Ever Attended College | | | Employed | | | Observation |
| **Male Only** | 0.016* | 0.005 | 0.021** | 0.026*** | 0.038** | 0.035*** | -0.010 | 0.007 | 0.006 | 68,221 |
| | (0.008) | (0.013) | (0.010) | (0.008) | (0.015) | (0.009) | (0.007) | (0.010) | (0.012) | |
| **Female Only** | 0.009 | 0.018*** | 0.006 | 0.018* | 0.028** | 0.026** | -0.012 | 0.002 | 0.008 | 67,750 |
| | (0.010) | (0.006) | (0.009) | (0.009) | (0.014) | (0.011) | (0.012) | (0.017) | (0.007) | |
| **Black** | 0.030** | 0.048*** | 0.019 | 0.020 | 0.058*** | 0.033** | -0.001 | -0.001 | 0.002 | 15,698 |
| | (0.013) | (0.016) | (0.014) | (0.015) | (0.019) | (0.016) | (0.012) | (0.020) | (0.014) | |
| **White** | 0.007 | 0.002 | 0.011 | 0.018*** | 0.026*** | 0.025*** | -0.015* | 0.005 | 0.004 | 107,265 |
| | (0.008) | (0.007) | (0.008) | (0.005) | (0.009) | (0.008) | (0.008) | (0.011) | (0.009) | |
| **Other Race** | 0.025 | 0.015 | 0.040* | 0.032 | 0.034 | 0.080** | 0.006 | 0.015 | 0.028 | 13,008 |
| | (0.020) | (0.027) | (0.021) | (0.034) | (0.054) | (0.037) | (0.018) | (0.035) | (0.024) | |
| **Age 18-22** | 0.012 | 0.017 | 0.013 | 0.018** | 0.031** | 0.024*** | -0.017 | -0.003 | -0.006 | 77,697 |
| | (0.009) | (0.011) | (0.008) | (0.007) | (0.013) | (0.008) | (0.012) | (0.014) | (0.011) | |
| **Age 22-25** | 0.016* | 0.008 | 0.017 | 0.028*** | 0.040** | 0.042*** | -0.003 | 0.013 | 0.024*** | 58,274 |
| | (0.009) | (0.010) | (0.011) | (0.010) | (0.015) | (0.015) | (0.009) | (0.016) | (0.008) | |
| **Drug Abuse** | 0.009 | -0.002 | 0.005 | 0.008 | 0.020 | 0.015 | 0.009 | 0.004 | 0.028** | 79,289 |
| **(top 50th quartile)** | (0.010) | (0.013) | (0.006) | (0.010) | (0.014) | (0.010) | (0.009) | (0.014) | (0.014) | |
| **Violent Crime Index** | 0.016* | 0.011 | 0.019* | 0.018 | 0.033** | 0.051*** | -0.008 | -0.012 | 0.011 | 79,148 |
| **(top 50th quartile)** | (0.008) | (0.010) | (0.010) | (0.011) | (0.014) | (0.015) | (0.009) | (0.015) | (0.012) | |
| **Property Crime Index** | 0.011 | 0.021** | 0.014 | 0.020** | 0.041** | 0.037*** | 0.009 | 0.011 | 0.023** | 82,799 |
| **(top 50th quartile)** | (0.009) | (0.010) | (0.010) | (0.010) | (0.016) | (0.013) | (0.007) | (0.019) | (0.011) | |
| State and Year FE | X | X | X | X | X | X | X | X | X | |
| Individual Controls | X | X | X | X | X | X | X | X | X | |
| Time-Varying State Controls | X | X | X | X | X | X | X | X | X | |
| Linear State Trends | | X | | | X | | | X | | |
| Region*Year Dummies | | | X | | | X | | | X | |

*Note.* Individual controls include: female, age, white, black, asian, married ever. State-time controls include metro area, violent crime, property crime, and drug abuse index, and seasonally adjusted unemployment rate. Huber-White robust standard errors are in parentheses and is clustered within each state. ***$p$ ¡ .01. **$p$ ¡ .05. *$p$ ¡ .1.

# Bibliography

[1] Adams, M. S., Robertson, C. T., Gray-Ray, P., & Ray, M. C. (2003). Labeling and delinquency. *Adolescence*, 38(149), 171.

[2] Aizer, A., & Doyle Jr, J. J. (2015). Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges. *The Quarterly Journal of Economics*, 130(2), 759–803.

[3] Apel, R., & Sweeten, G. (2010). The impact of incarceration on employment during the transition to adulthood. *Social Problems*, 57(3), 448-479.

[4] Bai, Y., Wells, R., & Hillemeier, M. M. (2009). Coordination between child welfare agencies and mental health service providers, children's service use, and outcomes. *Child abuse & neglect*, 33(6), 372–381.

[5] Beckett, K., & Western, B. (2001). Governing social marginality: Welfare, incarceration, and the transformation of state policy. *Punishment & Society*, 3(1), 43–59. doi: 10.1177/14624740122228249

[6] Bennett, . C. P., M. (2010). The law and economics of information sharing: The good, the bad and the ugly. *European Competition Journal*, 6(2), 311-337.

[7] Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1), 249–275.

[8] Brame, R., Bushway, S. D., Paternoster, R., & Turner, M. G. (2014). Demographic patterns of cumulative arrest prevalence by ages 18 and 23. *Crime & Delinquency*, 60(3),

471-486. Retrieved from https://doi.org/10.1177/0011128713514801 (PMID: 26023241) doi: 10.1177/0011128713514801

[9] Brame, R., Turner, M. G., Paternoster, R., & Bushway, S. D. (2012). Cumulative prevalence of 47 arrest from ages 8 to 23 in a national sample. *Pediatrics*, 129(1), 21–27. Retrieved from https://pediatrics.aappublications.org/content/129/1/21 doi: 10.1542/peds.2010-3710

[10] Chuang, E., & Wells, R. (2010). The role of inter-agency collaboration in facilitating receipt of behavioral health services for youth involved with child welfare and juvenile justice. *Children and youth services review*, 32(12), 1814–1822.

[11] Doleac, J. L., & Hansen, B. (2017). Moving to job opportunities? the effect of 'ban the box' on the composition of cities., *American Economic Review*, 107(5), 556 - 559.

[12] Feierman, J., Levick, M., & Mody, A. (2009). The school-to-prison pipeline... and back: Obstacles and remedies for the re-enrollment of adjudicated youth. *NYL Sch. L. Rev.*, 54, 1115.

[13] The Community Change & National Employment Law Project, T. C. (2014, August). *The "wild west" of employment background checks: A reform agenda to limit conviction and arrest history abuses in the digital age.* Retrieved from `https://nelp.org/wp-content/uploads/2015/03/Wild-West-Employment-Background-Checks-Reform-Agenda.pdf`

[14] Gottfredson, D. C., & Barton, W. H. (1993). Deinstitutionalization of juvenile offenders. Criminology, 31(4), 591–611. Griffin, P. (2000). *Separate tables: Interagency information sharing in real life.* National Center for Juvenile Justice.

[15] Hirschfield, P. J. (2008). The declining significance of delinquent labels in disadvantaged urban communities. *Sociological Forum*, 23(3), 575 - 601. Retrieved from `https://ezproxy.cul.columbia.edu/login?qurl=https%3a%2f%2fsearch.ebscohost.com%2flogin.aspx%3fdirect%3dtrue%26db%3da9h%26AN%3d33246268%26site%3dehost-live%26scope%3dsite`

[16] Hjalmarsson, R. (2008). Criminal justice involvement and high school completion. *Journal of Urban Economics*, 63(2), 613–630.

[17] Kirk, D. S., & Sampson, R. J. (2013, jan). Juvenile arrest and collateral educational damage in the transition to adulthood. *Sociology of Education*, 86(1), 36–62. doi: 10.1177/0038040712448862

[18] Lambie, I., & Randell, I. (2013). The impact of incarceration on juvenile offenders. *Clinical Psychology Review*, 33(3), 448–459.

[19] Lemert, E. M. (1951). *Social pathology; a systematic approach to the theory of sociopathic behavior.*

[20] Leone, P. E., & Weinberg, L. A. (2010). *Addressing the unmet educational needs of children and youth in the juvenile justice and child welfare systems.* Georgetown University, Center for Juvenile Justice Reform.

[21] Litwok, D. (2014). *Have you ever been convicted of a crime? the effects of juvenile expungement on crime, educational, and labor market outcomes.* (Unpublished Manuscript.)

[22] Lovenheim, M. F., & Owens, E. G. (2014). Does federal financial aid affect college enrollment? evidence from drug offenders and the higher education act of 1998. *Journal of Urban Economics*, 81, 1–13.

[23] Marsh, S. (2014). *School pathways to the juvenile justice system: The context for a practice guide for courts and schools.* National Council of Juvenile and Family Court Judges.

[24] Nagin, D., & Waldfogel, J. (1995). The effects of criminality and conviction on the labor market status of young british offenders. *International Review of Law and Economics*, 15, 109–126.

[25] Nye, F. I., Short Jr, J. F., & Olson, V. J. (1958). Socioeconomic status and delinquent behavior. *American Journal of Sociology*, 63(4), 381–389.

[26] The Concil of State Government (CSG) Justice Center, T. C. (2018). *The second chance act fact sheet.* Retrieved 2019-01-22, from `https://csgjusticecenter.org/wp-content/uploads/2018/07/July-2018_SCA_factsheet.pdf`

[27] OJJDP Statistical Briefing Book. Online. Available: `http://www.ojjdp.gov/ojstatbb/crime/JAR_Display.asp?ID=qa05201`. October 22, 2018.

[28] Pierce, M. W., Runyan, C. W., & Bangdiwala, S. I. (2014). The use of criminal history information in college admissions decisions. *Journal of School Violence*, 13(4), 359–376.

[29] Puzzanchera, C. and Kang, W. (2017). *"Easy Access to FBI Arrest Statistics 1994-2014"* Online. Available: `http://www.ojjdp.gov/ojstatbb/ezaucr/`

[30] Richardson, T., DiPaola, T., & Gable, R. K. (2012). *Former juvenile offenders re-enrolling into mainstream public schools.*

[31] Rosenthal, A., NaPier, E., Warth, P., & Weissman, M. (2015). *Boxed out: Criminal history screening and college application attrition.* Center for Community Alternatives, Inc. Retrieved from http://www.communityalternatives.org/fb/boxed-out.html

[32] Seigle, E., Walsh, N., & Weber, J. (2014). *Core principles for reducing recidivism and improving other outcomes for youth in the juvenile justice system.* Council of State Governments.

[33] Shah, R. S., Fine, L., & Gullen, J. (2014). Juvenile records: A national review of state laws on confidentiality, sealing and expungement. *Juvenile Law Center*, 1-50.

[34] Shah, R. S., & Strout, J. (2016, Febraury). Future interrupted: The collateral damage caused by proliferation of juvenile records. Retrieved from `https://jlc.org/sites/default/files/publication_pdfs/Future%20Interrupted%20-%20final%20for%20web_0.pdf`

[35] Sickmund, M., & Puzzanchera, C. (2014). Juvenile offenders and victims: 2014 national report. Stewart, R. (2015, November). Criminal records and college applications. (Paper presented at the annual meeting of the American Society of Criminology)

[36] Sickmund, M., Sladky, T.J., Kang, W., & Puzzanchera, C. (2017) ”Easy Access to the Census of Juveniles in Residential Placement.” Online. Available: `http://www.ojjdp.gov/ojstatbb/ezacjrp/`

[37] Stewart, R., & Uggen, C. (2018). *Criminal records and college admissions: A national experimental audit (Tech. Rep.).* Working paper draft.

[38] Sweeten, G. (2006). Who will graduate? disruption of high school education by arrest and court involvement. *Justice Quarterly*, 23(4), 462–480.

[39] Tanner, J., Davies, S., & O’Grady, B. (1999). Whatever happened to yesterday’s rebels? longitudinal effects of youth delinquency on education and employment. *Social Problems*, 46(2), 250-275.

[40] Teske, S. C. (2011). A study of zero tolerance policies in schools: A multi-integrated systems approach to improve outcomes for adolescents. *Journal of Child and Adolescent Psychiatric Nursing*, 24(2), 88–97.

[41] Teske, S. C., Huff, B., & Graves, C. (2013). Collaborative role of courts in promoting outcomes for students: The relationship between arrests, graduation rates, and school safety. *Family Court Review*, 51(3), 418 - 426.

[42] Torbet, P., Gable, R., Montgomery, I., & Hurst, H. (1996). *State responses to serious & violent juvenile crime.* DIANE Publishing.

[43] Wachter, H. H. D. T. . T. D., A. (2017). *Collecting data and sharing information to improve school-justice partnerships.* National Council of Juvenile and Family Court Judges.

[44] Willison, J. B., Mears, D. P., Shollenberger, T., Owens, C., & Butts, J. A. (2010). Past, present, and future of juvenile justice: Assessing the policy options (apo). *Washington DC: The Urban Institute.*

[45] Wojcik, L., Schmetterer, K., & Naar, S. (2008). From juvenile court to the classroom: The need for effective child advocacy. *Chicago: DLA Piper.*

# Chapter 2

# Understanding Dropouts Among Community College Students: Using Cluster Analysis and Data Mining

Rina Seung Eun Park

# I Introduction

Low completion rates at community colleges in the U.S. continue to present a grim picture. Only about 24% of first-time, full-time undergraduate students who began seeking a certificate or associate degree earned that degree within three years of initial enrollment (150% of scheduled time) at public 2-year institutions (McFarland, et al., 2017). Much of the prior research on low completion rates has focused on students who discontinue enrollment early (within one year) and on identifying factors that are associated with early-stage momentum toward college completion, such as student family background, high school preparation, college enrollment immediately after high school, committed goal to completing a degree, and full-time attendance (Adelman, 2006; Calcagno, Crosta, Bailey, & Jenkins, 2006). Jenkins and Bailey (2017) summarizes first-semester and first-year momentum indicators that relates to success: attempting at least 15 credits in the first semester and 30 credits in the first year, passing a college-level Math and English in the first year, particularly for students placed in remedial education, and passing at least 9 credits in college-level major courses in the first year. This research has illuminated early-stage retention barriers but provided little information about late dropouts who decide to leave college after their first year.

Unlike a handful of research on early dropouts, we know very little about students who drop out in later years throughout the college process. Even more, we know very little about the typologies of college dropouts. Recent empirical findings shed light on this search for heterogeneous dropout groups. For an example, Stratton, O'Toole, & Wetzel (2008) find that there are two groups of students who have distinctive characteristics among the second year withdraws, a short-term stop out (who returns within a year) and a long-term dropout (who do not return more than a year). Another study finds that, among students who drop out from a two-year or a non-selective four-year institution in Ohio and Florida, about one-third of the population leave college after completing 75% of their required degree credits (Mabel & Britton, 2018). These close-to-completion dropouts face different barriers from the traditional early dropouts, for example, the transition to upper division coursework. Prior findings suggest that different dropout groups can have very different reasons for leaving college and it is important to understand the population and employ different supports and interventions to help students achieve their completion goals.

In this paper, I use administrative data from a large community college system in a single state. This study uses two analytic approaches, cluster analysis and classification, to better understand dropouts at community colleges. The first analysis involves using cluster analysis to identify different dropout groups (clusters). The second analysis uses classification algorithms to build a reliable prediction model on each dropout group that can identify risk factors.

Cluster analysis and classification algorithms have gained popularity in the education data mining literature as the availability of electronic data tremendously grew, particularly in the e-learning environment. Specifically, cluster analysis is used to analyze student learning patterns or behaviors using web-login history, understand learning styles such as content or environment preferences, or identify different e-learning student profiles (Dutt, Ismail, & Herawan, 2017). Similarly, classification algorithms are widely used in an online environment to model student characteristics, behaviors, predict performance, or measure assessment (Pena-Ayala, 2014).[1] However, both methodological approaches have rarely been used applied in a traditional classroom setting, which is a common instruction form at US colleges. The novelty and complexity of the algorithms limited its usage to researchers who are in the field of computer science or statistics, however, have less familiarity in an educational context. Consequently, both methodologies have less been applied to administrative datasets, which consist of greater detail about a student's college experience. This paper contributes to the literature by illustrating a practical use of these algorithms with administrative data to answer the most pressing issues in a traditional community college classroom

[1] Pena-Ayala (2014), Romero and Ventura (2007), and Dutt, Ismail, and Herawan, (2017) provides a comprehensive review of education data mining literature using cluster analysis or classification algorithms.

settings.

Understanding dropouts can be particularly policy- and practice- relevant to the community college context. Community colleges enrollees are more likely to be on the margin of college attendance given a larger proportion of non-traditional students and of open-access admissions. Indeed, about 52% of students in the sample of this study remain as non-completers (did not transfer to 4-year degree institution, have not earned a certificate or diploma, and are not enrolled within six years after entry). This makes the community college population a key target population for understanding dropouts.

As the preview of results, using hierarchical clustering analysis, this paper identifies four typologies of dropouts: college trials (first semester dropouts), high-credit medium dropouts, low-credit medium dropouts, and late dropouts. The four dropout groups are then, matched to students who successively complete college and have similar enrollment history. Three prediction models, logistic regression with elastic net, random forest, and gradient boosting, are applied to model the prediction on completion by each dropout group. Using 10 fold cross-validation on the training set (70% of the sample), gradient boosting algorithm performed the best with .719 to .868 AUC values. Variable importances suggest that each group has a different set of variables that contribute most to completion. In specific, I find third-year enrollment variables (i.e., credits withdrawn, full-time status, and remedial credits attempted) largely contributing to predictions of completion among the matched late dropout group. On the contrary, math and science credits earned in the first and second year are most predictive of completion among the matched high-credit medium dropout group. Similarly, gradient boosting performed the best on out-of-sample (remaining 30%) prediction.

This paper contributes to the literature in three ways. First, this paper identifies four typologies of community college dropouts, who are in the margin of college attendance. I find that these four groups have distinguishable enrollment patterns in two dimensions: the first dimension is relative time since entry and the second dimension is attempted credits within each term. In addition, I find representative characteristics of the four cluster groups that institutions can practically use to identify dropout types among at-risk students. Second, I identify different sets of variables that are important in predicting completion within each cluster group. This set of variables can provide guidance to institutions toward developing targeted supports and interventions that address diverse needs. The targeted interventions should accompany with efficacy study in order to make a causal interpretation, which this paper cannot provide. In addition, I provide suggestive evidence that identifying dropout typology prior to building an early alert system can improve predictive power. Lastly, this paper contributes to the general education literature by introducing the usage of machine learning techniques in a traditional classroom setting and applying matching to link clustering assignment with predictions using high dimensional data.

The remainder of the paper proceeds as follows: Section 2 provides background on dropouts at community colleges. Section 3 introduces research questions and section 4 describes data and sample of this study. In section 5, I describe cluster analysis methodology and present results. Section 6, I describe matching and classification methodology and present results. Section 6 concludes.

# II Dropouts at Community College

Among 2013-2014 first-time, full-time, degree/certificate-seeking cohort at a public 2-year institution in 2013-2014, on average, 23.6 percent of students graduated with a certificate or associate degree within 150 percent of normal time.[2] For the past 13 years, graduation rate has not improved and ranged between 19 to 24 percent. Among full-time degree seeking students who entered a public 2-year institution between 2015-2016, on average, 62 percent of students returned to school or have received a degree in their subsequent year since entry.

Although the term ”college completion” is defined differently by institutions, the term

[2] Statistics retrieved from Digest of Education Statistics 2017, table 326.20.

generally involves a time metric (e.g. within relative years since entry, 150% of published time for the program), a particular subset of student composition (e.g. degree seeking or full-time), and a measure of completion (e.g. degree attainment or transferred to four year college). For an example, Federal Student Aid agency (FSA) defines graduation rate as percentage of first-time, first-year undergraduates who complete their program within 150% of published time for the program and IPEDS definition further restricts to full-time students obtaining a degree from their first institution.[3]

At community colleges, it is quite common to use longer than the conventional 150% program time (e.g., six years since entry) as matriculated students are largely non-traditional, less academically achieving, and pursuing a shorter degree program (one year for certificate and two years for an associate degree or four-year college transfer).[4] The community college system of this study also follows the six-year mark for their internal performance measures, which I follow in this paper. It is worth pointing out that community colleges, in addition to degree completion, have a mission to prepare students to transfer to a four-year institution. As a result, a student who has transferred to a four-year college is counted towards completion. In this paper, I follow the definition of dropout commonly used at community colleges; that is, my dropout population is defined as students who do not have a degree (certificates or associates), did not transfer to a four-year college, and is not enrolled in their fall term of the sixth year.

The process of student's dropout/persistence choice in college draws upon three theoretical frameworks; Tinto's Student Integration Model (1993), Bean's Student Attrition Model (1985), and College Choice Nexus Model (St. John, Cabrera, Nora, & Asker, 2000). Tinto's (1993) Student Integration model suggests that student's persistent choice is a function of individual characteristics, academic performance, financial constraint, and social integration

[3] Definition of completion for FSA and IPEDS is retrieved from `https://fafsa.ed.gov/help/fotw91n.htm` and `https://nces.ed.gov/fastfacts/display.asp?id=40`, respectively.

[4] For an example, using Beginning Postsecondary Students (BPS) survey for 2003 cohort, a report shows that about 8% received a certificate and about 14% received an associate degree within six years since entry, which is about 10% higher than IPEDS graduation rate for that particular cohort (Ma & Baum, 2016).

to college. Bean's (1985) Student Attrition model adds environmental factors (e.g. employment opportunities) as another factor in determining student retention choice. The most recent model, College Choice Nexus Model, is a combination of these two prior models and adds a dynamic process in student's choice.

Based on the three theoretical models, empirical research has demonstrated that good academic preparation, marital and parental status, and sequential enrollment histories are important predictors to completion (Adelman, 2006; Desjardins, Ahlburg, and McCall, 2002; Willett and Singer, 1991; Ishitani, 2006). In addition, accumulating credit milestones, passing gatekeeper courses, and financial aid, particularly in the first semester, have shown to have an important contribution to student's choice to persist in college (Adelman, 2006; Attewell, Heil, & Reisel, 2012; Bailey & Alfonso, 2005; Stratton, O'Toole, & Wetzel, 2008). These empirical studies rely on the premises that dropout choices are identical across all college students. However, recent studies find evidence that students who leave college can have very different reasons (e.g., short-term stop out versus long-term dropout or early versus close-to-completion dropouts) and their needs for completion can be different. Given these recent findings, it seems plausible to question the hypothesis on homogeneity decisions of leaving college. This paper aims to fill in the gap by asking the following questions, (1) are there distinguishable groups of students who leave college, (2) what are their different characteristics, and (3) what are their different needs to improve retention? Are the needs different by dropout groups?

## III Research Questions

The objective of this study is to understand the characteristics of students who drop out from community colleges. This paper has the following two research questions:

1. Among students who drop out, can we identify separable groups (clusters) of students by using enrollment historical patterns? How are the demographic characteristics of

each group different from each other? What dimensions (e.g., time) distinguish different groups of students?

2. What are the important variables that largely contribute to predictions on completion? And how do they differ by each dropout typologies?

## IV Data and Sample

The administrative data for this study comes from all of the community colleges in a single state (more than 50 individual institutions). The data consists of all first-time in college students during Fall 2002 through Fall 2004 with 6 years of follow up information. The data includes information on demographic data (e.g. dependency status, limited English proficiency indicator, citizenship, in-state, race, age at entry, has prior entry flag), transcript data (e.g. total term credits, college-level credits, and remedial credits attempted and earned, grades for each term, number of withdraw classes by semester), financial aid information (e.g. indicator and amount of pell, total grant, loan, and aid), credential data using the National Student Clearinghouse (NSC) (e.g. received aa, certificate, or ba, semesters received in four-year or for-profit schools within two through six years relative to entry), and earnings data (e.g. monthly earnings, number of employment in each semester).

I restrict my analysis to fall cohorts to be consistent when using relative terms. Furthermore, about 4% of students who have missing values in their first-term GPA because they had incomplete, withdrawn, unknown, or received other letter grades for all classes in the first term are dropped from the sample. Around 50 students who are missing age at entry are dropped from my sample, as well. The final sample included total 78,496 students.

The main population of interest is ”college dropouts” defined as students who did not transfer, do not have a degree, and are not enrolled by their first term of the sixth year. Figure 2.1 shows that about 52% of the sample are college dropouts according to this definition. Table 2.1 provides demographic characteristics, enrollment history, and financial aid information about the sample. The first two columns describe the average characteristics of completers and dropouts in my analysis sample, for comparison. The third and fourth column present national averages from the Beginning Postsecondary Students (BPS) 2003/2004 survey, restricting to students whose first institution is a public two-year college during the academic year 2003-2004. To closely align with the definition used in this paper, the dropout sample is selected as students who are not enrolled and have no degree within five years. The complete sample is selected as students who are either currently enrolled or not enrolled but, have AA or certificate or transferred to a 4-year within five years. I use the variable "cumulative persistence and attainment anywhere in 2008-2009 (within 5 years since entry cohort)" to distinguish between dropout and completer sample. This selection resulted in 44% of a dropout sample, which is slightly smaller than my analytic dropout sample. From Table 2.1, I find that my dropout sample, on average, has less female, older, fewer whites and more black, have fewer prior enrollments, and more in-state students compared to my sample completers. In addition, my dropouts received more TANF, more financial aid, earned less cumulative credits, and has a lower GPA, on average. Among my sample of completers, around 16% of students receive a certificate, 41% received an associate, 14% received a bachelors, and 56% ever transferred to a four-year institution within six-year since entry. Comparing to the national averages, my dropout sample has fewer females, more white and black population, more US citizens, and less likely to receive any financial aid. In addition, my sample dropouts have higher credits earned and higher GPA than the national averages.

# V Cluster Analysis

## A Methodology

The first step of my analysis is to find if any, there are distinct groups of students among the dropout population. Cluster analysis is an appropriate technique to perform this kind of analysis. Cluster analysis is used to group or segment a collection of objects into similar subgroups, where similarity/dissimilarity is measured by a choice of a distance matrix. Observations within the same cluster have smaller pairwise distances than with observations in a different cluster (Hartigan, 1975; Gordon 1999; Kaufman & Rousseeuw, 1990). It is worth mentioning that cluster analysis does not aid in identifying the ”relationship” between predictors and outcome class. Rather, cluster analysis is used to discover the underlying patterns of data without a target variable, outcome, and group data into meaningful categories. Traditionally, there are two clustering techniques, hierarchical and partitional. For my analysis, I choose Hierarchical Cluster Analysis (HCA) for my main analysis but, I compare my analysis using Clustering Large Applications (CLARA), which is a partitional approach in the appendix.

HCA has considerably been used in the education data mining literature to detect common data patterns in student profiles, activity, and learning patterns. For an example, Lee, et al., (2016) uses HCA and heatmap to understand the patterns of student activity (e.g.. number of attachment views, discussion participation, etc) and its correlation to a final course grade of a course. Nitkin (2018) uses HCA to examine pattern association between instruction method and longitudinal academic outcomes, such as exit slip scores and content levels for technology-based personalized instructions. The most relevant to this paper is Bowers (2010), who uses HCA to analyze historical K-12 grading patterns and has linked to dropout indicator. In the post-secondary context, Asif, et al., (2017) used HCA to group students by yearly progression patterns of their indicators, which are derived from predicting students’ performance from high school variables and first two years of college course history. However, Asif, et al., (2017) analysis is conducted on a single degree program from one institution resulting in a small sample size of 210 students in Pakistan.

The HCA algorithm particularly, an agglomerative HCA, is a bottom-up approach, where every data point initially starts as a cluster on its own. At each iterative step, two most similar clusters are grouped together building a hierarchical representation. The algorithm stops when all of the data points are grouped to a single cluster. The entire iterative process represents an ordered sequence of groupings.

Several decisions are made by the user when implementing HCA. First, the user should specify a dissimilarity measure between clusters that will be used to determine the next two merging clusters. As suggested in the literature, I use uncentered Pearson correlation to calculate the pairwise distance between observations and average linkage to calculate intergroup dissimilarity between clusters (groups of observations)(Bowers, 2010; Romesburg, 1984). Uncentered Pearson correlation is calculated as:

$$D(x_i, x_j) = \frac{1}{n}\sum_{n}^{i=1}(\frac{x_i}{\tilde{\sigma_{x_i}}})(\frac{x_j}{\tilde{\sigma_{x_j}}}) \tag{2.1}$$

$$\text{where, } \tilde{\sigma_{x_i}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i)^2}, \tilde{\sigma_{x_j}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_j)^2}$$

Using uncentered pearson correlation as distance metrix has several advantages when trying to detect patterns on multidimensional data. One that is particularly useful for this study is its ability to detect feature changes without neglecting absolute amplitude differences (Anderberg, 1973). To calculate dissimilarity between two clusters, $C_1, C_2$, the average linkage is calculated as below:

$$d(C_1, C_2) = \frac{N_{C_1}}{N_{C_2}}\sum_{x_i \in C_1}\sum_{x_j \in C_2} D(x_i, x_j) \tag{2.2}$$

The average linkage for calculating dissimilarity between clusters is particularly helpful when addressing missing data in this study. The average linkage takes the average of pairwise distances between observations in one cluster to another cluster. For an observation with missing values in the later relative years, the average linkage will pose heavier weights toward earlier non-missing values when calculating the pairwise distances that involve this observation. In other words, students who drop out at a similar time period will be grouped together, depending on values in the earlier periods, adding an additional time feature to the clustering algorithm (Bowers, 2010).

A second decision the user needs to specify is the stopping criterion. HCA algorithm continues to cluster until all observations are grouped into a single cluster. The recommended stopping criterion is to stop when within cluster observations are very similar and across cluster are sufficiently different. In my study, a good number of clusters should also have policy meaning, which is not too small or big to have no policy implication. I show my determination of the number of clusters through visualizing on the dendrogram.

One way to visualize the hierarchical structure of HCA is using a heatmap (Eisen et al., 1998; Weinstein et al., 1997). A heatmap represents student as rows and enrollment variables as columns, where the ordering of rows or/and columns aligns with cluster assignments. This paper will use both heatmap and dendrogram to visually represent HCA findings.

### Features(variables) used for clustering

In this paper, I use enrollment history, which is re-coded in relative years since the entry term ranging from one to five years. The longitudinal features used in this clustering analysis are term GPA, term credits attempted, term college-level credits earned, and the yearly number of attended semesters from one to five relative years since entry. In order to reduce correlation across variables, I use term variables instead of cumulative variables.

### Pre-processing

In order to align entering semester and its relative terms across cohorts, I restrict my sample to only fall entry cohorts.[5] The final dropout sample has 40,559 observations. For implementation, all variables are standardized as recommended in the literature to prevent over-weightening the similarity matrix (Romesburg, 1984; Xu, 2008).

[5] For an example, fall entrants' second semester is a winter semester and is not comparable to winter entrants' second semester, which is a summer semester.

### Missing Data

It is not surprising to have missing data for enrollment variables among the dropout sample. Since the definition of dropout is determined at their relative sixth year, students who drop out prior to their sixth year will have missing values for all of the consequent terms after departure. For an example, a student who leaves college after their first year will have missing values for term GPA, credits attempted, and college-level credits earned variables in their second, third, fourth, and fifth year.

Fortunately, HCA can address missing values through the average linkage, which puts more weight on non-missing variables for a student with missing value.

## B Clustering Result

### Hierarchal Clustering Analysis

The goal of the HCA analysis is to identify if any, a number of distinctive groups among students who drop out from community colleges. The historical enrollment variables used for HCA are: total credits attempted, college-level credits earned, and GPA in each term and yearly variables indicating the number of semesters attended in the past academic year. Figure 2.2 presents a dendrogram for a visual representation of the HCA result. The entire hierarchical organization of the HCA is summarized with a vertical line representing an individual cluster. The horizontal line indicates merging of two clusters at that iterative step. The height in the left sidebar of Figure 2.2 illustrates inter-cluster distance at the time of merging. At each iterative step (height), one can count the number of disjoint clusters by drawing a horizontal line and counting the number of vertical lines that intercept. This step is equivalent to stopping the HCA algorithm at a point when intra-cluster dissimilarity reaches a certain threshold (height level). It is recommended in the literature to select a height where there is a large height gap from the previous iteration (Hastie, Tibshirani, & Friedman, 2009). In Figure 2.2, the red horizontal line indicates my decision to cut the tree at 4 clusters, which illustrates a relatively large height gap.

Table 2.2 indicates the distribution of cluster assignment when cutting my HCA tree at 4. As shown in Table 2.2, cluster one, two, and four are each distributed by about 20% of the dropout sample. Cluster 3 has the largest share of 40%.

Figure 2.3 is a full visual representation of the four cluster assignment from the HCA using both heatmap and dendrogram. Columns are aligned left to right in sequential order, from earliest to latest relative term to entry. The x-axis is a list of variables that are ordered by college-level credits earned, total credits attempted, and total GPA by each term. At the end of each year, a total number of attended semesters within that year is listed. A dendrogram is visible on the left sidebar with cluster numbers that align with Table 2.2.[6] Each row in Figure 2.3 represents a student in the sample. The center of the heatmap displays 4-levels of standardized z-scores of each variable, where the blue indicates high quartiles, the red indicates low quartiles, and the gray indicates missing values. Figure 2.3 shows that summer enrollments are generally low. If any, the high-credit medium dropout group has slightly more summer enrollments compared to other clusters but still has a low enrollment rate.

For an easier visual representation, I remove all summer terms in Figure 2.4. From Figure 2.4, it is clearer to see distinct characteristics of the four clusters in the dropout population: Late Dropouts, Trials, Low-credit Medium Dropouts, and High-Credit Medium Dropouts. The first distinctive pattern is a time dimension, which indicates when students leave college (gray). Cluster 2, which is labeled as ”late dropouts”, are students who stay long (after three to four years) before making the decision to leave college. This late dropout group has been identified in Mable & Britton’s (2018) paper. Among my 2002 – 2004 cohorts, around 20% of students are late dropouts, which is a slightly smaller proportion than Mabel & Britton (2018), which was one-third of the dropout sample. One possible reason for the different proportion is that Mable & Britton’s (2018) sample includes both two-year

[6] Note, the dendrogram from Figure 2.2 is equivalent to the left-sidebar of heatmap in Figure 2.3. However, Figure 2.3 heatmap rearranges rows to order students by clustering groups.

and non-selective four-year institutions while my study is strictly restricted to community college students. Cluster 3, which is labeled as ”trials”, is the largest share of my dropout population (40%) and are students who decide to leave college immediately after their first semester. Cluster 1, which is labeled as ”Low-credit medium dropouts” and Cluster 4, which is labeled as ”High-credit medium dropouts”, are groups of students who decide to leave college after one or two years. A second distinguishing feature across the four clusters is the color of the heatmap body. The late dropouts have mixed academic performances, while the trial dropouts have dominantly red colors that represent low-credits and low GPA students. Color distinction is particularly clear between cluster 1 and 4, where I find that the low-credit medium dropout group is dominantly red for variables total credits attempted and college-level credits earned. In contrast, high-credit medium dropout group is dominantly blue for those variables. [7]

### Characteristics of HCA clusters

I explore further with other variables that may provide additional information on the differentiating characteristics of the four clusters. Table 2.3 – 2.5 present averages on demographic characteristics, academic outcomes, and financial aid by HCA cluster membership.

Table 2.3 shows that the four clusters have a few distinct demographic characteristics. The low-credit medium dropouts are older, less white, more high-school graduates, and are more likely to be working at the time of entry. The late dropouts are largely young females, across all race, and are in-state students. The trial group has a relatively large male population, slightly older, less white, and less in-state students. The high-credit medium dropouts are more white, less black, and more likely to be receiving TANF.

Table 2.4 reveals interesting enrollment patterns by cluster assignment. The low-credit medium dropouts have lower credits however, not so low average GPA. One possible expla-

[7] Appendix Figure A2.4 presents the heatmap that re-aligns observations by both row and column clusters. This figure shows that the number of attended semester, credits attempted, and college-level credits are clearly grouped to have distinctive patterns for each cluster.

nation for this is that a not-so-low GPA may be because students attempted fewer credits in the first place. One outstanding pattern for the late dropouts is that students in this group resemble closely to remedial students, who have high remedial credits and low college-level credits earned in their first year. Over time, the late dropouts catch up with total credits attempted/earned and total GPA to behave similarly with completers. The trial group attempts fewer credits and have more withdraws than any other groups. The high-credit medium dropouts are very similar to completers. If any, these students earn more credits in the first two years. Lastly, Table 2.5 describes financial aid characteristics by clusters. The late dropout cluster and the high-credit medium dropout cluster receive relatively more financial aid than other groups.

To summarize, HCA has identified four typologies of community college dropouts: trials, high-credit medium dropout, low-credit medium dropout, and late dropout. By comparing descriptive characteristics of each clusters, I find that the trials are largely male, older, attempting less credits, and withdrawing a lot of classes in their first semester; the low-credit medium dropouts are older, more people of color, are more likely to be working at the time of entry; the high-credit medium dropouts resemble closely to completers and if any difference, attempts too much in their first year; the late dropouts have high proportion of young females, in-state students, resembles a lot like remedial students, and are receiving high financial aid.

## VI Classification

One of the drawbacks of cluster analysis is that it does not reveal any underlying relationships between features and cluster class. Given that the two groups, the high-credit medium dropout and the late dropout, have similar descriptive characteristics with completers, it seems particularly interesting to identify what factors contribute to completion. In order to explore this relationship, I need to first find a matching group of students who complete and have similar enrollment pattern that represents each cluster. With a matched set, I can build a prediction model that will identify which variables contribute most to completion and whether or not these important variables are different by clusters.

## A Matching dropout clusters to completers

In order to build a prediction model, my first step is to find a matching completer (non-dropout) group to each dropout cluster. Note that these completing students were initially excluded when running the HCA cluster analysis. The goal of matching is to generate causal hypothesis for my predictive analysis. That is, among students who have similar enrollment patterns, what variables contribute most to having a higher probability of completion?[8] In order to create a matched completers by dropout clusters, my initial step is to identify the medoid point of each dropout cluster; that is, the medoid point will be a representative point of each cluster. I use a distance matrix, uncentered Pearson correlation, to calculate the four medoid points, one for each cluster. The concept of medoid point is adopted from the K-medoid algorithm, which assigns a point as the representative center of a cluster and use that point to assign clusters to the remaining observations at each iteration. A medoid point, therefore, has the minimum average pairwise dissimilarity with observations within that cluster.

The distance matrix is chosen arbitrarily by the author, however, is used to align with the distance matrix used in the previous HCA clustering. After I identify the four medoid points of each cluster, I adopt from the k-medoids algorithm by assigning cluster number to completer students with the smallest distance to its medoid points. Similarly, the minimum distance is calculated using the un-centered person correlation distance matrix. [9] In order

[8] The proposed idea is analogous to the idea of matching in social sciences.

[9] Note that the choice of distance matrix is arbitrary. In Appendix Figure A2.5, I show the matching results when using the Euclidean distance matrix instead of the un-centered correlation matrix. Both matching results are similar except that the Euclidean distance tries harder to match the distribution of cluster assignment with HCA dropout clusters. For an example, the Euclidean distance matching shows a greater proportion of trials than that from the matching results from the un-centered correlation matrix. There are

to avoid assigning a cluster membership to a completer student who has a very different enrollment pattern from any of the dropout groups, I restrict my matching to a distance with a caliper of 2 standard deviations away from each cluster mean. A completer student who has bigger than 2 standard deviation distance away from the cluster mean will be dropped from the sample. Among total sample of 37,937 completing students, 36,581 students are assigned to a cluster.

## B Classification Methodology

With this matched completers and dropouts groups, three classification models are used to build a prediction model on completion, by clusters. In particular, I use logistic regression with elastic net regularization, random forest, and gradient boosting classification algorithms to build a prediction model. Tree-based ensemble methods, such as random forest and gradient boosting, are appropriate for this analysis as my data have missing values and mixed type of data (Hastie, et al., 2009). Also, ensemble methods are preferred over a single weak classification tree because of the advantage of reducing variances. Up to date, Gradient Boosting and Random Forest are one of the top performing ensemble models that are frequently used in the machine learning applications (Olson, et al., 2017). I use {h2o}(2016) for running the three prediction models.

A wide set of variables such as demographics, financial aid, earnings, and other enrollment variables (e.g. remedial credits attempted and earned and math and science credits attempted) are used to build a prediction model. It is important to mention that I intentionally exclude enrollment variables used for HCA clustering from my classification models.[10]

two reasons why I chose the un-centered correlation matrix for my main analysis. First, it aligns well with my previous HCA analysis and taking correlation into account when estimating pairwise distance is more reasonable given that my variables are varying only over relative time. Secondly, euclidean effort to match the distribution between completing students and dropouts is not necessary given that the two populations are expected to have different enrollment patterns, in the aggregate. As one example, one would expect to see only a small fraction of students who leave college after their first semester among completers, while 40% of dropout population were trials who left college after their first semester.

[10] The proposed idea is analogous to excluding covariates that were used for matching the treatment and

My final sample includes total 77,140 students (40,559 dropouts and 36,581 completers) and 278 predictors.

When running the prediction models, I follow the common practice in the field to hold out a random set from the sample as a test set. In this study, I leave out 30% of the sample as a test set. In addition, to tune the hyperparameters, I use Cartesian grid search for logistic regression with elastic net and random grid search for the remaining two algorithms, random forest and gradient boosting. Further details on how I tuned my hyperparameters are explained below. Across the three algorithms, I consistently use Area Under The Curve (AUC) values as my main performance metric for both selecting the optimal set of hyperparameters and selecting the best model, as recommended in the literature (Bowers & Zhou, 2018).

One important dimension that contributed to distinguishing the four HCA clusters was longevity of enrollment. In order to account for differences in longevity, I use a different length of relative variables to include in each model. In particular, given the short enrollment availability for the trial group, I use only the first relative year variables to predict completion. For both the high-credits and the low-credits medium matched clusters, I use up to two relative years variables to predict completion. Lastly, for the matched late group, I use all variables up to five relative years to predict completion. As I vary the number of variables to include in the prediction models across clusters, I would expect to have different predictive powers by clusters. In particular, the group with more variables (e.g., late dropouts) would have higher predictive power than the model with fewer variables (e.g., matched trials), which is what I see in my analysis. However, as the goal of this analysis is not to compare performances across clusters but, to select the best predictive model for each cluster, I continue with my analysis with keeping in mind that I would expect differential performances across clusters.

control from regression analysis.

### Generalized Linear Model with Elastic Net Regularization

The simplest approach for prediction is to fit a linear model, logistic regression model. Given that I have a large set of predictors that can have strong correlations, I add elastic-net regularization to my logistic regression model (Zou & Hastie, 2005). The minimizing objective function is :

$$min_{\beta}\{\sum_{i=1}^{N}(-y_i\beta^T x_i + log(1+e^{\beta^T x_i})) + \lambda\sum_{j=1}^{p}\alpha\beta_j^2 + (1-\alpha)|\beta_j|\} \tag{2.3}$$

Elastic net regularization is particularly useful when using highly correlated data, as in most education data sets. By combining lasso ($L_1$) and ridge ($L_2$) regularizations, the elastic net adjusts by averaging the highly correlated features (lasso) first, and then, shrinking the coefficients of those averages to zero (ridge). Parameter $\lambda$ controls for the shrinkage level of the model, where a high value of $\lambda$ shrinks more and equate a large number of coefficients to zero. Parameter $\alpha$ controls for the balance between ridge and lasso regularization where, $\alpha = 0$ is equivalent to ridge only model, $\alpha = 1$ is equivalent to lasso only model, and $0 < \alpha < 1$ is equivalent to elastic net model.

For my analysis, I tune my hyper-parameter $\alpha$ by using a Cartesian grid search on a sequence of numbers between 0.1 and 0.9, by an interval of 0.1. The optimal $\alpha$ is selected with a model that has the largest AUC value. For $\lambda$, the h2o platform embeds a $\lambda$ search, which starts from a maximum lambda value and efficiently searches through by decreasing $\lambda$ at each iteration until the minimum level is reached. Appendix Table A2.2 shows the result from the grid search by varying alpha and its corresponding AUC values. The optimal $\alpha$ values are shown in bold.

### Random Forest

Random Forest was formally introduced by Brieman (2001) with the concept of bagging. As one of an ensemble method of Classification and Regression Trees (CART), the Random

Forest aggregates many single CART algorithms through bagging. The bagging algorithm is a way of taking averages of a bootstrap sample that is applied to a weak learning classifier, iteratively. By averaging many weak-learner classifiers, the bagging algorithm reduces the overall variance. Random forest improves from a general bagging algorithm by using a random subset of features at each terminal-node of a tree, which decreases the correlation between trees and reduces average variance. By decreasing correlation between trees, a random forest can reduce variance while, maintaining un-biasedness of a single tree. Random Forest algorithm is specified as follows:

For iteration $i = 1, .., B$ (bootstrap sample = B)

1. Take a random bootstrap sample (randomly selected observations with replacement).
2. Using bootstrap sample to grow a random forest tree Ti until minimum node size is reached following,
    (a) Select random m variables from p predictors'
    (b) Select best split-point of a variable among the selected m variables
    (c) At that split-point, make two child nodes

Finally, the average of T1, T2,...TB random forest trees for an observation x makes the outcome prediction (either numeric or class). A splitting point is evaluated differently depending on whether the outcome variable is continuous or categorical. In my example with a categorical outcome, the best-split point is measured by taking the majority vote.

Random Forest has another advantage of only needing to tune a few hyperparameters. For implementing the Random Forest, I follow closely from the Breiman & Cutler's website.[11] and Hastie, et al.(2009). I use {h2o.randomforest} for implementation (Wright & Ziegler, 2017).

The hyperparameters that I tune for Random forest are the number of trees to grow (ntrees), the number of random variables to sample (m), and the terminal node size. Given

[11] Accessed from: `https://www.stat.berkeley.edu/~breiman/RandomForests/`

a large possible combination of parameters, I use random grid search from the h2o platform. Random grid search improves efficiency from a Cartesian search by at each iteration, randomly selecting a set of hyperparameters from a user-specified hyperparameter space. My stopping criterion is set to maximum running time as 40 minutes or if AUC value doesn't improve in the consequent 5 rounds by at least a tolerance of 0.00001.

- Number of trees to grow (ntrees): A large number of trees allows the model to stabilize its error. On the other hand, growing too many trees can be inefficient since it takes a greater computing time. I set my hyperparameter space for the number of trees as a sequence of numbers between 1 to 1000 by an interval of 5.

- Number of variables to Sample (mtries): Number of variables to sample is an important hyperparameter that controls error rates. Small m reduces correlation between any two trees in the forest (which reduces the error rate) but also, reduces the strength of each tree (which increases the error rate) (Breiman & Cutler website). In addition, the proportion of relevant variable and m can be important for improving prediction power for a model with large number of variables. Hastie, et al. (2009, pg 596) shows that random forest can perform poorly for a small fraction of relevant variable and small m. The default value for is $sqrt(p)$. I allow mtries hyperparameter space to have a sequence of numbers between 1 to 25, by an interval of 5.

- Maximum Depth of tree: Large depth of tree controls for the depth and complexity of the trees. While a deep tree can improve prediction accuracy, however, too deep of a tree can overfit to the training data. I allow maximum depth hyperparameter space to vary between a sequence of 10 and 30, by an interval of 5.

- Terminal node size: A terminal node size is another way to control for how complex your trees will be. A terminal node size controls for minimum number allowed at the terminal node, which means that the smaller the node size, the deeper the trees will be. My hyperparameter space for the node size is a sequence of 1 to 30, by an interval of 1.

### Gradient Boosting

Gradient Boosting algorithm is originated from Ada boosting ensemble method (Friedman, 2001). A boosting algorithm, similarly as bagging, is an ensemble method that aggregates many weak classifiers. However, the boosting algorithm differs from bagging in that the algorithm builds up in sequential order by updating the weights from the previous iteration step. In particular, at each boosting step, the weights on observations are updated by computing the negative gradient of a loss function, which gives a higher weight to a misclassified observation and a lower weight to a correctly classified observation. The final class decision is made on a weighted majority vote of boosting step classifiers with higher weights on the more accurate classifier. Decision trees are ideal to use as a base classifier for gradient boosting. For implementation, I follow closely to the recommendation from Hastie, et al. (2009) and Clike, et al. (2018) for using {h2o.gbm} feature.

The hyperparameters I tune for gradient boosting algorithm are the number of iterations (ntrees: Number of trees), the size of each tree (*max_depth*: Maximum depth of a tree), the learning rate, and the fraction of the sample. It is quite common to apply shrinkage techniques on a gradient boosting algorithm to control for prediction risk in training data. Similarly to randomly forest, I use a random grid search with the same stopping rule of maximum running time at 40 minutes and maximum tolerance of 0.00001 improvements in AUC for the consequent five rounds.

- Number of Trees (M): A Large number of boosting iteration reduces the training risk. However, too big of a number of trees can result in overfitting to the training data. I allow hyperparameter space to be a sequence of a number between 1 and 1500 by 5.

- Size of Each Tree: Size of each tree controls for interaction effects. For $J = 2$, the model will only consider main effects and for $J = 3$ the model will also include two-way interaction effects. Hastie, et al.(2009) suggests $4 \leq J \leq 8$ as ideal range for this parameter. Consequently, I allow parameter space to be a sequence of numbers between 4 and 8, by an interval of 1.

- Learning Rate: Learning rate controls for shrinkage in the model. A smaller learning rate will shrink more and thus, increase the training risk for a given number M. Hastie, et al.(2009) suggest for setting a small learning rate ($v < 0.1$). I allow this hyperparameter space to have a sequence number between 0.01 and 0.1, by interval 0.1.
- Fraction of subsample ($\eta$): A typical choice for $\eta = 1/2$. However, for large N, Hastie, et al. (2009) suggest a smaller choice of this parameter. My parameter space for $\eta$ is 0.1, 0.25, and 0.5.

## C Classification Results

### Matching Results: Assigning cluster groups to completers

Figure 2.5 illustrates heatmap visual representation of the matched clusters for completer students. Table 2.6 shows the distribution of cluster assignments for students who do not drop out from community colleges. Both Figure 2.5 and Table 2.6 illustrates that proportions of the matched trial group and the matched low-credit medium group are smaller than that of HCA dropout groups (about 27% compared to 42% and about 18% compared to 20%, respectively). Also, the proportion of the matched late group and the matched high-credit medium group are bigger than that from the HCA groups (about 23% compared to 18% and about 33% compared to 21%, respectively). Given that matching is among students who do not drop out, it is not surprising to see a different share of students in each group that resembles dropout enrollment patterns. In particular, I find a smaller proportion of completer students with similar enrollment patterns with the dropout trials and the low-credit medium dropout group.

### 10-fold Cross Validation Performances

Using the matched groups, I run three prediction models on a random 70% of the sample as my training set. To tune my hyperparameters, I split this training set into 10-fold to calculate the averages of the 10-fold cross validation performance metrics for each set of hyperparameters. The best set of hyperparameters are selected with the model with the largest AUC value. Table 2.7 presents averages and standard errors of the 10-fold cross validations performances. The first two columns indicate performance measures among the matched low-credit medium group, columns three and four indicate performances among the late dropouts and its matched completers, columns five and six indicate performances among the trial dropouts and its matched students, and final two columns indicate performances among the high-credit medium dropout and its matched completed students. Top to bottom panel shows performance measures of the three prediction models: logistic regression with elastic net, random forest, and gradient boosting. Each model is illustrated using five performance metrics; accuracy, which is the proportion of true positive and negative predictions over all possible predictions made; AUC value, which is the area under the Receiver Operating Characteristics (ROC) curve; precision, which is the true positive prediction overall positive predictive values; recall, which is the true positive portion overall true positive cases; and specificity, which is the true negative portion of the all negative cases in the data.

Across the three models, the trial group has the lowest performance measures with ranging between .677 to .719. This is not surprising given that I am only using one relative year variables to predict completion, as mentioned before. The low-credit medium matched group is the second lowest with AUC values ranging between .748 to .759 and is the group with the smallest difference across models. On the other hand, the matched high-credit medium and the matched late groups have large AUC values across all of the three models that are ranging between .815 to .849 and .829 to .868, respectively.

Gradient boosting model is by far the best performing model across the four clusters. Random forest, generally, is the second best and the logistic model with elastic net performing the worst. One exception where the logistic model with elastic net outperforms random forest is among the matched late trials. Overall, ensemble methods slightly improved performance compared to the generalized linear model by AUC values of around .011 to .042. Comparing to the previous findings from Knowles (2015), my findings are consistent in that gradient boosting out-performs random forest and the generalized linear model. In terms of AUC value, my AUC values are generally lower than Knowles (2015) except for the prediction on completion among the matched late dropout group, which had an AUC value of .868. One possible reason for a generally lower prediction power may be because Knowles (2015) uses high school data to create an early warning system while this paper uses college-level data to predict completion that is determined in the sixth year. It is worth mentioning that despite my AUC values differ slightly from Knowles (2015), it generally falls under a wide range of performance measures evaluated by Knowles(2015) using different algorithms.

### Variable Importance

In order to better understand the relationship between variables and completion, I choose the best performing gradient boosting model (with the highest AUC values) and compute the relative variable importances. Variable importance is embedded in tree-based models and is calculated using the number of times a variable is selected at each splitting node. A variable is selected at each node when that variable reduces the maximum amount of squared error risk. Variable importance in a single tree is measured by taking the sum of this reduced squared error risk every time that variable is selected as a node. In an ensemble method, relative variable importance is calculated by averaging the importance measures for that variable obtained from a single tree (Hastie, et al., 2009). It is worth mentioning that important variables have no causal interpretation of the relationship between variables and outcome. The variables are measured in relative values with 100 being the most important variable and then scaling down.

Figures 2.6–2.9 show top 20 variables with the highest values in relative variable importance. Figure 2.6 present important variables for the matched trials group, Figure 2.7 for the matched low-credit medium group, Figure 2.8 for the matched high-credit medium group, and Figure 2.9 for the matched late group. As shown in Figure 2.6, the top three most important variables for explaining completion among the matched trial group are net tuition in their first year and age at entry. The top three most important variables for explaining completion among the matched low-credit dropout group are % of credits withdrawn in the spring of the 2nd year, age at entry, and remedial credits attempted in the spring term of the 2nd year (see Figure 2.7). For the matched high-credit dropout group, the most important variables are math and science cumulative credits earned by spring of 2nd and 1st year and % of credits withdrawn in the spring of the 2nd year (see Figure 2.8). This suggests that among students who attempt and earn low-credits during their first two years, students who withdraw less and attempts more remedial classes in their second year have a higher probability in completing college. On the contrary, among students who attempt and earn high-credits within their first two years, students with high cumulative credits earned in math and science have a high probability of completion. It is also worth noting that net tuition appears as the next important variable for predicting completion. As shown in Figure 2.9, the matched late dropout group have credits withdrawn, full-time status in their spring semester of third year, and remedial credits attempted in the fall of their third year as the most important three variables in predicting completion.

### Out-of-sample prediction and ROC curves

Up til now, my performance measures were based on the predictive model using 10-fold cross-validation from the training sample (70% of the entire sample). Table 2.8 present performance measures (AUC value) for out-of-sample prediction model (the remaining 30% of the sample) using the three models: logistic regression and elastic net, random forest, and gradient boosting. The hyperparameters are obtained from the best performing model from the previous 10-fold cross-validation models. Reassuringly, the out-of-sample AUC values are very similar to 10-fold cross-validation performances suggesting that the models are not overfitting to training dataset. Figures 2.10 – 2.13 draw ROC curves for the three classification methods using out-of-sample predictions. The x-axis indicate false positive rates and the y-axis indicate true positive rates. ROC curves for the trial matched group (see Figure 2.10) indicates that gradient boosting and random forest algorithms outperform logistic regression with elastic net algorithm. However, for the matched low-credit medium group (see Figure 2.11), the three predictive models are not very distinguishable. This is consistent with my prior findings from the 10-fold cross-validation. On the other hand, Figure 2.12 indicates that there is a clear performance difference for the matched high-credit medium group with gradient boosting model being the best performing model. Figure 2.13 is interesting as the performance for the logistic regression model with elastic net is better than that for the random forest. Nonetheless, gradient boosting algorithm performed best for this group, as well. From all four matched clusters, gradient boosting outperformed any other prediction models even with the 30% leave-out sample. This is consistent with my previous findings from the10-fold cross-validation.

# VII Conclusion

This paper identified four typologies of community college dropout using hierarchical cluster analysis. The four dropout groups are trials (dropout after 1st semester), high-credit medium dropouts, low-credit medium dropouts, and late dropouts. Among my dropout sample, around 20% are each distributed to the low-credit medium, high-credit medium, and late clusters. The remaining 40% of the sample is assigned to the trial group.

Exploring demographic characteristics of each cluster, I find that the trial and the low-credit medium clusters have more males, older population, and fewer whites. On the other hand, the high-credit medium cluster has more whites and the late cluster has more young females across all race. In terms of academic achievement, the trials and the low-credit medium clusters attempt and earn fewer credits. Despite low credits earned, the low-credit medium dropout cluster does not necessarily have low GPAs. The high-credit medium cluster aligns very closely with students who complete college and if any, attempts too much during the first year. The late dropouts characterize closely to remedial students who take a lot of remedial classes in their first year.

Given the distinctive characteristics of dropout clusters, I match students who complete and have similar enrollment history to dropout clusters. Within the matched clusters, I predict completion using three classification algorithms, logistic regression with elastic net, random forest, and gradient boosting. Across the three prediction models, I find that gradient boosting performs the best. The AUC values for both 10-fold cross-validation and out-of-sample using gradient boosting ranges between .719–.868 and .713–.867, respectively. Examining variable importance measures from the best gradient boosting model, I find that each of the four clusters has a different set of important variables that contribute largely to predicting completion. In particular, for the trial group, net tuition in their first semester contribute most to predicting completion. For the matched low-credit group, I find that credits withdrawn and age at the entry contributes most in predicting completion. For the matched high-credit group, I find that math and science cumulative credits earned by their second year explains most in predicting completion. Finally, for the matched late dropout group, credits withdrawn and full-time status in their third year explained most in predicting completion.

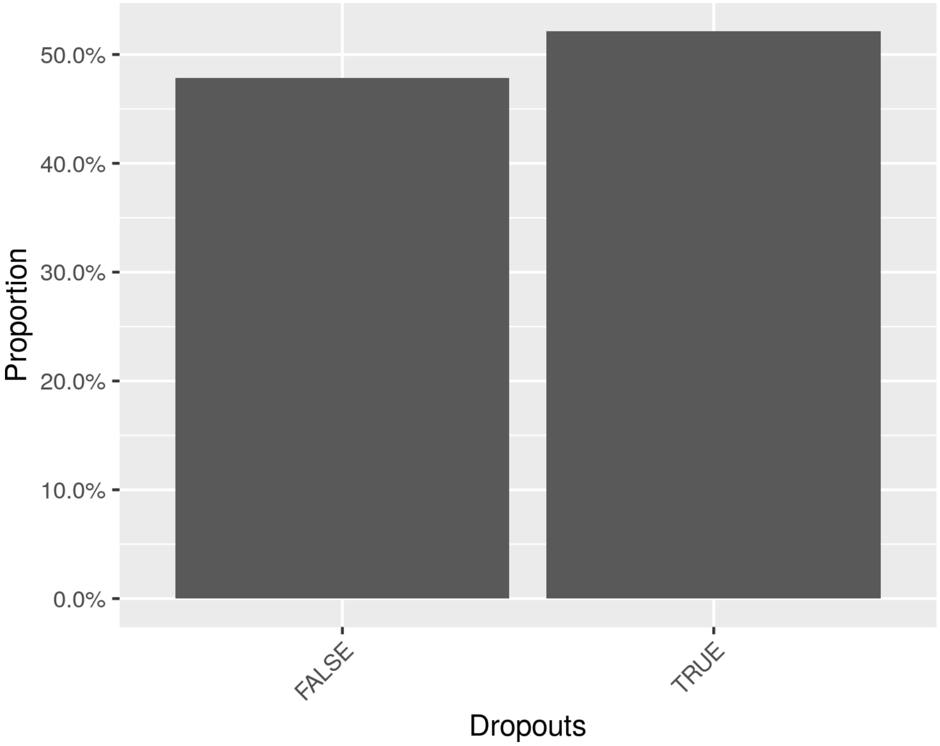
This paper provides suggestive evidence that categorizing the dropout population first, helps to better understand the needs for completion. First, the variable importances suggest that dropout groups (i.e., clusters) have a different set of variables that contribute largely to completion. This indirectly suggests that students who dropouts can have different needs for completion and grouping at-risk population by similar enrollment patterns could help identify the heterogeneous needs. Secondly, to compare the predictive power between grouping dropouts as a single group versus separating into four clusters, I compare AUC performances when using the same logistic regression with elastic net but, when treating dropout as one population. Similarly, to address longevity, I run five models with a different set of longevity variables; that is, using only first-year variables, one and second-year variables, and so on, up to including all five relative year variables. Table 2.9 shows that average AUC values of the 10-fold cross validation using dropout as one group has smaller prediction power than my AUC values using clustering and matching. The only exception is the matched trial group, which performed less than using one-year variables and treating dropout as a single group. For example, both high-credit and low-credit medium dropout groups have higher AUC (.815 and .748) than a prediction model that uses up to second relative year variables (.740). My late dropout prediction power is even far better than a prediction model that uses up to five relative year variables (0.846 versus 0.788). This comparison provides suggestive evidence that separating out the dropout population by four clusters and then building a prediction model can have higher prediction power.

It is worth mentioning the limitation of this paper. Although cluster analysis is useful in identifying the four groups of dropout, it is not clear how institutions can identify which group at-risk students are in prior to dropping out of college as one of the two metrics that identify groups is time of departure. This suggests for future work to develop a good predictive model that can identify dropout typologies using college first-year information.

To conclude, this paper identifies that there are four distinct groups in the dropout population at a large community college system within a single state. In specific, the four distinct groups are trials, high-credit medium dropouts, low-credit medium dropouts, and late dropouts. The findings of this paper will provide guidelines to community colleges in understanding the typologies of college dropouts. Institutions can use two dimensions of enrollment patterns to distinguish at-risk student types: relative time of entry and total credits attempted and earned by term. Furthermore, by matching students who complete to each of the four distinct dropout groups by using similarity in enrollment patterns, I find that different groups have different set of variables that are contributing most in predicting completion. Particularly, for late dropouts, I find that percentage of credits withdrawn and having full-time enrollment in the third year matters most in improving probability of completion. This suggests that providing additional support in student's third year can help in reducing the probability of dropping out among students who are enrolled beyond their second year. Second, I find that higher credits earned in math and science during the first two years contributes importantly in increasing the probability of completion among students who have relatively good grades and high credits in the first two years. This suggests that institutions should promote students in taking math and science courses early in the first two years even for those who may be doing relatively well. Lastly, the differential needs for completion for each dropout typology suggests that it is important for institutions to create standard practices in identifying at-risk dropout types and employ group-specific supports and interventions to help students achieve their completion goal.

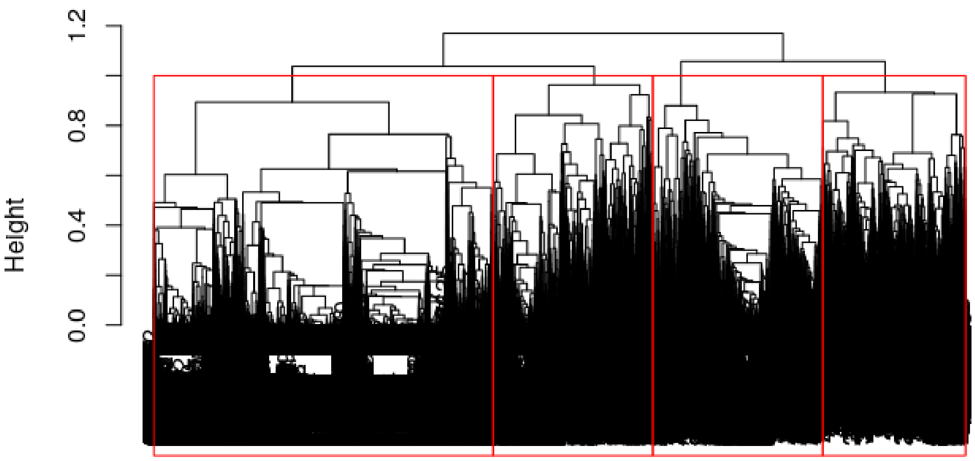
# VIII Figures

Figure 2.1: Dropouts Proportion (51.7%)



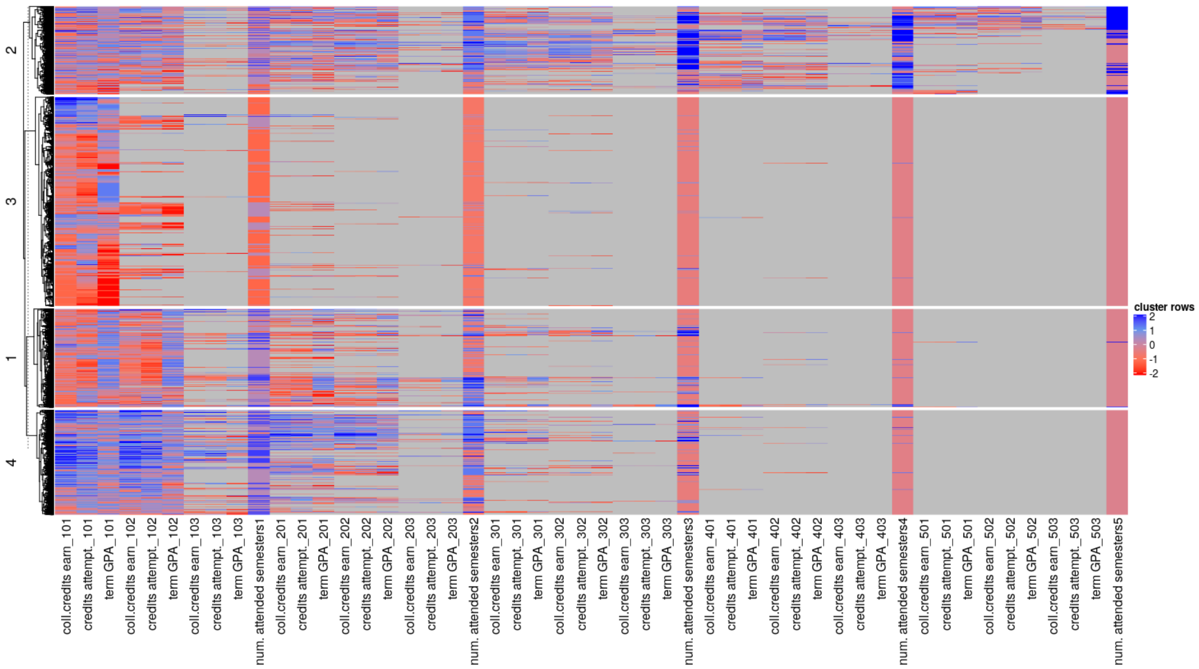
This figure shows the proportion of dropouts in the analysis sample, which are restricted to students in 2002 - 2004 fall entry cohorts. Dropouts is defined as students who do not receive a degree or certificate, did not transfer to a four-year institution, and is not enrolled in first semester of the sixth year of community college entry.

Figure 2.2: HCA Cluster Dendrogram



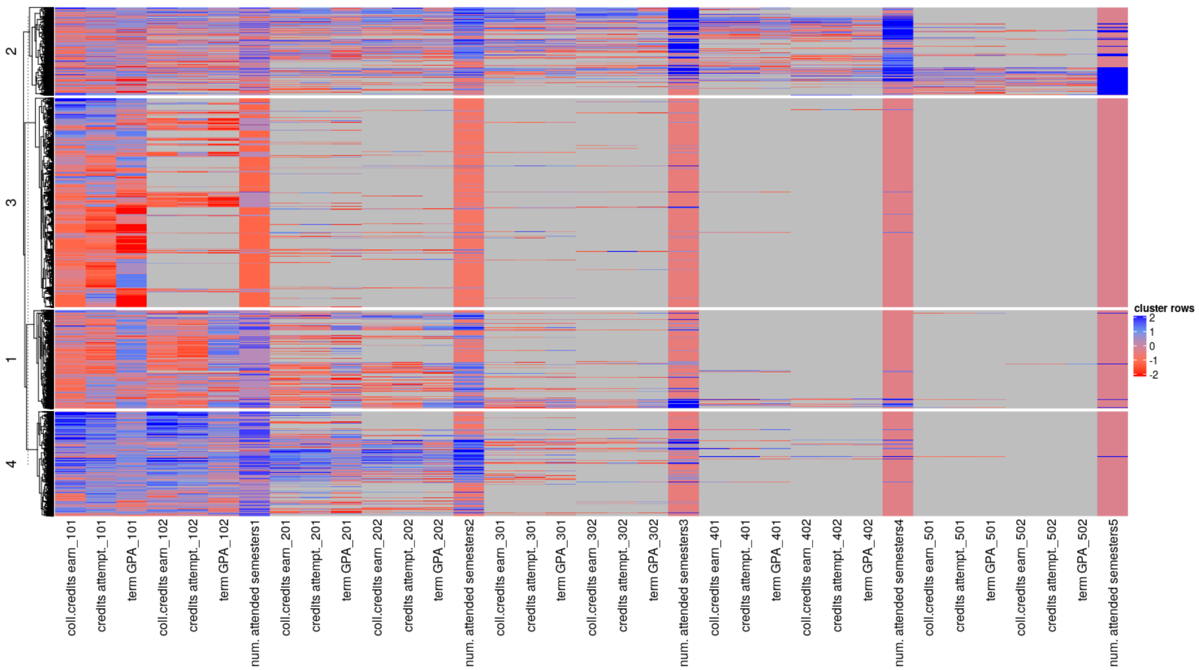
This figure shows dendrogram for my hierarchical clustering analysis. Each vertical line represent each cluster and vertical line show cluster merges. The height represent intra-cluster distance at the iteration of the merge. The red box indicates my cutoff tree at 4 clusters.

Figure 2.3: HCA Heatmap (Cluster Row Only)



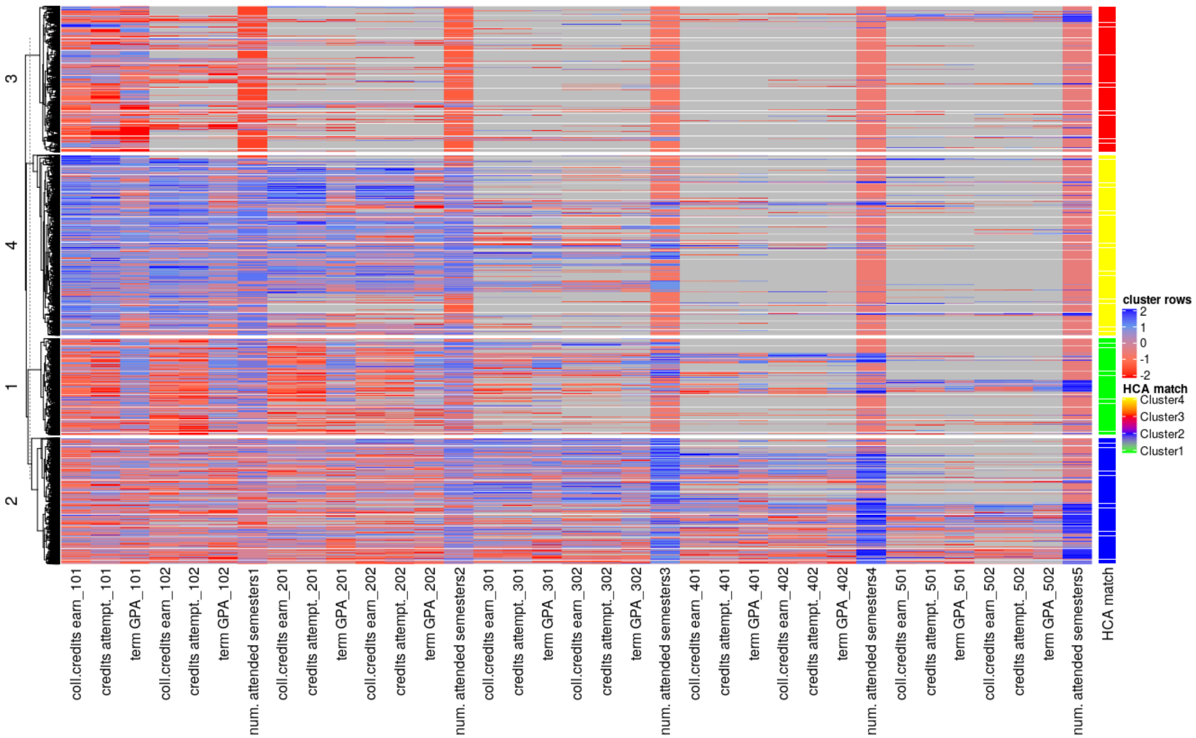
This figure is a visual representation of hierarchical clustering analysis result using a heatmap. Sample is restricted to dropouts among 2002-2004 fall entry cohorts. Each student represent rows. Each columns are listed in the order from left to right from year one to five: college-level credits earn (fall semester), total credits attempted (fall semester), term GPA (fall semester), college-level credits earn (spring semester), total credits attempted (spring semester), term GPA (spring semester), college-level credits earn (summer semester), total credits attempted (summer semester), term GPA (summer semester),and number of attended semesters (out of total three semesters). Column values are standardized and represented in four quartiles, where red to blue represent lowest to highest quartiles. The ordering is represented by row clusters only.

Figure 2.4: HCA Heatmap excluding Summer Semester (Cluster Row Only)



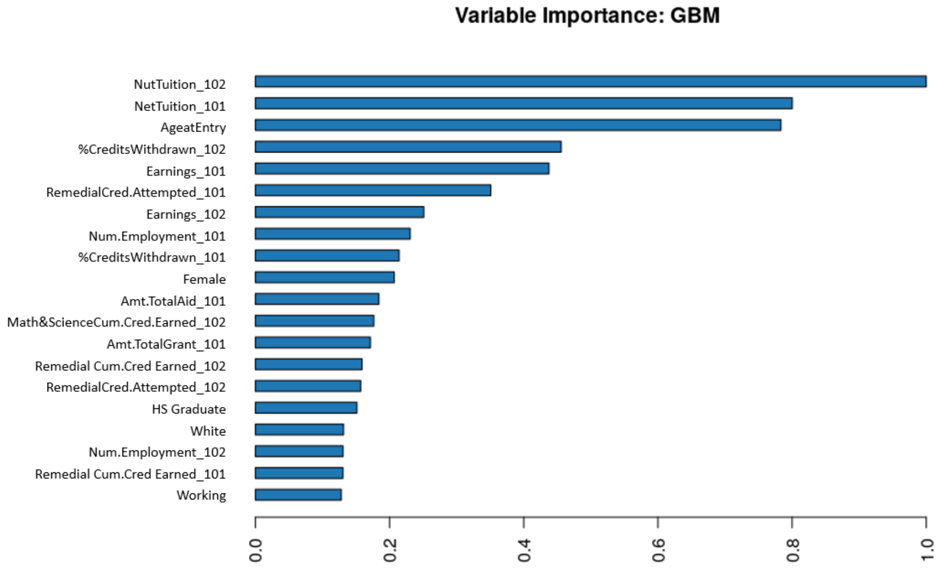
This figure is a visual representation of hierarchical clustering analysis result using a heatmap. Sample is restricted to dropouts among 2002-2004 fall entry cohorts. Each student represent rows. Each columns are listed in the order from left to right from year one to five: college-level credits earn (fall semester), total credits attempted (fall semester), term GPA (fall semester), college-level credits earn (spring semester), total credits attempted (spring semester), term GPA (spring semester),and number of attended semesters (out of total three semesters). Column values are standardized and represented in four quartiles, where red to blue represent lowest to highest quartiles. The ordering is represented by row clusters only.

Figure 2.5: Matched Completers



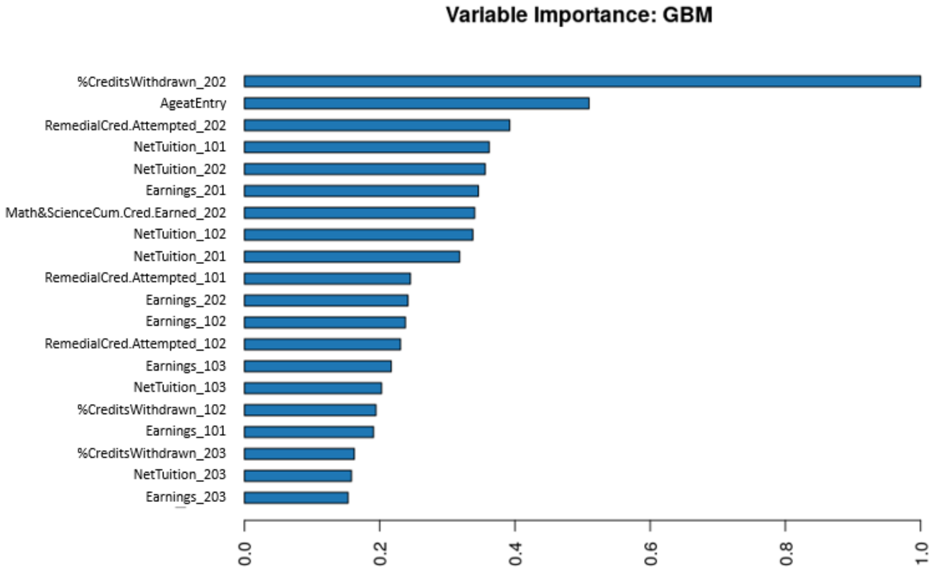
This figure is a heatmap representation of matched sample of completing students assigned to a cluster number with similar enrollment patterns. Sample is restricted to completers among 2002-2004 fall entry cohorts. Each student represent rows. Each columns are listed in the order from left to right from year one to five: college-level credits earn (fall semester), total credits attempted (fall semester), term GPA (fall semester), college-level credits earn (spring semester), total credits attempted (spring semester), term GPA (spring semester),and number of attended semesters (out of total three semesters). Column values are standardized and represented in four quartiles, where red to blue represent lowest to highest quartiles. The ordering is represented by row clusters only.

Figure 2.6: Variable Importance Plot for Trials



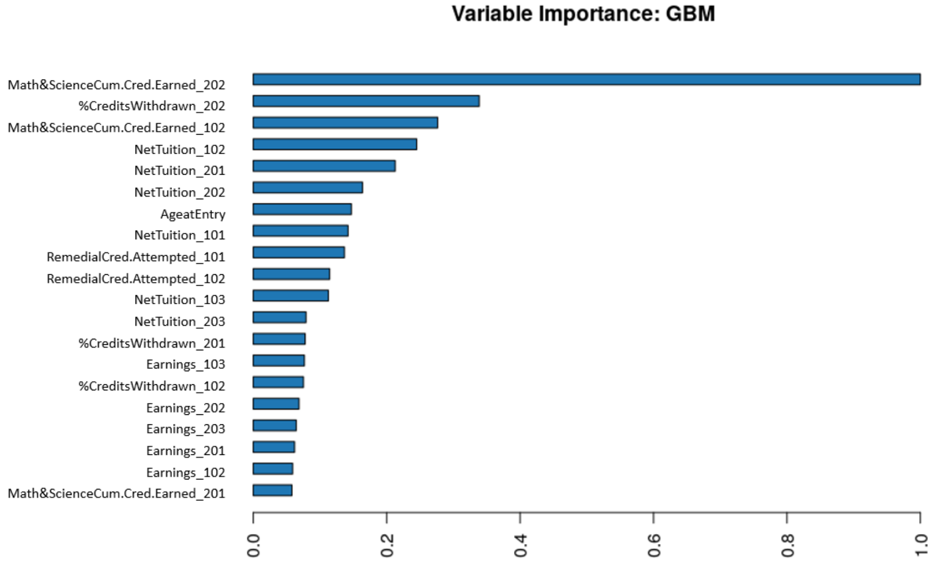
This figure show top 20 variable importance plot for prediction model on completion using gradient boosting algorithm among Trials dropout and matched students who complete. Variables are ranked from top to bottom by its relative influence on outcome. The x-axis indicate scaled relative importance that lies between zero and one.

Figure 2.7: Variable Importance Plot for Low-Credit Medium Dropouts



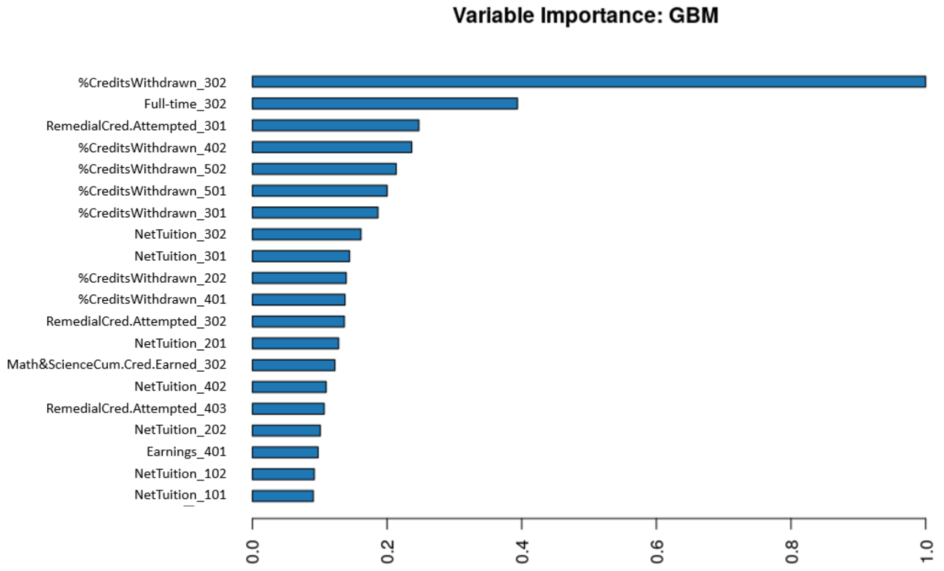
This figure show top 20 variable importance plot for prediction model on completion using gradient boosting algorithm among low-credit medium dropout and matched students who complete. Variables are ranked from top to bottom by its relative influence on outcome. The x-axis indicate scaled relative importance that lies between zero and one.

Figure 2.8: Variable Importance Plot for High-Credit Medium Dropouts (bottom)



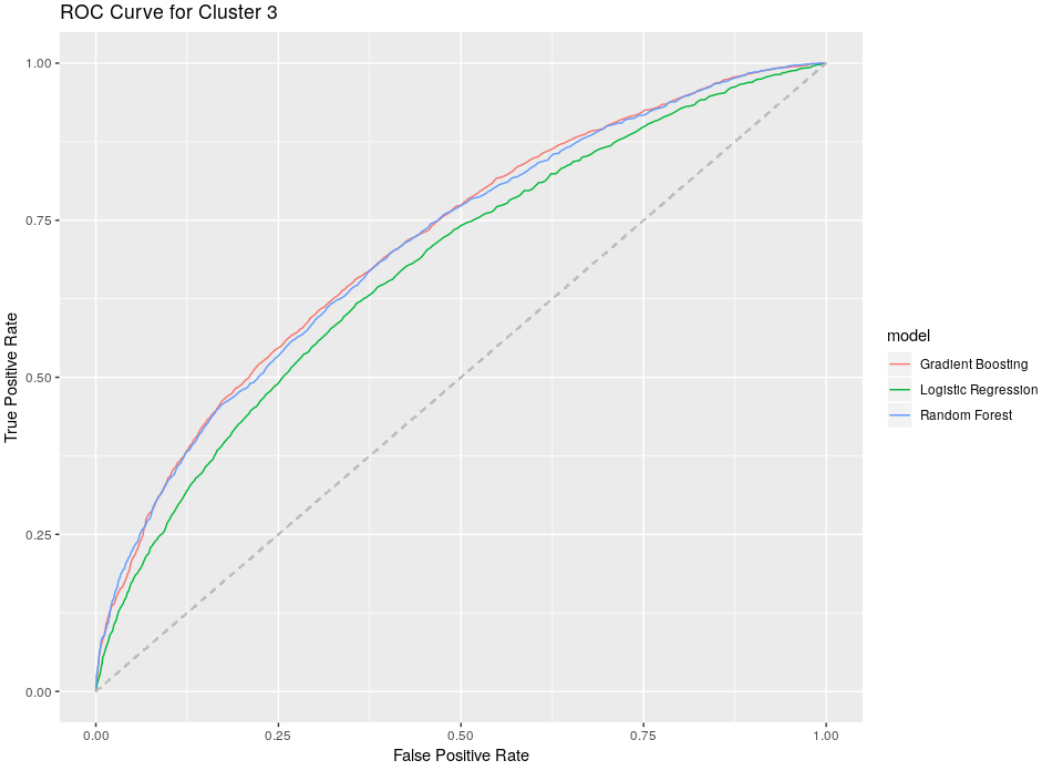
This figure show top 20 variable importance plot for prediction model on completion using gradient boosting algorithm among high-credit medium dropout and matched students who complete. Variables are ranked from top to bottom by its relative influence on outcome. The x-axis indicate scaled relative importance that lies between zero and one.

Figure 2.9: Variable Importance Plot for Late Dropouts



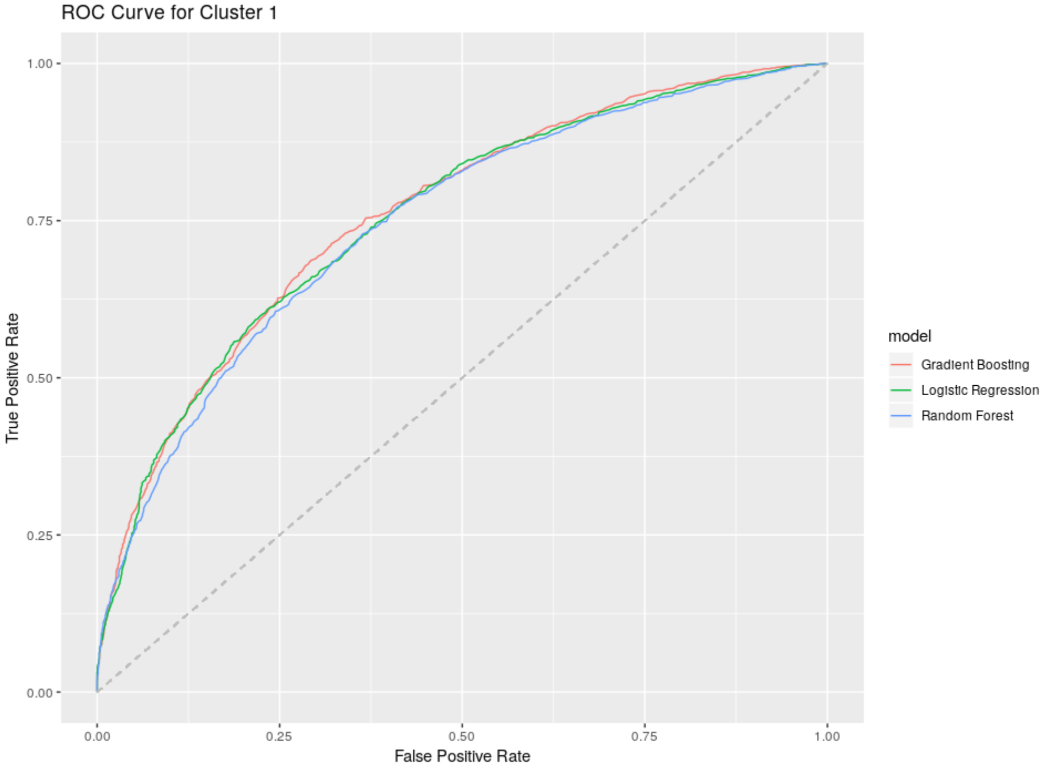
This figure show top 20 variable importance plot for prediction model on completion using gradient boosting algorithm among late dropout and matched students who complete. Variables are ranked from top to bottom by its relative influence on outcome. The x-axis indicate scaled relative importance that lies between zero and one.

Figure 2.10: ROC Curves for Trials



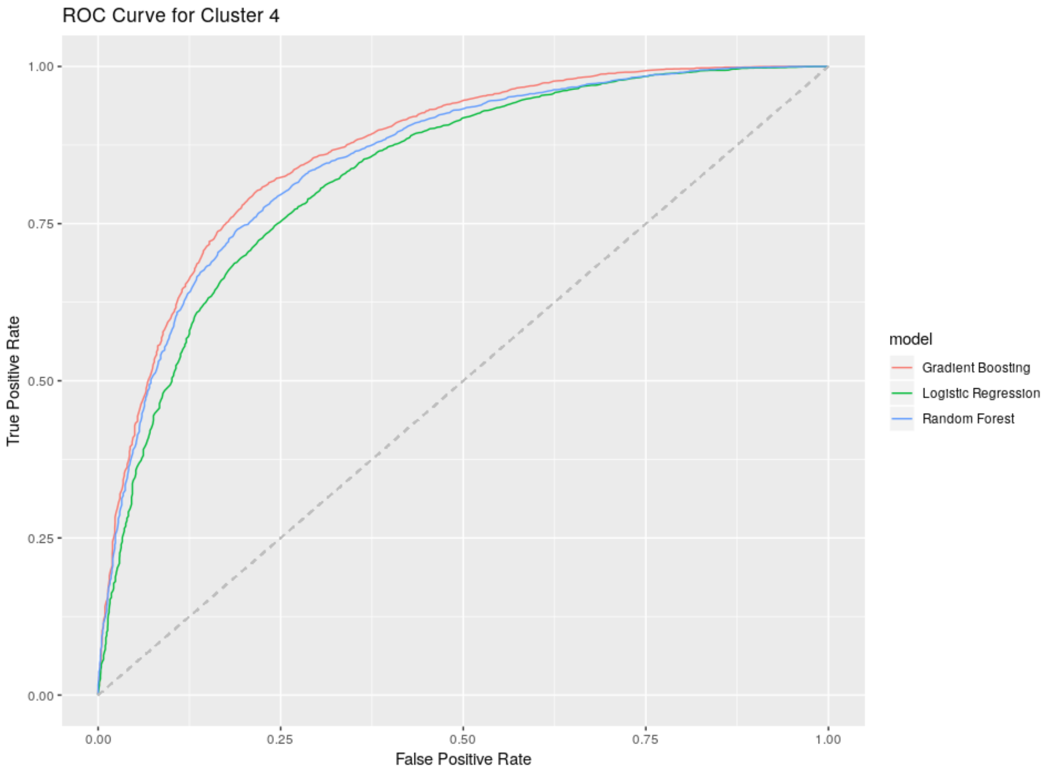
This figure show receive operating characteristic (ROC) curve for out-of-sample prediction models using logistics regression with elastic net (green), random forest (blue), and gradient boosting (red). Sample is restricted to trial dropout and matched students who complete, among 2002-2004 fall entry cohorts.

Figure 2.11: ROC Curves for Low-Credit Medium Dropouts



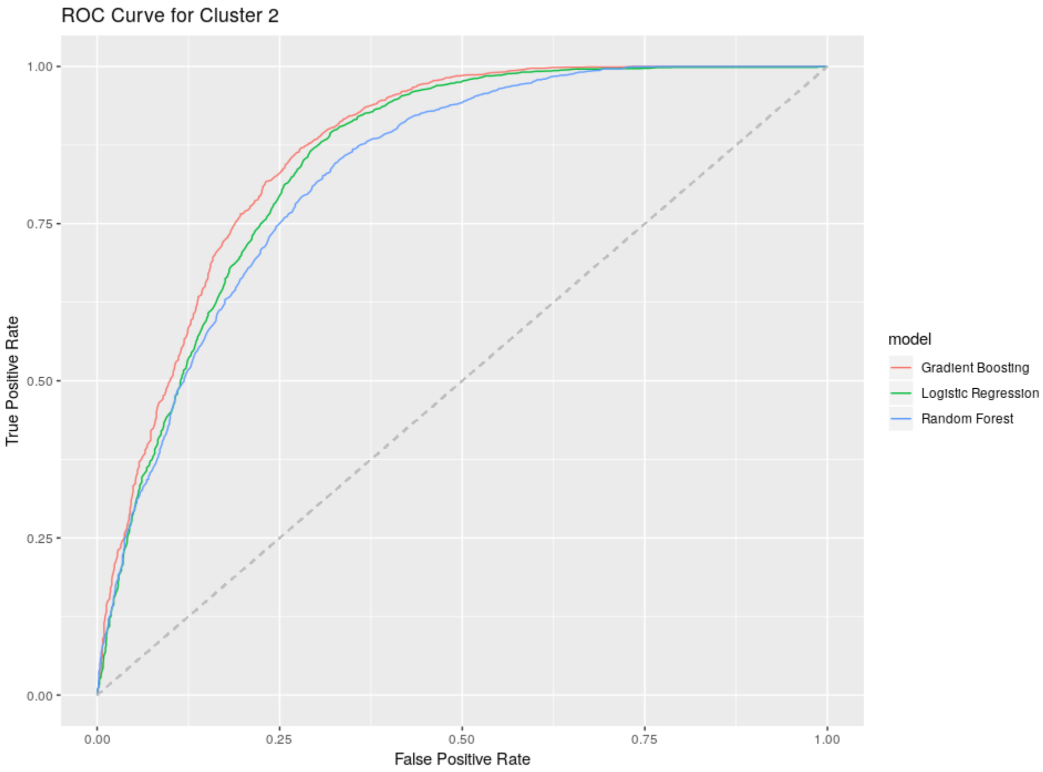
This figure show receive operating characteristic (ROC) curve for out-of-sample prediction models using logistics regression with elastic net (green), random forest (blue), and gradient boosting (red). Sample is restricted to low-credit medium dropout and matched students who complete, among 2002-2004 fall entry cohorts.

Figure 2.12: ROC Curves for High-Credit Medium Dropouts



This figure show receive operating characteristic (ROC) curve for out-of-sample prediction models using logistics regression with elastic net (green), random forest (blue), and gradient boosting (red). Sample is restricted to high-credit dropout and matched students who complete, among 2002-2004 fall entry cohorts.

Figure 2.13: ROC Curves for Late Dropouts



This figure show receive operating characteristic (ROC) curve for out-of-sample prediction models using logistics regression with elastic net (green), random forest (blue), and gradient boosting (red). Sample is restricted to late dropout and matched students who complete, among 2002-2004 fall entry cohorts.

# IX Tables

Table 2.1: Sample Characteristics of 2002-2004 Cohort by Dropout Indicator

| Variable | Sample Completers | Sample Dropouts | National Avg. Completers | National Avg. Dropouts |
|---|---|---|---|---|
| Female | 0.584 | 0.543 | 0.574 | 0.572 |
| Age at entry (years) | 23.492 | 25.184 | 22.319 | 25.911 |
| Race | | | | |
| White | 0.707 | 0.676 | 0.641 | 0.578 |
| Black | 0.206 | 0.236 | 0.128 | 0.148 |
| Hispanic | 0.03 | 0.031 | 0.195 | 0.135 |
| Other Race | 0.057 | 0.058 | 0.09 | 0.078 |
| Citizenship | 0.986 | 0.988 | 0.920 | 0.937 |
| High School Graduates (%) | 0.884 | 0.895 | 0.983 | 0.993 |
| Inmate | 0.023 | 0.019 | N/A | N/A |
| Instate | 0.907 | 0.932 | N/A | N/A |
| Limited English Proficiency | 0.004 | 0.004 | N/A | N/A |
| Has prior enrollment | 0.116 | 0.084 | 0.213 | 0.143 |
| Received TANF | 0.075 | 0.091 | N/A | N/A |
| Working | 0.566 | 0.565 | 0.729 | 0.787 |
| Financial Aid (first term) | | | | |
| Received any aid | 0.364 | 0.385 | 0.544 | 0.530 |
| Amount total aid ($, Incl.0s) | 545.605 | 596.292 | 1776.676 | 1407.05 |
| Received any grant | 0.335 | 0.366 | 0.476 | 0.458 |
| Amount total grant ($, Incl.0s) | 485.291 | 546.705 | 1241.872 | 942.219 |
| Has any loan | 0.024 | 0.023 | 0.114 | 0.114 |
| Among total loan ($, Incl.0s) | 34.655 | 31.583 | 317.766 | 304.505 |
| Enrollment | | | | |
| Total missed semesters by year 3 | 4.428 | 6.022 | N/A | N/A |
| Cum. credits earned by year 3 | 60.2 | 47.8 | 51.872 | 23.719 |
| Cum. college-level credits earned by year 3 | 55.7 | 43.0 | N/A | N/A |
| Cumulative GPA by year 3 | 2.92 | 2.634 | 2.937 | 2.558 |
| Degree Attainment | | | | |
| Received certificate by year 6 | 0.159 | 0 | 0.115 | 0 |
| Received associates by year 6 | 0.413 | 0 | 0.259 | 0 |
| Received bachelors by year 6 | 0.136 | 0 | 0.227 | 0 |
| Ever transferred to 4-year by year 6 | 0.556 | 0 | 0.517 | 0 |
| Sample Size | 37,937 | 40,599 | 9,028 | 7,094 |

*Note.* Columns 1-2 are the analysis sample of 2002-2004 fall entry cohort students. Columns 3-4 show averages for the national representative BPS 2003/2004 sample, restricted to those whose first institution is at a public two-year college for the first time in academic year 2003-2004. Column 3 is averages of a sample who is enrolled or not enrolled and have AA, certificate, or transferred to a 4-year or students enrolled, no degree from variable, cumulative persistence and attainment anywhere in 2008-2009. Column 4 is averages of a sample who is not currently enrolled and do not have a degree using variable, cumulative persistence and attainment anywhere in 2008-2009.

Table 2.2: HCA Clustering Distribution

| HCA Clustering | Sample Size | % |
|---|---|---|
| 1 (Low-credit Medium Dropouts) | 7978 | 19.67 |
| 2 (Late Dropouts) | 7127 | 17.57 |
| 3 (Trials) | 16959 | 41.81 |
| 4 (High-credit Medium Dropouts) | 8495 | 20.94 |

*Note.* Samples are restricted to dropouts among 2002-2004 fall entry cohort students. This table shows distribution of cluster assignment from hierarchical clustering analysis. First column indicates cluster assignments, second column is sample size of each cluster, and third column indicates sample proportions of each clusters.

Table 2.3: HCA Demographic Characteristics

| Variables | Low-credit Medium Dropout | Late Dropouts | Trials | High-credit Medium Dropouts | Completers |
|---|---|---|---|---|---|
| Female | 0.604 | 0.632 | 0.484 | 0.531 | 0.584 |
| Age at entry | 27.492 | 22.333 | 25.529 | 24.717 | 23.492 |
| White | 0.673 | 0.681 | 0.650 | 0.724 | 0.707 |
| Black | 0.235 | 0.233 | 0.260 | 0.189 | 0.206 |
| Hispanic | 0.033 | 0.025 | 0.033 | 0.029 | 0.030 |
| Other Race | 0.059 | 0.061 | 0.056 | 0.057 | 0.057 |
| Citizenship | 0.984 | 0.987 | 0.990 | 0.990 | 0.986 |
| High School Graduates | 0.910 | 0.879 | 0.897 | 0.892 | 0.884 |
| Inmate | 0.012 | 0.006 | 0.027 | 0.021 | 0.023 |
| Instate | 0.932 | 0.953 | 0.916 | 0.943 | 0.907 |
| Limited English Proficiency | 0.004 | 0.003 | 0.004 | 0.005 | 0.004 |
| Has prior enrollment | 0.078 | 0.098 | 0.076 | 0.093 | 0.116 |
| Received TANF | 0.083 | 0.093 | 0.087 | 0.103 | 0.075 |
| Working | 0.621 | 0.581 | 0.561 | 0.506 | 0.566 |

*Note.* Columns 1–4 are restricted to dropouts among 2002-2004 fall entry cohorts. Column 1 present demographic averages among low-credit medium dropout cluster, column 2 is for late dropout cluster, column 3 is for trial cluster, and column 4 is for high-credit medium dropout cluster. Column 5 shows averages for completers among 2002-2004 fall entry cohorts.

Table 2.4: HCA Enrollment History

| | Low-Credit | Late | | High-Credit | |
|---|---|---|---|---|---|
| Variables | Medium Dropout | Dropouts | Trials | Medium Dropouts | Completers |
| total credits attempted | | | | | |
| year1,fall | 8.058 | 10.318 | 8.292 | 12.962 | 11.102 |
| year1,spring | 14.682 | 18.336 | 10.279 | 25.367 | 20.230 |
| end of year 1 | 15.680 | 19.333 | 10.346 | 28.362 | 22.160 |
| end of year2 | 22.551 | 32.319 | 11.171 | 41.161 | 37.718 |
| end of year3 | 24.599 | 44.406 | 11.525 | 43.251 | 46.680 |
| end of year4 | 25.003 | 52.706 | 11.624 | 43.586 | 51.942 |
| remedial credits attempted | | | | | |
| year1,fall | 2.039 | 2.411 | 1.138 | 1.652 | 1.515 |
| year1,spring | 1.077 | 1.184 | 0.169 | 0.808 | 0.704 |
| year2,fall | 0.401 | 0.448 | 0.042 | 0.192 | 0.252 |
| year2,spring | 0.181 | 0.244 | 0.028 | 0.092 | 0.134 |
| number of course withdraws | | | | | |
| year1,fall | 0.254 | 0.316 | 0.570 | 0.152 | 0.227 |
| year1,spring | 0.442 | 0.366 | 0.258 | 0.228 | 0.241 |
| year2,fall | 0.320 | 0.283 | 0.056 | 0.215 | 0.182 |
| year2,spring | 0.208 | 0.256 | 0.027 | 0.211 | 0.159 |
| cum. college-level credits earned | | | | | |
| year1,fall | 4.257 | 5.530 | 4.512 | 9.730 | 8.130 |
| year1,spring | 8.574 | 10.732 | 5.167 | 20.145 | 15.675 |
| end of year1 | 9.332 | 11.455 | 5.212 | 22.781 | 17.400 |
| end of year2 | 14.077 | 22.003 | 5.731 | 33.926 | 31.690 |
| end of year3 | 15.734 | 32.849 | 5.988 | 35.612 | 40.169 |
| end of year4 | 16.091 | 40.369 | 6.063 | 35.904 | 45.189 |
| cum. GPA | | | | | |
| end of year1 | 3.112 | 2.686 | 1.886 | 2.942 | 3.159 |
| end of year2 | 2.888 | 2.615 | 1.773 | 2.918 | 3.058 |
| end of year3 | 2.926 | 2.621 | 1.815 | 2.702 | 2.920 |
| end of year4 | 3.019 | 2.515 | 1.896 | 2.576 | 2.822 |

*Note.* Columns 1–4 are restricted to dropouts among 2002-2004 fall entry cohorts. Column 1 presents averages of enrollment history among low-credit medium dropout cluster, column 2 is for late dropout cluster, column 3 is for trial cluster, and column 4 is for high-credit medium dropout cluster. Column 5 shows averages for completers among 2002-2004 fall entry cohorts.

Table 2.5: HCA Financial Aid

| Variables | Low-credit<br>Medium Dropout | Late<br>Dropouts | Trials | High-credit<br>Medium Dropouts | Completers |
|---|---|---|---|---|---|
| Has Financial Aid | | | | | |
| year1, fall | 0.348 | 0.411 | 0.352 | 0.463 | 0.364 |
| year2, fall | 0.325 | 0.373 | 0.257 | 0.421 | 0.240 |
| year3, fall | 0.258 | 0.341 | 0.268 | 0.298 | 0.152 |
| year4, fall | 0.227 | 0.300 | 0.308 | 0.216 | 0.089 |
| Amt.Total Aid($) | | | | | |
| year1, fall | 493.998 | 655.676 | 538.799 | 757.313 | 545.605 |
| year2, fall | 474.435 | 621.098 | 382.859 | 715.348 | 397.823 |
| year3, fall | 315.772 | 585.071 | 399.376 | 485.658 | 262.550 |
| year4, fall | 274.700 | 514.916 | 519.884 | 295.377 | 163.162 |
| Has Grant | | | | | |
| year1, fall | 0.336 | 0.395 | 0.332 | 0.440 | 0.335 |
| year2, fall | 0.314 | 0.355 | 0.240 | 0.398 | 0.225 |
| year3, fall | 0.253 | 0.320 | 0.245 | 0.282 | 0.141 |
| year4, fall | 0.221 | 0.285 | 0.288 | 0.194 | 0.083 |
| Amt.Total Grant($) | | | | | |
| year1, fall | 455.901 | 603.839 | 493.547 | 690.171 | 485.291 |
| year2, fall | 431.713 | 553.312 | 353.188 | 636.407 | 343.479 |
| year3, fall | 292.476 | 494.448 | 352.626 | 412.740 | 216.133 |
| year4, fall | 240.787 | 436.969 | 445.822 | 253.333 | 129.750 |
| Has Loan | | | | | |
| year1, fall | 0.019 | 0.025 | 0.022 | 0.029 | 0.024 |
| year2, fall | 0.020 | 0.027 | 0.023 | 0.028 | 0.019 |
| year3, fall | 0.010 | 0.032 | 0.017 | 0.025 | 0.015 |
| year4, fall | 0.014 | 0.027 | 0.021 | 0.018 | 0.011 |
| Amt.Total Loan($) | | | | | |
| year1, fall | 25.657 | 31.967 | 28.126 | 43.725 | 34.655 |
| year2, fall | 31.014 | 43.230 | 23.462 | 48.250 | 32.580 |
| year3, fall | 18.915 | 58.018 | 28.349 | 53.369 | 30.392 |
| year4, fall | 33.283 | 54.594 | 45.911 | 25.575 | 23.928 |

*Note.* Columns 1–4 are restricted to dropouts among 2002-2004 fall entry cohorts. Column 1 presents averages of financial aid among low-credit medium dropout cluster, column 2 is for late dropout cluster, column 3 is for trial cluster, and column 4 is for high-credit medium dropout cluster. Column 5 shows averages for completers among 2002-2004 fall entry cohorts.

Table 2.6: Distribution of Cluster Assignment for Completers

| Matched Clusters | Sample Size | % |
|---|---|---|
| 1 (Low-credit Medium Dropouts) | 6438 | 17.60 |
| 2 (Late Dropouts) | 8393 | 22.94 |
| 3 (Trials) | 9726 | 26.58 |
| 4 (High-credit Medium Dropouts) | 12024 | 32.87 |

*Note.* Samples are restricted to completer among 2002-2004 fall entry cohort students that are matched to a cluster (i.e., distance to cluster medoid points fall within caliper). This table shows distribution of cluster assignment from hierarchical clustering analysis. First column indicates cluster assignments, second column is sample size of each cluster, and third column indicates sample proportions of each clusters.

Table 2.7: 10-fold Cross Validation Performance Metrics

| **CV Performance Metrics** | **Low-cred. Dropouts** | | **Late Dropouts** | | **Trials** | | **High-cred. Dropouts** | |
|---|---|---|---|---|---|---|---|---|
| | mean | sd | mean | sd | mean | sd | mean | sd |
| **Logistic with Elastic Net** | | | | | | | | |
| accuracy | 0.660 | 0.018 | 0.788 | 0.010 | 0.552 | 0.032 | 0.743 | 0.009 |
| **auc** | **0.748** | **0.008** | **0.846** | **0.008** | **0.677** | **0.006** | **0.815** | **0.008** |
| precision | 0.591 | 0.022 | 0.740 | 0.013 | 0.444 | 0.019 | 0.725 | 0.016 |
| recall | 0.801 | 0.029 | 0.943 | 0.008 | 0.820 | 0.049 | 0.907 | 0.016 |
| specificity | 0.546 | 0.053 | 0.602 | 0.022 | 0.396 | 0.078 | 0.512 | 0.039 |
| **Random Forest** | | | | | | | | |
| accuracy | 0.660 | 0.02 | 0.760 | 0.009 | 0.621 | 0.022 | 0.759 | 0.007 |
| **auc** | **0.751** | **0.01** | **0.829** | **0.005** | **0.715** | **0.006** | **0.832** | **0.007** |
| precision | 0.587 | 0.02 | 0.724 | 0.019 | 0.492 | 0.021 | 0.746 | 0.013 |
| recall | 0.818 | 0.04 | 0.911 | 0.022 | 0.767 | 0.037 | 0.895 | 0.020 |
| specificity | 0.530 | 0.06 | 0.580 | 0.044 | 0.536 | 0.054 | 0.567 | 0.044 |
| **Gradient Boosting** | | | | | | | | |
| accuracy | 0.662 | 0.021 | 0.805 | 0.007 | 0.615 | 0.020 | 0.776 | 0.008 |
| **auc** | **0.759** | **0.015** | **0.868** | **0.008** | **0.719** | **0.008** | **0.849** | **0.006** |
| precision | 0.590 | 0.025 | 0.759 | 0.012 | 0.487 | 0.016 | 0.760 | 0.013 |
| recall | 0.825 | 0.033 | 0.941 | 0.011 | 0.786 | 0.037 | 0.903 | 0.021 |
| specificity | 0.531 | 0.058 | 0.642 | 0.021 | 0.514 | 0.054 | 0.595 | 0.045 |

*Note.*The table was generated using dropout and matched completer samples of 2002-2004 fall entry cohorts. The values are averages and standard errors of performance metrics using 10-fold cross validation using 70% of training data. The first two columns are for low-credit dropout cluster and matched completers, columns 3–4 are for late dropout cluster and its matched completers, columns 5–6 are for trial dropouts and its matched completers, and columns 7–8 are for high credit dropouts and its matched completers. Top panel estimates are for logistic regression with elastic net model, middle panel estimates are for random forest models, and bottom panel estimates are for gradient boosting models.

Table 2.8: Out-of-Sample Prediction Performance: AUC values

| **Prediction** | **Low-cred. Dropouts** | **Late Dropouts** | **Trials** | **High-cred. Dropouts** |
|---|---|---|---|---|
| **GLM** | | | | |
| **auc** | 0.755 | 0.850 | 0.676 | 0.829 |
| **Random Forest** | | | | |
| **auc** | 0.746 | 0.831 | 0.709 | 0.851 |
| **Gradient Boosting** | | | | |
| **auc** | 0.760 | 0.867 | 0.713 | 0.865 |

*Note.*The table was generated using dropout and matched completer samples of 2002-2004 fall entry cohorts. The values indicate AUC prediction performance value using 30% out-of-sample prediction using the best performing hyper parameters from 10-fold cross validation models. The first column is for low-credit dropout cluster and matched completers, second column is for late dropout cluster and its matched completers, third column is for trial dropouts and its matched completers, and fourth column is for high credit dropouts and its matched completers. Top panel estimates are for logistic regression with elastic net model, middle panel estimates are for random forest models, and bottom panel estimates are for gradient boosting models.

Table 2.9: Cross Validation Performance Metrics using Relative Year Variables as Predictors

| **CV Performance Metrics** | **Year 1** | | **Year 2** | | **Year 3** | | **Year 4** | | **Year 5** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | mean | sd | mean | sd | mean | sd | mean | sd | mean | sd |
| **Logistic with Elastic Net** | | | | | | | | | | |
| accuracy | 0.586 | 0.010 | 0.634 | 0.010 | 0.662 | 0.009 | 0.691 | 0.009 | 0.709 | 0.009 |
| **auc** | **0.691** | **0.005** | **0.740** | **0.005** | **0.759** | **0.004** | **0.770** | **0.004** | **0.788** | **0.005** |
| precision | 0.539 | 0.008 | 0.579 | 0.010 | 0.608 | 0.012 | 0.646 | 0.011 | 0.663 | 0.013 |
| recall | 0.877 | 0.021 | 0.841 | 0.017 | 0.815 | 0.016 | 0.775 | 0.020 | 0.785 | 0.017 |
| specificity | 0.322 | 0.039 | 0.449 | 0.034 | 0.525 | 0.030 | 0.615 | 0.035 | 0.639 | 0.032 |

*Note.*The table was generated using full sample of 2002-2004 fall entry cohorts. The values are averages and standard errors of performance metrics using 10-fold cross validation using 70% of training data. The first two columns use only year one relative variables, columns 3–4 use year one and two relative variables, columns 5–6 use year one, two, and three relative variables, columns 7–8 use year one, two, three, and four relative variables, and columns 9–10 use years one through five relative variables. The prediction model used is logistic with elastic net.

# Bibliography

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

[2] Alexander, K., Entwisle, D., & Kabbani, N. S. (2001). The Dropout Process in Life Course Perspective. *Teachers College Record,* 103, 760-882.

[3] Allensworth, E. M., Nagaoka, J., & Johnson, D. W. (2018). High School Graduation and College Readiness Indicator Systems: What We Know, What We Need to Know. Concept Paper for Research and Practice. *University of Chicago Consortium on School Research.*

[4] Anderberg, M. R. (1973). *Cluster analysis for applications.* New York: Academic Press

[5] Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, 113, 177-194.

[6] Attewell, P., Heil, S., & Reisel, L. (2012). What is academic momentum? And does it matter?. *Educational Evaluation and Policy Analysis*, 34(1), 27-44.

[7] Attewell, P., & Monaghan, D. (2016). How many credits should an undergraduate take?. *Research in Higher Education*, 57(6), 682-713.

[8] Bailey, T. R. % Alfonso, M. (2005). Paths to persistence: An analysis of research on program effectiveness at community colleges. *Lumina Foundation for Education New Agenda Series*, 6(1). Indianapolis, IN: Lumina Foundation for Education.

[9] Bean, J. P. (1983). The application of a model of turnover in work organizations to the

student attrition process. *The review of higher education*, 6(2), 129-148.

[10] Bettinger, E. (2004). How financial aid affects persistence. In College choices: The economics of where to go, when to go, and how to pay for it (pp. 207-238). *University of Chicago Press.*

[11] Borsato, G. N., Nagaoka, J., & Folley, E. (2013). College Readiness Indicator Systems Framework. *Voices in Urban Education*, 38, 28-35.

[12] Bowers, A. J. (2007). Grades and Data Driven Decision Making: Issues of Variance and Student Patterns. *Online Submission.* http://eric.ed.gov/?id=ED538574

[13] Bowers, A.J. (2015-2017) Using Big Data to Investigate Longitudinal Education Outcomes through Visual Analytics. *National Science Foundation, Directorate for Computer & Information Science & Engineering (CISE), Division of Information & Intelligent Systems (NSF IIS-1546653).* http://www.nsf.gov/awardsearch/showAward?AWD_ID=1546653

[14] Bowers, A.J. (2010) Analyzing the Longitudinal K-12 Grading Histories of Entire Cohorts of Students: Grades, Data Driven Decision Making, Dropping Out and Hierarchical Cluster Analysis. *Practical Assessment, Research & Evaluation (PARE)*, 15(7), 1-18. http://pareonline.net/pdf/v15n7.pdf (Preprint available)

[15] Bowers, A.J., Zhou, X. (2018) Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes. *Journal of Education for Students Placed at Risk (JESPAR)*, 24(1), 20-46.

[16] Breiman, L. (2001). Random forests, *Machine Learning* 45: 5–32.

[17] Calcagno, J. C., Crosta, P., Bailey, T., & Jenkins, D. (2007). Stepping stones to a degree: The impact of enrollment pathways and milestones on community college student outcomes. *Research in Higher Education*, 48(7), 775-801.

[18] Click, C., Malohlava, M., Parmar, V., Roark, H., and Candel, A. (Nov 2017). Gradient

Boosting Machine with H2O. *H2O.ai*, `http://h2o.ai/resources/`.

[19] Denning, J. T., Marx, B. M., & Turner, L. J. (2017). *ProPelled: The Effects of Grants on Graduation, Earnings, and Welfare* (No. w23860). National Bureau of Economic Research.

[20] DesJardins, S. L., Ahlburg, D. A., & McCall, B. P. (2002). Simulating the longitudinal effects of changes in financial aid on student departure from college. *Journal of Human resources*, 653-679.

[21] Doyle, W. R. (2011). Effect of increased academic momentum on transfer rates: An application of the generalized propensity score. *Economics of Education Review*, 30(1), 191-200.

[22] Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access*, 5, 15991-16005.

[23] Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25), 14863-14868.

[24] Eisen, M. B., & DeHoon, M. (2002). *Cluster 3.0 Manual.* Palo Alto, CA: Stanford University

[25] Gordon, A. (1999). *Classification* (2nd edition), Chapman and Hall/CRC Press, London.

[26] Gurantz, O., & Borsato, G. N. (2012). Building and implementing a college readiness indicator system: Lessons from the first two years of the CRIS Initiative. *Voices in Urban Education*, 35, 5-15.

[27] H2O.ai (Oct. 2016). R Interface for H2O, R package version 3.10.0.8. *H2O.* `https://github.com/h2oai/h2o-3`.

[28] Hastie, H., Tibshirani, R., & Friedman, F. (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer Series in Statistics.

[29] Ishitani, T. T. (2006). Studying attrition and degree completion behavior among first-

generation college students in the United States. *The Journal of Higher Education*, 77(5), 861-885.

[30] Janosz, M., Le Blanc, M., Boulerice, B., & Tremblay, R. E. (2000). Predicting different types of school dropouts: A typological approach with two longitudinal samples. *Journal of Educational Psychology*, 92(1), 171-190. doi:10.1037/0022-0663.92.1.171

[31] Jenkins, D., & Bailey, T. (2017). *Early Momentum Metrics: Why They Matter for College Improvement* (Brief No. 65). Community College Research Center, Columbia University. Retrieved from http://ccrc.tc.columbia.edu/media/k2/attachments/early-momentum-metrics-college-improvement.pdf

[32] John, E. P., Cabrera, A. F., Nora, A., & Asker, E. H. (2000). Economic influences on persistence reconsidered. *Reworking the student departure puzzle*, 29-47.

[33] Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York.

[34] Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning* (Vol. 1). STHDA.

[35] Knowles, J. E. (2015). Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin. *Journal of Educational Data Mining*, 7(3), 18-67.

[36] Lee, J., Recker, M., Bowers, A.J., Yuan, M. (2016). Hierarchical Cluster Analysis Heatmaps and Pattern Analysis: An Approach for Visualizing Learning Management System Interaction Data. In *EDM* (pp. 603-604).

[37] Ma, J., & Baum, S. (2016). Trends in community colleges: Enrollment, prices, student debt, and completion. *College Board Research Brief*, 4, 1-23.

[38] Mabel, Z., & Britton, T. A. (2018). Leaving late: Understanding the extent and predictors of college late departure. *Social science research*, 69, 34-51.

[39] McFarland, J., Hussar, B., de Brey, C., Snyder, T., Wang, X., Wilkinson-Flicker, S., ... & Bullock Mann, F. (2017). The Condition of Education 2017. NCES 2017-144. *National*

*Center for Education Statistics.*

[40] Nitkin, D. (2018). *Technology-Based Personalization: Instructional Reform in Five Public Schools* (Doctoral dissertation, Columbia University).

[41] Olson, R. S., La Cava, W., Mustahsan, Z., Varik, A., & Moore, J. H. (2017). Data-driven advice for applying machine learning to bioinformatics problems. *arXiv preprint arXiv:1708.05070.*

[42] Park, R. S. E., & Scott-Clayton, J. (2018). The impact of Pell Grant eligibility on community college students's financial aid packages, labor supply, and academic outcomes. *Educational Evaluation and Policy Analysis*, 40(4), 557-585.

[43] Pena-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, 41(4), 1432-1462.

[44] Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1), 135-146.

[45] Romero, C., Ventura, S., & Garcia, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, 51(1), 368-384.

[46] Romesburg, H. C. (1984). *Cluster analysis for researchers.* Belmont, CA: Lifetime Learning Publications.

[47] Stratton, L. S., O'Toole, D. M., & Wetzel, J. N. (2008). A multinomial logit model of college stopout and dropout behavior. *Economics of education review*, 27(3), 319-331.

[48] Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition* (2nd ed.). Chicago, IL: University of Chicago Press.

[49] Tucker, M., Daro, P., Snow, C., Pellegrino, J., Everson, H., Glasper, R., ... & Fain, P. (2013). What Does It Really Mean to Be College and Work Ready? The Mathematics and English Literacy Required of First Year Community College Students.

[50] Weinstein, J. N., Myers, T. G., O'Connor, P. M., Friend, S. H., Fornace, A. J., & Kohn, K. W., et al. (1997). An information-intensive approach to the molecular pharmacology

of cancer. *Science*, 275(5298), 343-349

[51] Willett, J. B., & Singer, J. D. (1991). From whether to when: New methods for studying student dropout and teacher attrition. *Review of educational research*, 61(4), 407-450.

[52] Wright & Ziegler (2017) Wright MN, Ziegler A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*. 2017;77(1):1–17.

[53] Xu, R., & Wunsch, D. (2008). /em Clustering (Vol. 10). John Wiley & Sons.

[54] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2), 301-320.

# Chapter 3

# The Impact of Pell Grant Eligibility on Community College Students' Financial Aid Packages, Labor Supply, and Academic Outcomes

Rina Seung Eun Park and Judith Scott-Clayton

(The authors' names are listed alphabetically. R. Park and J. Scott-Clayton contributed equally to this work.)

# I Introduction

In 1965, President Lyndon Johnson signed into law the Higher Education Act of 1965, which initiated the precursors to today's Pell Grant and Stafford Loan programs and solidified the federal government's role in higher education finance. Since then, the importance of federal financial aid policy has only increased. In 2014-15, the federal government provided over $120 billion in student loans, grants, and other forms of financial aid for undergraduates - more than four times the level of support provided in 1990-91.

The federal Pell Grant program is the largest single source of grant aid, providing $30.3 billion in grants to over 9 million students annually in 2014-15, up to $5,775 each per year. Students can use the grant at any eligible institution, and receive the same amount regardless of where they go. Although the eligibility formula is complex, family income is the main component: those with family income below $30,000 typically receive the maximum award, while only about 5 percent of those with family incomes above $70,000 receive any award. If the award exceeds tuition and fees, students can use the extra amount for books, food, or other living expenses.

Although a large body of research convincingly demonstrates that financial aid programs can influence student enrollments and completion (Page & Scott-Clayton, 2016; Deming & Dynarski, 2009; Long, 2008), evidence on the effects of Pell Grants specifically is more mixed. Two studies of the effect of the introduction of Pell Grants found no evidence that college enrollments increased any faster for Pell-eligible students relative to ineligible students (Hansen, 1983; Kane, 1995). More recently, a regression-discontinuity analysis of urban community college students just above and below the eligibility cutoff for Pell finds no impact on college choice, course credits or degree completion (Marx & Turner, 2015). On the other hand, Pell Grants appear to positively influence enrollment rates for adult students (Seftor & Turner, 2002) and may increase persistence and acceleration in graduation conditional on enrollment (Bettinger, 2004; Denning, 2016).

The ambiguous evidence regarding Pell has led researchers to investigate possible explanations. Several studies have suggested that the complexity of the federal aid application process and late notice of Pell eligibility may undermine the ability of the program to reach students who need aid most (Dynarski & Scott-Clayton, 2006; Bettinger, Long, Oreopoulos, & Sanbonmatsu, 2012; Dynarski & Wiederspan, 2012; Scott-Clayton, 2013).

While progress has been made over the past few years to simplify the federal aid application process and allow students to apply for aid earlier, another potential explanation for the mixed effects of Pell has received comparatively less attention: state and institutional aid policies may interact with the federal aid formula in a way that makes it difficult to isolate the effect of Pell. The interaction of multiple governments' fiscal decisions in a redistributive program like Pell is an example of fiscal vertical externalities (Boadway & Tremblay, 2012; Johnson, 1988): the federal government acts as the "first mover" by establishing Pell as the foundation of financial aid packages (Pell Grants are never reduced as a result of other aid eligibility), but states or institutions as second movers can reduce or retarget their own aid dollars in response.

For example, research by Lesley Turner (2014) finds that selective nonprofit institutions capture, via reductions in institutional aid, 67 cents of every Pell dollar received by their students. Bettinger and Williams (2013) also find a negative correlation between Pell Grants and state aid, while McPherson and Schapiro (1991) find a positive correlation between Pell Grants and overall institutional aid.[1] Finally, studies have found that students may adjust their own borrowing decisions in response to grant eligibility, such that receiving an extra dollar of grant aid often leads to less a dollar of total additional aid received (Marx & Turner, 2015; Goldrick-Rab, Kelchen, Harris, & Benson, 2016). Interactions with state and institutional aid programs may also help explain why the estimated effects of Pell are

[1] Tuition levels are another channel through which the impact of Pell could be diminished (this is often referred to as the "Bennett hypothesis" after former Secretary of Education William Bennett), although empirical research on this question has found mixed results (Singell & Stone, 2007; Rizzo & Ehrenberg 2004; Turner 2014).

not consistent from study to study, because state and institutional aid programs can vary substantially from context to context.

The availability of large administrative datasets facilitates research designs that were not feasible in decades past. In this paper, we utilize such a dataset from a single state on a population of particular interest: community college enrollees. We implement a regression-discontinuity design that examines the effects of just barely qualifying for a Pell Grant on the composition of recipients' overall financial aid package, students' labor supply, and subsequent academic outcomes.

We find that even at community colleges, other sources of student aid do shift substantially around the cutoff for Pell, consistent with Turner (2014) and Marx and Turner (2015). We find distinctive patterns of financial packaging depending on whether or not institutions participated in federal loan programs. At institutions that participate in the federal student loan programs, students above the cutoff (who are ineligible for Pell) borrowed 55% more than those below the cutoff. This pattern replicates the findings in previous research by Marx and Turner (2015), though it appears even more strongly in our sample. On the other hand, at institutions that did not offer loans, students just above the Pell cutoff received state/institutional grants that offset the discontinuity in Pell Grants (that is, at school not participating in the loan programs, there is no discontinuity in overall grant aid around the Pell cutoff).

For our analysis of student labor supply and academic outcomes, we limit the sample to students attending only loan-offering schools, and interpret the estimates as showing the effects of shifting students' aid packages from federal loans to Pell grants. [2] We find that qualifying for the minimum Pell increases the intensity of enrollment, with recipients 4-7 percentage points more likely to enroll full-time from the spring of their first year to the spring of their second year. We also find evidence that those who are just barely eligible for Pell

[2] We distinguish loan-offering school by looking at average loan take-up rates across cohorts. Although no-loan schools include those with non-zero take up rates, however, the rates were very close to zero. Schools that offered loans, no-loans, and switchers were clearly distinguishable.

earn less in the first two years after entry, suggesting a reduction of labor supply equivalent to perhaps one or two hours per week. This is consistent with previous findings that grants decrease the need to work for pay and allow student to shift their time allocation from work to school (Benson & Goldrick-Rab, 2011; Schudde, 2013). For cumulative outcomes at the end of three years - on cumulative GPA, cumulative credits earned, degree completion, and transfer within three years of entry - we cannot detect statistically significant effects, though the point estimates are positive and of a magnitude consistent with the impacts on enrollment intensity throughout the first two years.

After presenting our main results, we examine their sensitivity to possible selection bias. Our analysis uses data on community college entrants, but Pell eligibility may shift who chooses to enroll in a community college in the first place. Indeed, we find a discontinuity in the density of observations around the cutoff that suggests students who qualify for Pell are disproportionately induced not to enroll in community college (perhaps because they attend either a four-year or for-profit institution instead). While we are reassured that student characteristics do not appear to shift around the cutoff, we also address the problem using two methods introduced in the literature: 1) limiting our analysis to a subset of schools where we do not observe any evidence of differential selection, and 2) performing a bounding analysis under extreme assumptions about the missing population.

Unfortunately, because our main estimates are modest to begin with, they are not particularly robust to these rigorous sensitivity checks, leaving open the possibility that some of the positive effects we find may be due to differential selection into community colleges around the Pell grant cutoff. Still, because we find no differences in observed characteristics around the cutoff, we still view our main results as a reasonable ”best guess” regarding the impact of receiving a small Pell grant. In addition, a valuable side effect of examining the potential selection problem is that we can provide some insight on how Pell grant eligibility may influence institutional choice: the selection patterns we find are much more concentrated in areas with many nearby for-profit institutions.

Our paper contributes to the literature in three ways. First, we take a step towards understanding how the nation's largest need-based grant interacts with other aid programs. We find that other aid programs do respond to federal Pell Grant. Not only so, we find clear distinctive patterns of financial aid packaging between institutions that participate in federal loans versus those they do not. Second, our paper is one of the few that looks into the interaction of Pell eligibility with employment intensity during enrollment. Much interest on Pell Grant program has focused particularly on the impacts on college enrollment of low-income students. We show that students who are just below the cutoff (Pell eligible) seem to shift their time allocation, reducing work while increasing their enrollment intensity. Finally, our results provide indirect evidence that Pell grants may influence student enrollment decisions, in contrast to the findings of Marx and Turner (2015).

The remainder of the paper proceeds as follows: Section 2 provides background on financial aid at community colleges and on the Pell Grant eligibility formula. Section 3 describes our data and sample. In section 4, we describe our regression discontinuity strategy and highlight key identification assumptions. Section 5 presents our results, and section 6 discusses implications and open questions.

# II Financial Aid at Community Colleges

Among community college students enrolled in 2011-2012, on average, 38% of student enrolled received Pell and 17% received federal student loans with an average amount of $1,140 and $781 per enrollee, respectively.[3] Students qualify for the same amount of Pell regardless of where they enroll, and if the Pell Grant exceeds tuition and fees, students can receive the remainder back as a refund to cover other educational and living expenses.

Pell is by far the largest source of grant aid for community college students, but approximately 12% of students also receive state grant aid and 13% receive institutional grant aid.

[3] Authors' tabulations using NCES Quick Stats with NPSAS:2012 data split by institution type.

While the average amounts of state and institutional aid (approximately $190 and $120, respectively) distributed per enrollee are much smaller than for Pell, our analysis below will suggest that these smaller programs can be particularly important for students around the margin of Pell eligibility. Moreover, institutions may have some discretion about how to distribute state grant aid. In the state we examine here, the state's need-based grant is given as a lump sum to institutions, which can then use their own formula to provide aid to students, as long as it is need-based.

To qualify for any federal aid, students must file a Free Application for Federal Student Aid (FAFSA). This application collects detailed information on students' income and assets, as well as similar information from the parents of dependent students. This information is used in a complex formula that provides an ”Expected Family Contribution” or EFC as its output. While over a hundred pieces of information are required to precisely calculate the EFC, for the vast majority of students, the EFC is determined by income, family size, and number in college (Dynarski, Scott-Clayton, & Wiederspan 2013). Lower income students will have lower EFCs. The EFC is used to distribute not just federal aid, but frequently state and institutional aid as well.

Pell eligibility is directly related to EFC: in general, Pell eligibility equals the maximum Pell in a given year, minus EFC. However, in most years, there is a minimum grant size such that the Pell does not decline continuously to zero, but may drop from several hundred dollars to zero at a certain point in the EFC distribution. The precise formula varies from year to year. In many years prior to 2008, the minimum grant size was $400 (those with eligibility between $200 and $399 were rounded up, while those with eligibility below $200 received nothing). In years since 2011, the minimum grant has been $200. However, between 2008 and 2010, the minimum grant size was much larger than usual, in part due to additional American Reinvestment and Recovery Act funding. In 2008-09 the minimum was $690, rising to $976 in 2009-10, and falling back to $555 in 2010-11. We thus focus on these years for our regression discontinuity analysis.

Eligibility for subsidized student loans is calculated as the total cost of attendance (including estimated living expenses, for students attending at least half-time), minus the EFC and other aid already received by the student, subject to annual loan maximums. Students are eligible for unsubsidized loans regardless of EFC. Between 2008 and 2010, the combined limit of subsidized and unsubsidized loans for first-year students was around \$5,500 annually for dependent students and \$9,500 annually for independent students.[4] It is also worth pointing out that total costs of attendance are high enough even at community colleges such that students receiving the minimum Pell Grant are very unlikely to have their state financial aid limited by the cost of attendance (in 2008, for example, average total cost of attendance for full-time students at community colleges was \$9,700).[5]

Not all students at community college receive a federal loan offer in their financial aid packages. Colleges sometimes choose to opt out from Stafford loan program in fear of sanctions by the federal government.[6] For students who are eligible for Pell Grant, attending colleges that include (relative to colleges that do not include) federal loan offer in their financial aid package had higher likelihood and amount of borrowing as well as attempted credit hours in the first year (Wiederspan, 2016).

Examining the effect of a modest Pell Grant for students at community colleges has two advantages. First, the monetary incentive is sharpest for these students: the minimum Pell Grant, which averaged \$750 between 2008 and 2010, represented a more than 25% discount on tuition and fees during that time period.[7] Second, because of open-access admissions, community college enrollees are arguably more likely to be on the margin of college attendance and persistence (i.e., potentially more likely to change behavior as a result of aid), and thus represent a key target population for need-based aid.

[4] Federal loan limits are resourced from `http://www.finaid.org/loans/historicallimits.phtml`

[5] 2008 figure based on NPSAS:2008 data, using ”student budget (attendance adjusted)” variable for full-time students.

[6] If an institution have more than 30 percent cohort default rate for three consecutive years, that school is prohibited to offer any federal financial aid, including Pell Grant, for three years (Widerspan, 2016).

[7] Based on estimated average tuition and fees of \$2,713 in 2010-11 (Baum & Ma, 2011).

# III Data and Sample

The administrative data we use includes more than 20 community colleges in a single state. The data includes four types of information: student demographics, first-year financial aid eligibility and receipt, transcript data, degree/transfer information, and quarterly earnings. Student demographics include race, gender, age, family income, and dependency status. Financial aid information includes the expected family contribution or EFC (the summary measure of financial need which determines eligibility for Pell and other federal aid), and amounts of federal, state, and institutional aid actually received (broken out into detailed types of aid). Transcript data include remedial placement test scores for those who took such tests, credits attempted and earned, and grades for each term enrolled in any of the states' community colleges. Credential completion and transfer to four-year institutions are measured using data from the National Student Clearinghouse (NSC), which includes data for students who leave the community college system. Finally, student records are matched to quarterly earnings records, which we use to measure of student labor supply during the first two years post-entry.

The data are limited to first-time, fall entrants to the community college system. We focus on the 2008-2010 entry cohorts because of particularly large discontinuities in the Pell formula during those years (in earlier and later years, minimum awards were much smaller). In these years, the data include a total of 89,000 students. We further limit our sample to the 57% of students who filed a FAFSA (and thus have the financial information we need for the regression discontinuity analysis) and have EFCs within $2000 of the Pell cutoff in the relevant year.

Table 3.1 shows the characteristics and financial aid measures of our sample. The first three columns describe our analysis sample, while the fourth column provides statistics on the full sample of enrollees (regardless of EFC and including those who did not file a FAFSA) during these years, for comparison.[8] The majority of students in our sample are White students, about equally distributed in gender. On average, students in entry cohorts are slightly above 21 years old. About 60 percent of students in our analysis sample persisted to the subsequent fall, and about one-third transferred or received a degree within three years of entry. The final column provides national averages from the Beginning Postsecondary Students (BPS) 2012/2014 survey, representing first-time students who entered a public two-year college during academic year 2011-12. On average, compared to the BPS sample, our main analysis sample (column 3) has fewer Hispanic students, and has lower family income. In terms of financial aid, students in our sample received less state aid and borrowed less compared to the BPS sample.

Table 3.1 indicates that students above and below the EFC cutoff for receiving Pell are actually quite similar along most demographic dimensions other than family income. This confirms large differences in Pell receipt around the cutoff, but also highlights that students who are ineligible for Pell are also much more likely to take out student loans, and somewhat more likely to receive state grant aid. We will examine these patterns in more detail below.

# IV Empirical methodology

## A Regression discontinuity design

We use a Regression Discontinuity (RD) design to estimate the causal effect of Pell Grant eligibility for those near the EFC cutoff, using EFC as our forcing variable. The statutory discontinuity in Pell for a full-time student was $690 in 2008-09, $976 in 2009-10, and $555 in 2010-11 (awards are prorated for less-than-full-time enrollment).[9] The formula is reflected

[8] For dependent status, family income, family size, and EFC, our data has information only on those who have filed a FAFSA.

[9] In 2008 and 2009, Pell simply rises linearly below the cutoff until it reaches the maximum. In 2010, the formula takes a particularly weird shape, with eligibility fixed at $555 for students within a range below the threshold, then rising linearly for a range, then discontinuously jumping again by about $327 at an EFC approximately $500 below the cutoff. This odd pattern in 2010 can be detected in Figure 3.1.

in Figure 3.1, which plots students' estimated Pell eligibility based on their EFC. We use estimated Pell eligibility here instead of actual Pell amounts received, because amounts received are endogenous to enrollment intensity. Later graphs that show actual Pell received will reflect a similar, if slightly muted pattern (since amounts received can only be equal to or less than estimated eligibility).

The intuition behind the RD is that if we can assume that the relationship between EFC and an outcome variable is continuous as we approach the cutoff from either direction, then any discontinuity in the outcome at the cutoff can be attributed to the discontinuity in treatment. Formally, using Rubin's (1974) potential outcomes framework, let $Y_{0i}, Y_{1i}$ be potential outcomes for an individual i without treatment and with treatment, respectively. Let $PellEligible_i \in \{0, 1\}$ indicate treatment status.

We can then model outcomes as:

$$\begin{aligned} Y_i &= Y_{0i} + (Y_{1i} - Y_{0i})PellEligible_i \\ &= f(EFC_i) + \beta_i PellEligible_i + v_i \end{aligned} \tag{3.1}$$

where, $\beta_i$ is the treatment effect, $E[Y_{0i}|EFC_i] = f(EFC_i)$, and $v_i = Y_{0i} - E[Y_{0i}|EFC_i]$. The idea behind the RD design is that Pell eligibility is deterministic by EFC (i.e. $PellEligible = g(EFC)$).

Causal inference in the RD model relies on two assumptions: (1) a discontinuity in treatment assignment $E[PellEligible_i|EFC_i = c]$ exists at the cutoff ($c_0$) and (2) ($EFC_i$) is continuous in the neighborhood of the cutoff ($c_0$) (Hahn, Todd, and Vander Klaauw, 2001; Imbens & Lemieux, 2008).

If all assumptions above hold, then the local average treatment effect is:

$$\beta = lim_{c \to c_{0}+} E[Y_i|EFC_i = c] - lim_{c \to c_{0}-} E[Y_i|EFC_i = c] \tag{3.2}$$

In words, the RD estimate is the difference of two regression functions at the cutoff ($c_0$).

We use a "sharp" RD estimator, since the treatment of interest is Pell eligibility rather than receipt, and eligibility is completely determined by the forcing variable. Refer to Appendix Figure 3.1 for the relationship between EFC and actual probability and amount of Pell receipt. We implement the RD using a local linear regression estimator with a rectangular kernel (that is, with all observations weighted equally) for observations within ±2000 from the EFC cutoff (Hahn, Todd, & Klaauw, 2001; Imbens & Lemieux, 2008).[10] Specifically, we estimate:

$$\begin{aligned} Y_{ist} = \alpha + \beta_1(PellEligible_i) + \beta_2(Dist_i \times Above_i) + \beta_3(Dist_i \times Below_i) \\ + X_i\delta + \phi_s + \tau_t + \varepsilon_{ist} \end{aligned} \tag{3.3}$$

where, $Dist_{it}$ is distance from the EFC cutoff for Pell eligibility in the relevant year ($Dist_{it} = EFC_{it} - c_{0t}$) $Above_{it}/Below_{it}$ is a binary outcome indicating whether individual i in year t has EFC that is above or below the cutoff; $X_i$ is a vector of individual-level covariates including race dummies, age, income, dependent status, whether the student had dual enrollment credits from high school, and placement math, reading, and writing scores (with flags for missing scores); $\phi_s$ is a vector of school fixed effects; and $\tau_t$ is vector of dummies for each cohort. If the RD assumptions hold, adding covariates ($X_i$) is not necessary for identification of causal effects, but will adjust for small sample bias and reduce standard errors.

In addition to testing for sensitivity across different bandwidths, we also use three bandwidth selection methods: cross validation (Ludwig & Miller, 2005) and two plug-in rules - Imbens and Kalyanaraman (hereafter, IK) (2012) and Calonico, Cattaneo, and Titiunik (hereafter, CCT) (2014) - as a comparison to our baseline specification.[11] We estimate op-

[10] When using a subset of points to fit a local regression, different weights can be used to the fit data points (mostly, weight is given as a function of distance to the point estimator). This weight function is referred to a kernel. In the regression discontinuity literature, there is no consensus in an optimal choice of kernel because in practice different weight functions should have little impact on the estimator (Fan & Gijbels, 1996; McCrary & Royer, 2003; DesJardins & McCall, 2008; Lee & Lemieux, 2010). For consistency, we use a rectangular kernel, giving equal weights to all local points within the bandwidth, throughout the paper as suggested by Lee and Lemieux (2010).

[11] Lee and Lemieux (2010) also introduce rule-of-thumb bandwidth introduced by DesJardins-McCall

timal bandwidths under each method for all the outcomes separately and examine their distribution.

## B Threats to Validity

A key assumption for an unbiased RD estimator is that individuals should not be able to systematically manipulate whether they fall above or below the cutoff of the forcing variable. Because of the opaque nature of the EFC calculation, the fact that both the EFC formula and the relevant cutoffs change from year to year, and the fact that a high proportion of financial aid applicants will have to submit tax documents to verify their income, we are skeptical that students/families can manipulate their EFCs very precisely.

However, another way that the assumption of continuity in $f(EFC_i)$ can be violated is if there is differential sample selection around the cutoff. This is a bigger concern in this context, because our sample includes only students who ultimately enrolled in the community college system, and most students learn their aid eligibility prior to initial enrollment. If Pell eligibility induces some individuals to enroll in college who would not have otherwise, or if it influences students' choice of institution, this will cause a discontinuity in $f(EFC_i)$ within our sample frame.

This assumption can be tested by examining the density of observations around the cutoff. As shown in Figure 3.2, which plots density using $100 EFC bins, we can see that there is a jump in the number of observations just to the right of the cutoff; that is, students are more likely to appear in our community colleges sample if they are ineligible for Pell. The direction of this enrollment jump is counterintuitive to what we would expect if Pell Grant induced student's enrollment choices. To confirm this discontinuity, we conduct a McCrary (2008) test, which rejects the null hypothesis that the density is smooth. Given the direction of enrollment jump, we hypothesize that the "missing" students to the left of the cutoff may be using their Pell grants to attend other schools than community colleges. We explore on

(2008). We also run rule-of-thumb bandwidth and find similar to IK but slightly smaller.

this hypothesis further in the section following our main results.

Another approach to evaluating selection bias around the cutoff is to test for discontinuities in the baseline covariates around the EFC cutoff. Appendix Table A1 illustrates the relationship between covariates and EFC where we use a version of equation (3.3) above with covariates on the left-hand side to test for any significant discontinuities. Reassuringly, despite the substantial discontinuity in the density, we find no evidence of discontinuities in any baseline covariates at the cutoff in our preferred 2000 bandwidth, including not just age, race, and gender, but also family income, dependency status, and placement test scores.[12] This conclusion holds even after limiting the sample to loan schools, which have the largest discontinuity in density.

Our primary strategy to mitigate selection bias is to control for observable characteristics around the cutoff. In addition, to assess the possible role of selection on unobservable dimensions, we test the sensitivity of our results by following two procedures introduced in the literature: 1) analysis of impacts for a subset of institutions for which no discontinuity in the density of observations is observed (as proposed by Calcagno and Long, 2008), and an RD bounding analysis (as proposed by Gerard, Rokkanen, and Rothe [hereafter, GRR], 2016). We describe these strategies in more detail after presenting our main results.

Finally, it is worth noting that while this discontinuity is problematic for an analysis of outcomes among community college enrollees, it also provides indirect evidence that Pell eligibility does influence initial enrollment decisions, which is an important margin of impact on its own. This is in contrast to findings in Marx and Turner (2014), who find no evidence that Pell eligibility affects either the enrollment margin or the choice of two- versus four-year college for students who applied to CUNY colleges.[13]

[12] For 4000 bandwidth specification, we see dual enrollment, age, and dependent variables as significantly different.

[13] The CUNY system is substantially more expensive, and arguably more stratified by ability, than the system under consideration in this paper. While purely speculative, this provides possible explanations for why Pell eligibility may impact college choice in this context but not in the CUNY context.

# V Results

## A Effects of Pell Grant eligibility on the composition of overall financial aid package

The two panels of Figure 3.3 illustrate how different components of students' aid packages change around the Pell eligibility cutoff, with observations grouped into \$100 EFC bins and the size of each circle reflecting the number of observations. All panels plot data for students at loan schools and no-loan schools separately, for reasons that will become clear. The left panel shows actual Pell Grant amounts received, and indicates an increase of approximately \$500 just to the left of the cutoff, with no difference between loan and no-loan schools.[14] However, a clear difference between these two institution types emerges when we look at the right panel, plotting average total grants by EFC. Across most of the EFC distribution, the institutions that do not offer student loans give out more in total grants. They also use state grant aid to compensate students just above the cutoff for Pell, such that at these institutions, there is no discontinuity in total grant aid around the Pell cutoff. A large discontinuity in total grant aid exists only for institutions that participate in the student loan programs.

The left panel of Figure 3.4 shows student loan receipt by EFC. Of course, at no-loan schools, student loans are zero throughout the distribution.[15] At loan-schools, we see a sizable jump in average loan amounts for students just above the Pell eligibility threshold. Considering all aid together, the bottom-right panel shows that for neither institution type is there any discontinuity in total aid received. For no-loan schools, state grant aid smoothes out the discontinuity in Pell, while for loan-schools, the discontinuity is smoothed out by

[14] This amount is less than the statutory discontinuity in Pell eligibility largely because of less-than-full-time enrollment.

[15] We suspect that the few observations off the line are either data errors or possibly students who switched institutions mid-year.

loans. (Also note that the higher state grant aid at no-loan schools does not completely make up for the lack of loans: students at no-loan schools receive substantially less in total aid than students at loan schools.)

Table 3.2 shows the regression results corresponding to the panels of Figure 3.5, with the top portion of the table showing results for loan schools and the bottom portion showing results for no-loan schools. Confirming what is visible in the pictures, there is a large discontinuity in Pell Grants in both cases, but at no-loan schools, there is no significant discontinuity in total grant aid, loans, or total aid. For loan schools, there is a significant discontinuity in total grant aid (coefficient = \$560, $p < 0.01$), but an equal-and-opposite discontinuity in loan aid (coefficient = −\$592, $p < 0.01$), leading to no discontinuity in total aid.[16] The pattern of loan take-up at these schools replicates that found in previous research by Marx & Turner (2015), though it appears even more strongly in our sample.

For no-loan schools, which in our sample represent about half of the institutions but only about one-quarter of students enrolled, we have no first stage: Pell eligibility has no discontinuous effect on any treatment we expect to matter (unless we think a dollar of Pell Grants affects students differently than a dollar of state grants).[17] Therefore, we limit our subsequent analyses to students attending only loan-offering schools, where we do observe a significant discontinuity in overall grant aid. Even at loan institutions, these findings alter how we think about the treatment. In interpreting the effects that follow, it is important to recognize that we are estimating the effect of receiving \$500 in grants instead of loans.[18]

[16] Note that total aid includes some other small aid programs, so that it may be slightly more than the sum of grants and loans.

[17] In results not shown here, we can confirm that there are no impacts on any outcome when we run our models for students at no-loan institutions. Moreover, there is no discontinuity in the density of observations around the cutoff for these schools.

[18] Moreover, as noted by Marx and Turner (2015), these averages mask important heterogeneity, because everyone to the left of the cutoff qualifies for a \$500 Pell Grant, but to the right of the cutoff, some students take out large loans while others take out nothing. Thus, some students who are bumped just below the cutoff will experience an increase in total aid, while others may actually take up less total aid than if they had not been Pell-eligible.

# B Effect of Pell Grant eligibility on academic outcomes and labor supply while enrolled

Table 3.3 shows our estimated impacts on academic outcomes and student labor supply. We examine re-enrollment and enrollment intensity, cumulative GPA and credits completed, and earnings during each of the first two years. We also examine GPA, credits attained, credentials, and transfer at the end of our three-year follow-up period. Note that for all outcomes, the difference in treatment is based on the first year difference in aid received; this does not measure the cumulative effect of receiving Pell for more than one year.[19]

With a few exceptions, our results are mostly in a positive direction, but small and not statistically significant. Among the notable exceptions are that we do find significant positive effects on full-time enrollment in the spring of the first year (5 percentage point increase from a base of 52 percent), full-time enrollment in the fall of the second year (7 percentage point increase from a base of 37 percent), and full-time enrollment in the spring of the second year (4 percentage point increase from a base of 33 percent). In contrast, we find a negative effect on summer term enrollment between Years 1 and 2 (of about 5 percentage points), which is surprising taking into account that these include years in which summer Pell Grants were available.[20]

We also find consistently negative earnings effects during the first two years, though the reduction is only statistically significant in the first year. The negative earnings effects translate into about $12-$20 less per week and are of the same order of magnitude as the increase in grant aid for Pell-eligible students. These reductions are consistent with a story in which Pell allows students to shift their time allocation, perhaps an hour or two per week, from work to school. If true, we might expect to see increases not just in credits but in GPA.

[19] Though we cannot confirm it in our sample because we only have one year of aid data, Marx and Turner (2015) find no discontinuities in subsequent years' Pell Grants for students around the EFC cutoff in a given year.

[20] When we focus on the cohort most likely to have been eligible for summer pell, the negative effect is no smaller.

While effects on cumulative GPA were in a positive direction (between 0.06 to 0.08 points), they were not statistically significant (though very close by the end of our follow-up period).

Effects on cumulative credits earned, degree completion, and transfer, measured three years after entry, were generally in a positive direction and of a magnitude consistent with the positive effects observed in the time periods closest to the treatment. However, we do not have power to detect small effects on these distal outcomes, and it may simply be unrealistic to expect to see anything other than small effects given the treatment, which amounts to replacing $500 in loans with $500 in grants. In some respects, it might be considered surprising to find any effects of such a modest treatment.

## C Sensitivity checks

### Optimal Bandwidth

Tables 3.2 and 3.3 assess the sensitivity of our RD estimators using bandwidths of 1/2 and two times our baseline bandwidth of ±2000 ($pm1000$ and ±4000, respectively). The general pattern and sign of our main results holds across different bandwidths; however, both magnitude and significance level fluctuates. For the wide bandwidth, coefficients are generally smaller. We also calculated optimal bandwidths under three different methods - cross-validation, IK, and CCT - separately for each outcome considered (see Appendix Table A2 for a summary of these results).[21] Across outcomes, the average bandwidth suggested by cross-validation and IK is around ±4000, while CCT suggests ±1366. Our baseline ±2000 bandwidth lies at the lower end for cross-validation and IK but at the upper end for CCT. Given these results, we think our baseline bandwidth of ±2000 bandwidth is reasonable.

[21] For implementation, we use rdbwselect() function in Stata rdrobust package provided by Calonico et al. (2014) under 2014 released version. Note that this package is upgraded in 2016. We use the older version of that directly estimates across three methods.

### Degree of Polynomial

Misspecification of functional form can generate bias in our treatment estimator when calculating using linear regression (Lee & Lemieux, 2010). Thus, the last column of Tables 3.2 and 3.3 also provide results using a quadratic specification (with our widest bandwidth). Again, the overall pattern of results is similar to baseline but magnitudes shift and here we see some negative results (on spring/summer enrollment in Year 1) become significant. To explore optimal degree of polynomial, we conduct a degree of polynomial test following Lee and Lemieux (2010) non-parametric approach by adding bin dummies to the polynomial regression and testing for joint significance of the bin dummies (equivalent to an F-test using R-square from with and without the bin dummies regression, see Appendix Table A3 for full results).[22] For each outcome, polynomial degree is determined by the degree where adding a higher order term no longer makes the bin dummies jointly significant. In some cases, bin dummies remain significant even regardless of the order of polynomial.[23] However, for variables where functional form does matter, a linear specification (polynomial of degree one) is generally supported.

## D Addressing Sample Selection Bias

### Limit analysis to subgroup where no discontinuity is present

We first use a subgroup selection method introduced by Calcagno and Long (2008) to address the problem of discontinuous density in a different RD setting. Calcagno and Long (2008), examine the impact of a test-score based assignment to remediation, and find discontinuities in the density of observations around the cutoff at some institutions in their sample, but not others. They conduct a separate McCrary test for each institution and select only

[22] Lee and Lemieux (2010) also uses Akaike information criterion (AIC) model selection for selection of degree of polynomial, however, recommends non-parametric F-test because lack of visibility to compare across different models (see Lee & Lemieux (2010) (pg. 326)).

[23] We run this test including up to polynomial degree 6. There are no major changes by added extra two degrees. Ask authors for results for all polynomials up to degree 6.

a subset of institutions with smooth densities for further analysis. When we follow a parallel approach, we find that nine smaller institutions exhibit no discontinuity in enrollments around the Pell grant cutoff, while three large institutions do. Hereafter, we refer to the former group of institutions as the continuous group, and the latter as the non-continuous group.

Table 3.4, which examines how these two groups of institutions differ, is revealing in itself. Table 3.4 compares characteristics across two subgroups, continuous and non-continuous institutions. Initially, we look at averages of pre-treatment covariates for all of our 2008-2010 cohorts. Students at non-continuous schools have more students of color (Black, Hispanic, and Asians) and substantially less white. Non-continuous schools have more students who took remedial tests and have slightly higher writing and math scores, on average.[24] Exploring counts and distance of local schools, we find striking differences between the two groups. On average, continuous schools have no community colleges, 0.4 four-year schools, and 1.8 for-profit institutions within 10 miles. Schools with discontinuous enrollment around the Pell cutoff also have no community colleges, but more four-year schools and many more for-profit schools within 10 miles (1.7 and 12.7, respectively). On average, a student at one of these schools is only about three miles away from either a four-year or a for-profit institution, while at continuous schools the nearest alternatives in these sectors are about 20 miles away. (As one might expect, non-continuous schools are located in more urban areas.) The large difference in nearby for-profit alternatives, in particular, suggests that perhaps the missing students who are eligible for Pell may have switched their enrollment to attend at for-profit schools instead of community colleges. This would be consistent with Cellini's (2010) finding that increases in Pell awards increased enrollment at for-profit colleges. It is also possible, however, that students are using the Pell grant to attend four-year colleges as well.

Unfortunately, the large differences in demographics across the two groups of institu-

[24] One relatively large school from the continuous group has essentially zero remedial test take-up rate, which seems to drive the average down for the continuous group.

tions makes any differences in impacts hard to interpret. While it would be reassuring if our analyses held up within our subset of continuous-density schools, if they do not, it is not clear whether this indicates that our results are driven by selection, or simply that Pell grants have heterogeneous effects for different student populations. Nonetheless, we present our results for these two subsets of schools separately in Table 3.5. Our estimated effects on financial aid packages (top four rows) are consistent with our main results in Table 3.2. However, for academic and labor market outcomes, we see distinctive regression results between continuous and non-continuous density groups. The general pattern is that few results are significant within the continuous group and some outcomes even have the opposite sign. The positive results that we observe in our main results appear concentrated within the three large institutions with non-continuous density around the Pell cutoff. The fact that results are concentrated in the group where selection bias is most severe is not reassuring, but for the reasons explained above, neither is it definitive. The two groups are very demographically different and it is possible that effect of Pell Grant is larger for younger, non-white students with higher test scores.

### Bounding Analysis

Another way to account for potential selection bias is to bound our estimates as introduced by Gerard et al. (2015). GRR introduces a way to identify partial treatment effects through estimating upper/lower bounds by making worst/best assumptions about the missing population.[25] For further details about this methodology, see the Appendix. GRR define ”selectors” as those individuals, in this context, whose enrollment decision is influenced by whether or not they fall above or below the Pell cutoff. In this case, the selectors who fall below the cutoff, and hence qualify for Pell, are unobserved. Above the cutoff are a mix of non-selectors and selectors who would have enrolled elsewhere had they qualified for Pell.

[25] GRR bounding exercise is an extension to Lee (2009) bounding exercise in the Sharp RD case. GRR requires two additional assumptions regarding what they call the ”selectors” (those students whose enrollment decisions shift as a result of their Pell eligibility): that the direction of selection is one-sided and that the conditional density is left-differentiable.

The goal of the GRR method is to estimate upper and lower bounds of the effects for only non-selectors by trimming the mixed side (in this case, above the cutoff which includes both selectors and non-selectors) of the estimated proportion of selectors.

We first estimate the proportion of selectors ($\tau$) by calculating the jump in enrollment at the cutoff from height of the density curve using local polynomial smoothing with rectangular kernel (and degree 1 polynomial). Second, assuming selectors have the best (worst) observed outcomes, the upper (lower) bound is estimated by the difference in expectation of outcome between the left and right side of the cutoff, where the side with more observations has been trimmed of observations below (above) the $\tau$ (or,respectively,$1-\tau$) quantile. See appendix for further details.

We perform two versions of this bounding analysis. First, we trim separately based on for each individual outcome, as indicated by the GRR method. This produces the widest bounds, but is overly conservative in practice because different individuals are trimmed from the sample for each outcome (it is not the case that the best students on one outcome are the best students on all outcomes). So, as an alternative, we also examine results when we trim the sample just once, based on cumulative GPA in the first semester of first year, and then calculate bounds on all outcomes using that same sample.

Table 3.6 reproduces our baseline regression estimates ($\pm 2000$ bandwidth including covariate controls), and then shows the results from these two versions of our bounding analysis. As expected, the GRR bounds in column 2 (in which the sample is trimmed separately for each sample) are very wide. In column 3, we tighten our bounds by trimming only once, based on a single outcome variable, then calculating bounds on different outcome variables using that same trimmed sample. We choose cumulative GPA in the fall semester of entrance to college, under the logic that whatever are the unobservable factors that influence enrollment decisions (e.g., student motivation) may correlate with academic performance as observed after enrollment. Our bounding results (column 3) are tighter with more zero-excluding bounds (indicated in bold brackets). Effects of Pell eligibility on financial aid packaging holds with all zero-excluding bounds. The bounds on full-time enrollment still fail to exclude zero, but is shifted towards more positive impacts. Academic earnings and summer earnings in both year 1 and year 2 remain negative with bounds that exclude zero.

# VI Discussion

In this paper, we examine the effect of being eligible for Pell on financial aid packages, student outcome, and labor supply among those who are around the Pell Grant eligibility cutoff. First, we find that even at community colleges which have relatively little institutional aid to distribute, non-Pell aid awards are influenced by differences in Pell eligibility. Moreover, the pattern of response is distinctive depending on whether an institution offers federal student loans: for schools that offer loans, students who just miss qualifying for Pell borrowed more (almost equivalent to Pell eligibility at the cutoff), such that students just above and below the Pell cutoff received similar amounts of aid in total. For schools that do not offer loans, students who don't qualify for Pell receive higher state grants to compensate.

We next examine the effect of receiving a modest Pell grant (instead of loans) for students attending loan-offering schools. We find that students who just barely qualify for Pell are more likely to enroll full-time (about 4-7 percentage points more likely, depending upon the term) and at the same time reduce their labor supply by about $12-20 per week. These patterns are consistent with a story in which Pell allows students to shift their time allocation, perhaps an hour or two per week, from work to school.

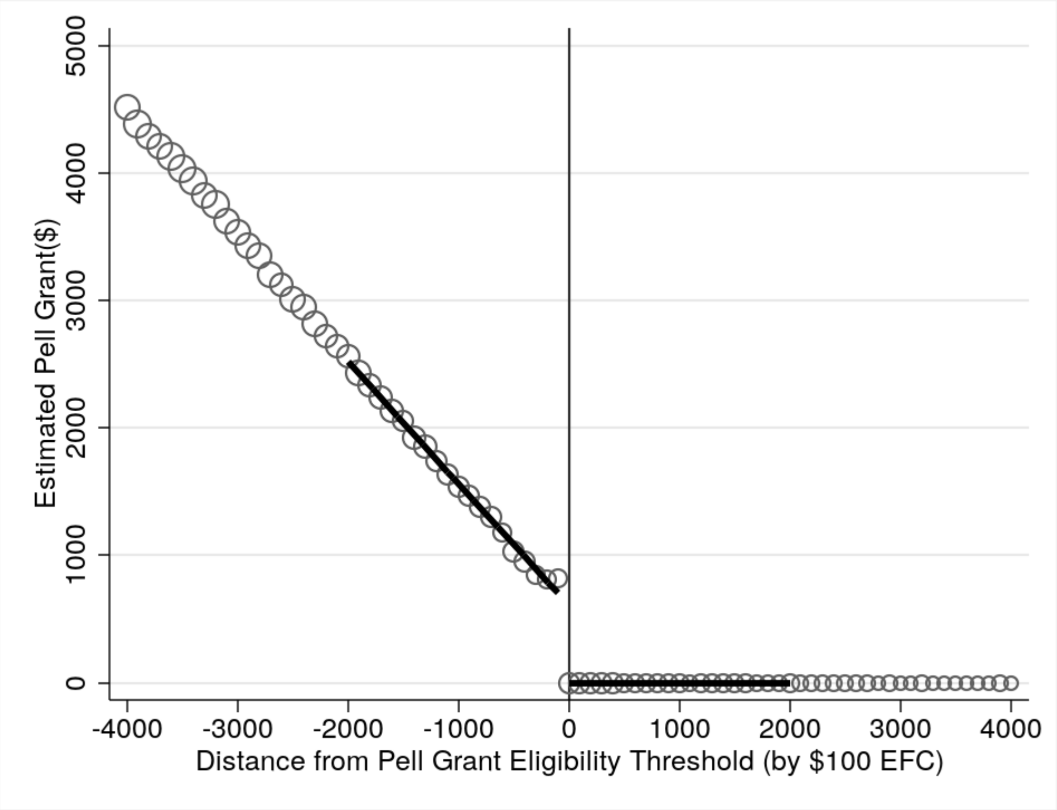
We also find a discontinuity in enrollments around the Pell cutoff (within loan-offering schools), which suggests that Pell eligibility may independently affect enrollment decisions as well. We find that this discontinuity in enrollments is concentrated at three large urban community colleges, which have a lot of local market competition, particularly from for-profit institutions.

Unfortunately, this pattern of enrollments may introduce bias into our regression discontinuity estimates. To examine this, we follow two methods in the literature: re-estimating impacts only for the subset of schools with continuous density through the cutoff, and a bounding analysis that makes extreme assumption about the missing population. In both cases, our results are not entirely robust. While this is not reassuring, neither does it provide affirmative evidence that our main results are biased. Our best guess regarding the likely effects of receiving a modest Pell, in comparison to an equivalent amount of additional loans, is still drawn from our main results in Tables 3.2 and 3.3, which control for a rich set of observable characteristics at entry. Still, the lack of robustness suggests these results should be interpreted cautiously and alongside evidence from other studies.

Our research has two implications. First, even at community colleges, which typically have very little ”institutional” aid to distribute, institutions may have discretion to determine how Pell interacts with other state and federal aid programs. In our sample, we find a complex web of interactions, with state grants smoothing over the discontinuity in the Pell schedule at no-loan schools, and loans smoothing over the discontinuity at loan schools. Second, although the resulting treatment is relatively small-essentially implying a shift of $500 from loans to grants-we nonetheless find some evidence that this alters some student behaviors. Students who are just below the cutoff (receiving Pell) seem to shift their time allocation, reducing work while increasing their enrollment intensity; we find significant increases in full-time enrollment and suggestive (but not significant) evidence of increases in GPAs. Moreover, we find indirect evidence that Pell eligibility may alter students' initial enrollment choices: students just barely eligible for Pell are less likely to show up in our sample of community college enrollees.

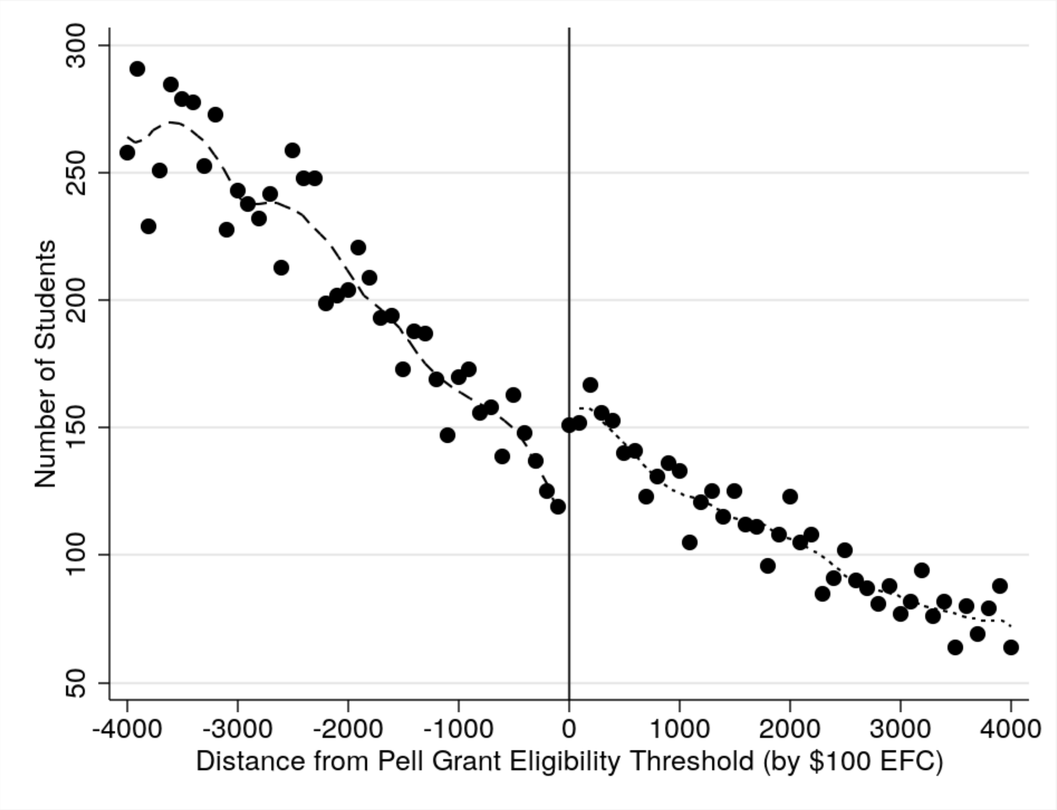
# VII Figures

Figure 3.1: Estimated Pell Grant by EFC (2008-10 Cohort)



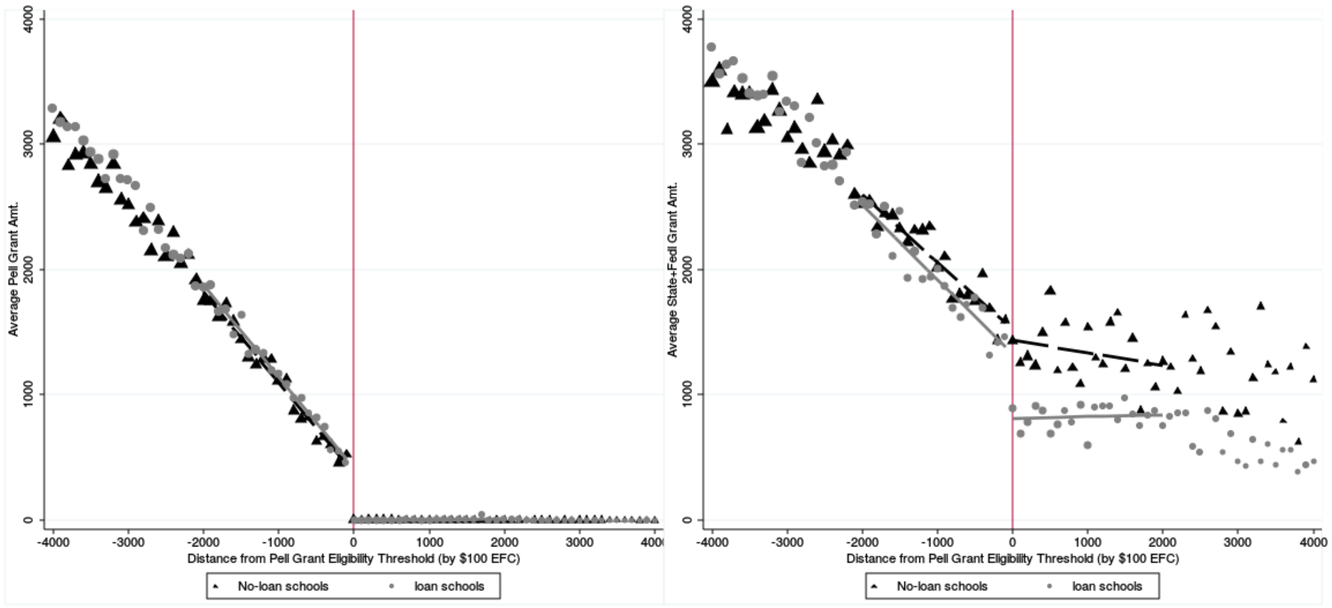
*Note.* Samples are restricted to 2008-2010 cohort students who have filed FAFSA, for whom race/ethnicity is not missing, and who are non-dual enrollees. Estimated Pell amount is computed by EFC assuming full-time enrollment intensity. Each point is a mean value of the outcome that falls within a bin of size $100 EFC. Graph shows only points that fall within the ±$4, 000 bandwidth. Gray line is a fitted line of mean points within a ±$2, 000 bandwidth.

Figure 3.2: Density Plot for All Schools



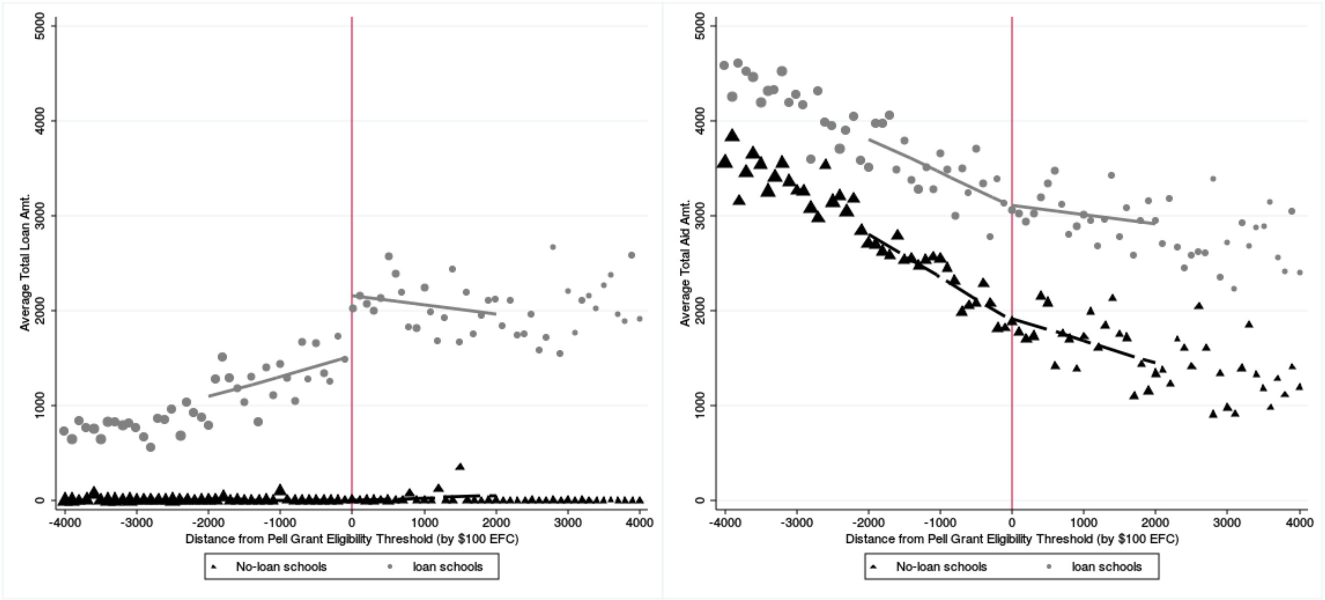
*Note.* Samples are restricted to 2008-2010 cohort students who have filed FAFSA, for whom race/ethnicity is not missing, and who are non-dual enrollees. Points represent number of students (sum count) that fall within a bin of size $100 EFC. Points within a ±$4,000 bandwidth are included in the figure. Gray line is a local smoothed polynomial line with degree 2, using points within the ±$4,000 bandwidth.

Figure 3.3: Grant Amounts ($) for Loan and No-Loan Schools



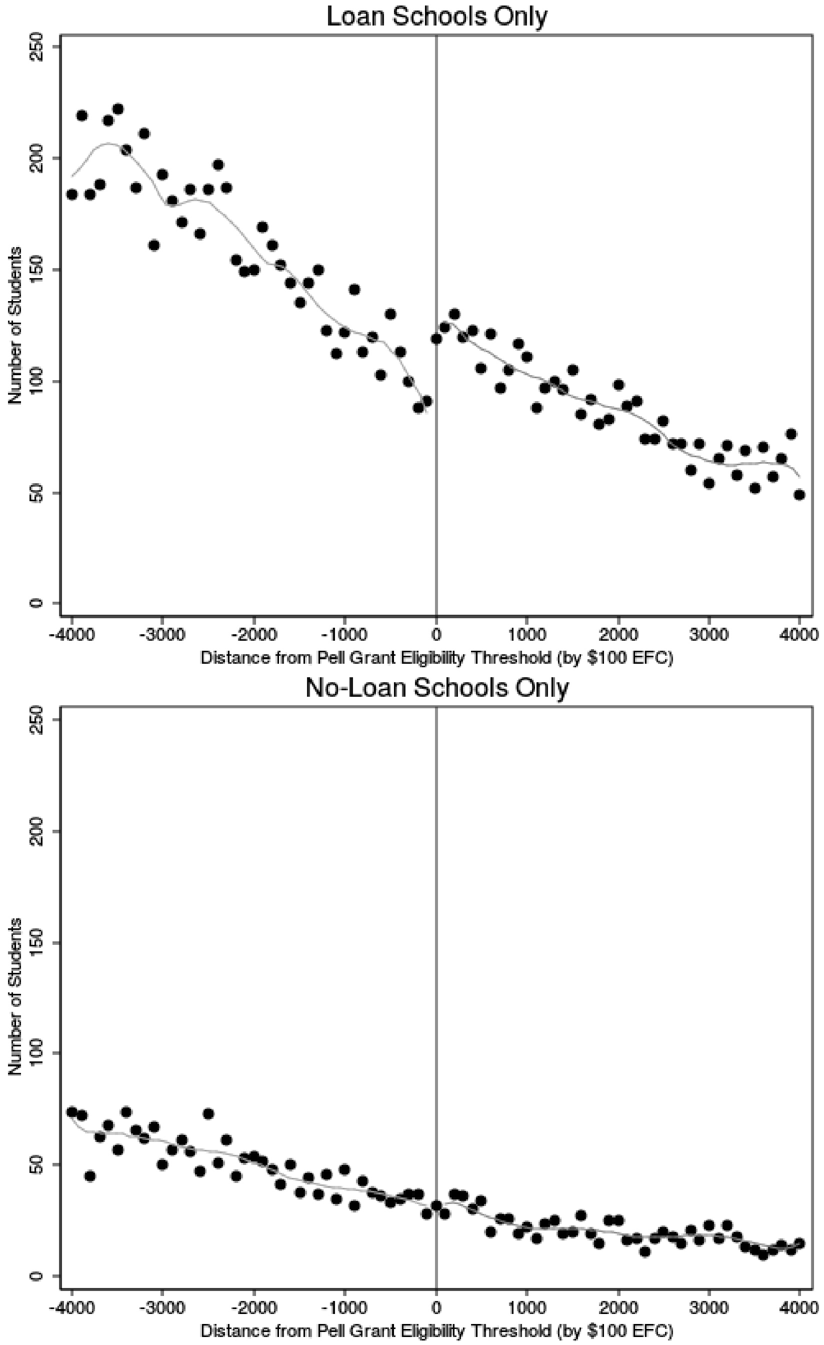
*Note.* Samples are restricted to 2008-2010 cohort students who have filed FAFSA, for whom race/ethnicity is not missing, and who are non-dual enrollees. Averages are plotted separately for loan schools (triangle points) and no-loan schools (circle points). Each point represents mean outcomes for students that fall within a bin of size $100 EFC. Only points within a ±$4,000 bandwidth are in the figure. Gray solid (loan schools) and black dashed (no-loan schools) lines are the linear fitted value of these points that fall within the ±$2,000 bandwidth.

Figure 3.4: Loan and Total Aid Amounts ($) for Loan and No-Loan Schools



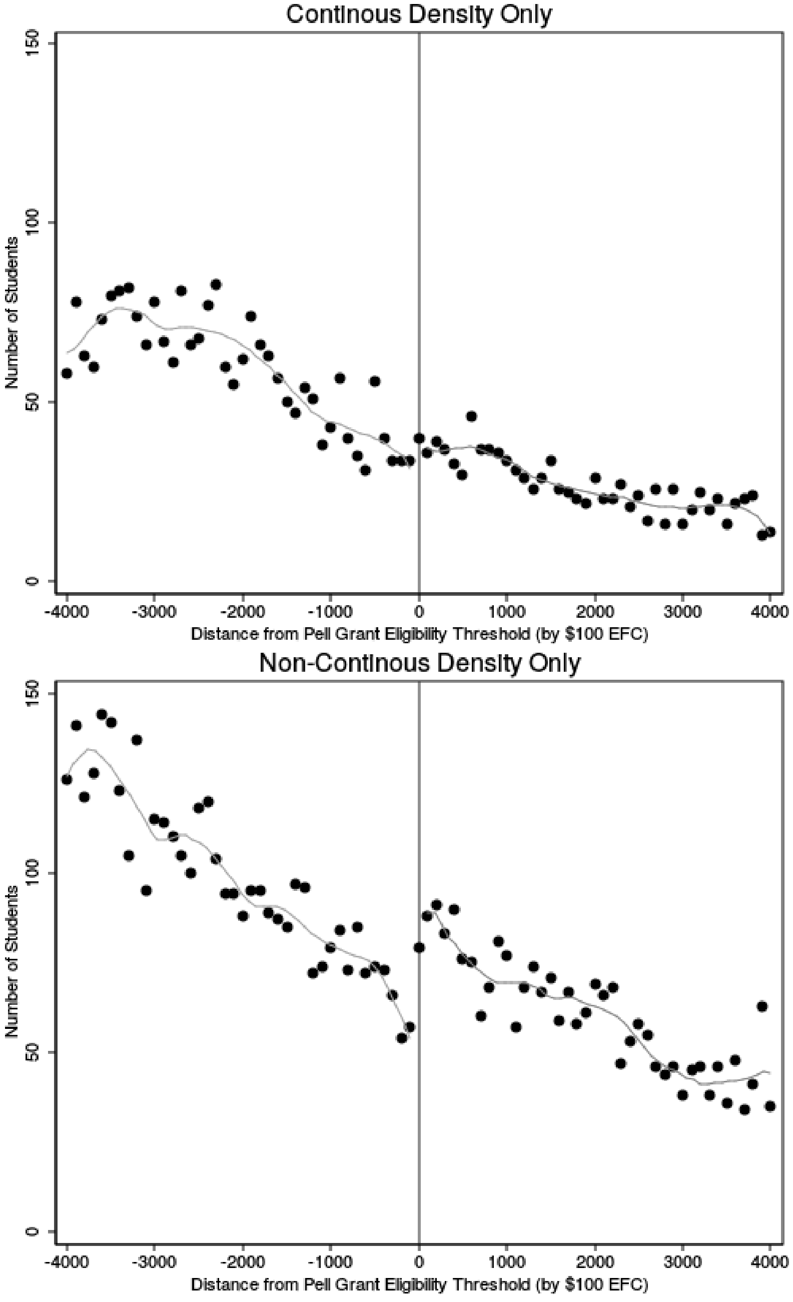
*Note.* Samples are restricted to 2008-2010 cohort students who have filed FAFSA, for whom race/ethnicity is not missing, and who are non-dual enrollees. Averages are plotted separately for loan schools (triangle points) and no-loan schools (circle points). Each point represents mean outcomes for students that fall within a bin of size $100 EFC. Only points within a $\pm\$4,000$ bandwidth are in the figure. Gray solid (loan schools) and black dashed (no-loan schools) lines are the linear fitted value of these points that fall within the $\pm\$2,000$ bandwidth.

Figure 3.5: Density Plot for Loan Schools (Top) and No-Loan Schools (Bottom)



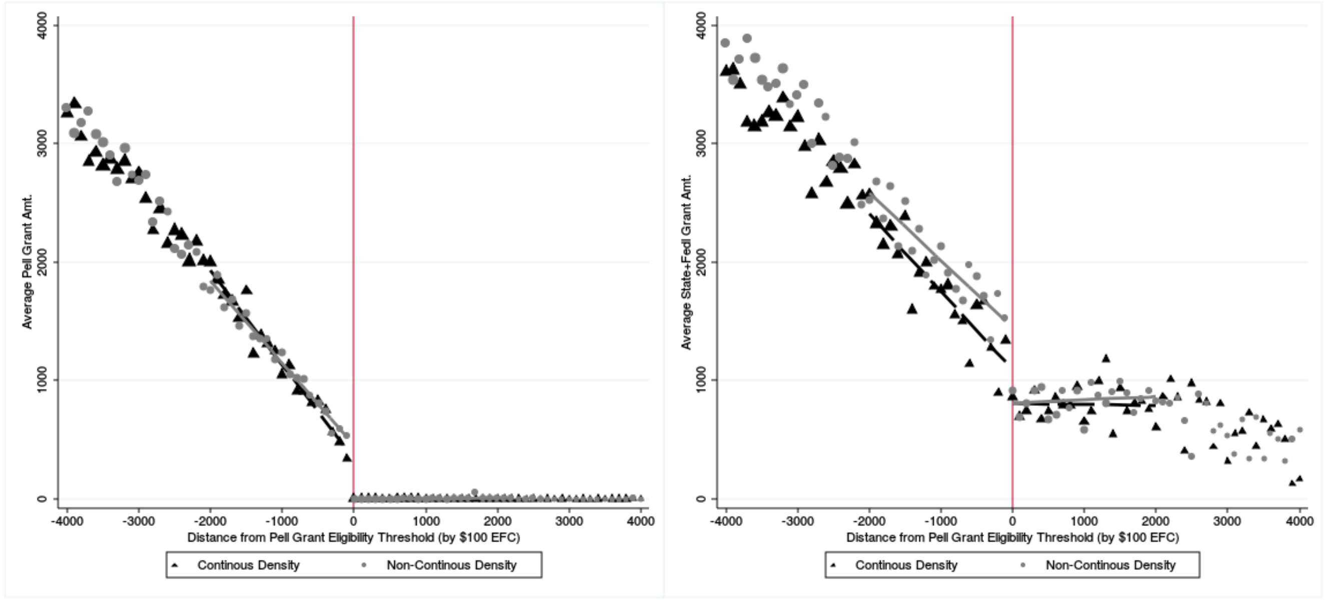
*Note.* Samples are restricted to 2008-2010 cohort students who have filed FAFSA, for whom race/ethnicity is not missing, who are non-dual enrollees, and only for students attending loan schools (top) or no-loan schools (bottom). Points represent number of students (sum count) that fall within a bin of size $100 EFC. Points within a ±$4,000 bandwidth are included in the figure. Gray line is a local smoothed polynomial line with degree 2, using points within the ±$4,000 bandwidth.

Figure 3.6: Density Plot for Continuous Schools (Top) and Non-Continuous Schools (Bottom)



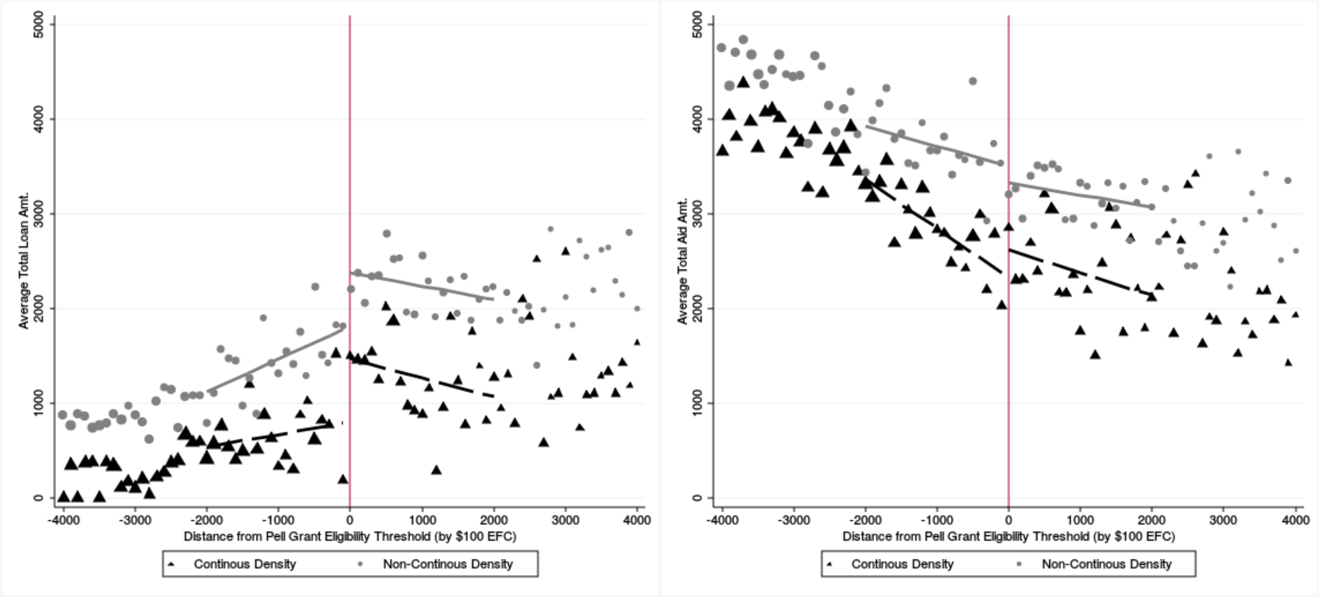
*Note.* Samples are restricted to 2008-2010 cohort students who have filed FAFSA, for whom race/ethnicity is not missing, who are non-dual enrollees, and who are attending loan schools. Points represent number of students (sum count) that fall within a bin of size $100 EFC. Points within ±$4,000 bandwidth are included in the figure. Gray line is a local smoothed polynomial line with degree 2, using points within the ±$4,000 bandwidth.

Figure 3.7: Grant Amounts ($) for Continuous and Non-Continuous Schools



*Note.* Samples are restricted to 2008-2010 cohort students who have filed FAFSA, for whom race/ethnicity is not missing, and who are non-dual enrollees. Averages are plotted separately for continuous schools (triangle points) and non-continuous schools (circle points). Each point represents mean outcomes for students that fall within a bin of size $100 EFC. Only points within a ±$4,000 bandwidth are in the figure. Gray solid (continuous schools) and black dashed (non-continuous schools) lines are the linear fitted value of these points that fall within a ±$2,000 bandwidth.

Figure 3.8: Loan and Total Aid Amounts ($) for Continuous and Non-Continuous Schools



*Note.* Samples are restricted to 2008-2010 cohort students who have filed FAFSA, for whom race/ethnicity is not missing, and who are non-dual enrollees. Averages are plotted separately for continuous schools (triangle points) and non-continuous schools (circle points). Each point represents mean outcomes for students that fall within a bin of size $100 EFC. Only points within a ±$4,000 bandwidth are in the figure. Gray solid (continuous schools) and black dashed (non-continuous schools) lines are the linear fitted value of these points that fall within a ±$2,000 bandwidth.

# VIII Tables

Table 3.2: RD Estimates of Effect of Pell Eligibility on Composition of Financial Aid Packages

| | Mean Outcomes | (1) Basic 2000bw. | | | (2) Without cov. | | | (3) 1000bw. | | | (4) 4000bw. | | | (5) 4000bw, Quadratic | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Outcome | Just Above Cutoff | Coef. | (S.E.) | | Coef. | (S.E.) | | Coef. | (S.E.) | | Coef. | (S.E.) | | Coef. | (S.E.) | |
| *Institutions Offering Federal Loans* | | | | | | | | | | | | | | | | |
| Amount of Pell received | $0 | $459 | (17) | *** | $467 | (17) | *** | $442 | (20) | *** | $445 | (16) | *** | $426 | (22) | *** |
| Amount of Pell+State grants received | $869 | $560 | (64) | *** | $574 | (66) | *** | $598 | (92) | *** | $413 | (46) | *** | $513 | (69) | *** |
| Amount of loans received | $1,953 | -$592 | (113) | *** | -$639 | (118) | *** | -$534 | (163) | *** | -$451 | (79) | *** | -$574 | (120) | *** |
| Amount of total aid received | $2,993 | $89 | (129) | | $56 | (131) | | $201 | (185) | | $107 | (91) | | $33 | (137) | |
| Sample size | 1,421 | 5,753 | | | 5,753 | | | 2,828 | | | 11,944 | | | 11,944 | | |
| *Institutions Not Offering Federal Loans* | | | | | | | | | | | | | | | | |
| Amount of Pell received | $0 | $434 | (25) | *** | $427 | (25) | *** | $409 | (32) | *** | $481 | (25) | *** | $435 | (34) | *** |
| Amount of Pell+State grants received | $1,640 | $132 | (105) | | $97 | (118) | | $66 | (151) | | $203 | (77) | *** | $90 | (116) | |
| Amount of loans received | $4 | $3 | (8) | | $6 | (8) | | -$4 | (11) | | -$9 | (9) | | $7 | (7) | |
| Amount of total aid received | $2,044 | $153 | (123) | | $113 | (136) | | -$40 | (180) | | $268 | (89) | *** | $100 | (136) | |
| Sample size | 456 | 2,102 | | | 2,102 | | | 1,048 | | | 4,421 | | | 4,421 | | |

*Note.* Samples are restricted to students in the 2008-2010 fall entry cohorts who filed FAFSA and for whom race/ethnicity is not missing. Top panel estimates use only loan schools and bottom panel estimates use only no-loan schools. Coefficients indicate beta values for indicator of treatment status (i.e., 1 if eligible for Pell and 0 otherwise). Huber-White robust standard errors are in parentheses. Columns 1 and 2 are for samples within ±$2,000 bandwidth, column 3 for ±$1,000 bandwidth, columns 4 and 5 for ±$4,000 bandwidth. All specifications control for cohort fixed effects. All columns except column 2 control for covariates-female, Black, Hispanic, Asian, American-Indian, age, income, dependent, dual enrollment, reading, writing, math score prior to entry, and flags on whether they have these test scores-and college fixed effects. All columns except for column 5 use local linear polynomial regression, while column 5 uses quadratic polynomial specification. Rectangular kernel is used in all specifications. ***$p$ ¡ .01. **$p$ ¡ .05. *$p$ ¡ .1.

Table 3.3: RD Estimates of Effect of Pell Eligibility on Academic Outcomes and Student Labor Supply (Loan Schools)

| | Mean Outcomes | (1) Basic 2000bw. | | | (2) Without cov. | | | (3) 1000bw. | | | (4) 4000bw. | | | (5) 4000bw, Quadratic | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Outcome | Just Above Cutoff | Coef. | (S.E.) | | Coef. | (S.E.) | | Coef. | (S.E.) | | Coef. | (S.E.) | | Coef. | (S.E.) | |
| *Year 1 Outcomes* | | | | | | | | | | | | | | | | |
| Enrolled full-time, Year 1 Fall | 0.657 | 0.020 | (0.024) | | 0.030 | (0.025) | | -0.003 | (0.034) | | 0.002 | (0.017) | | 0.008 | (0.025) | |
| Re-enrolled, Year 1 Spring | 0.842 | -0.016 | (0.020) | | -0.012 | (0.020) | | -0.028 | (0.029) | | -0.001 | (0.014) | | -0.048 | (0.021) | ** |
| Enrolled full-time, Year 1 Spring | 0.520 | 0.048 | (0.026) | * | 0.058 | (0.027) | ** | 0.026 | (0.037) | | 0.034 | (0.018) | * | 0.019 | (0.027) | |
| Enrolled, Year 1 Summer | 0.290 | -0.046 | (0.023) | ** | -0.049 | (0.024) | ** | -0.091 | (0.033) | *** | -0.022 | (0.017) | | -0.065 | (0.025) | *** |
| Cum. GPA, End of Year | 2.473 | 0.061 | (0.056) | | 0.060 | (0.060) | | -0.027 | (0.081) | | 0.007 | (0.040) | | 0.044 | (0.060) | |
| Cum. Credits Completed, End of Year | 17.462 | 0.480 | (0.559) | | 0.628 | (0.589) | | -0.647 | (0.808) | | -0.102 | (0.390) | | -0.312 | (0.591) | |
| Cum. Year 1 earnings (Q4-Q3) | $4,873 | -$806 | (393) | ** | -$911 | (406) | ** | -$710 | (538) | | -$727 | (282) | ** | -$616 | (420) | |
| *Year 2 Outcomes* | | | | | | | | | | | | | | | | |
| Re-enrolled, Year 2 Fall | 0.616 | 0.003 | (0.026) | | 0.005 | (0.026) | | -0.014 | (0.037) | | 0.014 | (0.018) | | -0.015 | (0.027) | |
| Enrolled full-time, Year 2 Fall | 0.371 | 0.074 | (0.026) | *** | 0.079 | (0.026) | *** | 0.065 | (0.036) | * | 0.036 | (0.018) | ** | 0.047 | (0.027) | * |
| Re-enrolled, Year 2 Spring | 0.580 | 0.005 | (0.026) | | 0.005 | (0.026) | | 0.034 | (0.037) | | -0.003 | (0.018) | | -0.002 | (0.028) | |
| Enrolled full-time, Year 2 Spring | 0.328 | 0.044 | (0.025) | * | 0.047 | (0.025) | * | 0.044 | (0.036) | | 0.021 | (0.017) | | 0.026 | (0.026) | |
| Enrolled, Year 2 Summer | 0.226 | -0.004 | (0.022) | | -0.004 | (0.022) | | 0.013 | (0.031) | | -0.011 | (0.015) | | 0.005 | (0.023) | |
| Cum. GPA, End of Year | 2.401 | 0.074 | (0.053) | | 0.084 | (0.057) | | 0.027 | (0.075) | | 0.022 | (0.037) | | 0.051 | (0.056) | |
| Cum. Credits Completed, End of Year | 29.075 | 1.243 | (1.063) | | 1.463 | (1.123) | | -0.033 | (1.530) | | 0.011 | (0.745) | | -0.202 | (1.127) | |
| Cum. Year 2 earnings (Q4-Q3) | $5,323 | -$534 | (445) | | -$627 | (455) | | -$552 | (620) | | -$364 | (318) | | -$385 | (475) | |
| SAP: Earned 2.0+ GPA, Year 1 | 0.690 | -0.005 | (0.024) | | 0.001 | (0.025) | | -0.039 | (0.035) | | -0.007 | (0.017) | | -0.016 | (0.026) | |
| SAP: Earned 67%+ of credits attempted, Year 1 | 0.626 | 0.039 | (0.025) | | 0.040 | (0.026) | | 0.003 | (0.036) | | 0.001 | (0.018) | | 0.017 | (0.027) | |
| SAP: Earned 2.0+ GPA, Year 2 | 0.703 | 0.008 | (0.024) | | 0.014 | (0.025) | | 0.002 | (0.034) | | 0.001 | (0.017) | | -0.003 | (0.025) | |
| SAP: Earned 67%+ of credits attempted, Year 2 | 0.594 | 0.060 | (0.025) | ** | 0.060 | (0.026) | ** | 0.026 | (0.036) | | 0.014 | (0.018) | | 0.040 | (0.027) | |
| SAP: Earned 2.0+ GPA, Year 3 | 0.711 | 0.013 | (0.023) | | 0.019 | (0.024) | | 0.020 | (0.033) | | 0.006 | (0.016) | | -0.002 | (0.025) | |
| SAP: Earned 67%+ of credits attempted, Year 3 | 0.591 | 0.065 | (0.025) | *** | 0.065 | (0.026) | ** | 0.023 | (0.036) | | 0.016 | (0.018) | | 0.035 | (0.027) | |
| *End of Year 3 Attainment Outcomes* | | | | | | | | | | | | | | | | |
| Cum. GPA | 2.392 | 0.084 | (0.052) | | 0.097 | (0.056) | * | 0.046 | (0.074) | | 0.019 | (0.036) | | 0.059 | (0.055) | |
| Cum. credits earned | 35.205 | 1.741 | (1.342) | | 1.937 | (1.406) | | 0.946 | (1.935) | | -0.041 | (0.940) | | 0.312 | (1.423) | |
| Ever transferred to 4-Yr | 0.215 | 0.026 | (0.021) | | 0.028 | (0.022) | | -0.002 | (0.031) | | 0.008 | (0.015) | | 0.005 | (0.023) | |
| Earned any degree/cert | 0.206 | 0.010 | (0.021) | | 0.016 | (0.022) | | 0.000 | (0.030) | | -0.011 | (0.014) | | -0.012 | (0.022) | |
| Earned any degree/cert or transferred | 0.317 | 0.026 | (0.024) | | 0.032 | (0.025) | | 0.005 | (0.035) | | -0.003 | (0.017) | | 0.002 | (0.025) | |
| Sample size | 1,421 | 5,753 | | | 5,753 | | | 2,828 | | | 11,944 | | | 11,944 | | |

*Note.* Samples are restricted to students in the 2008-2010 fall entry cohorts who filed FAFSA and for whom race/ethnicity is not missing. Top panel estimates use only loan schools and bottom panel estimates use only no-loan schools. Coefficients indicate beta values for indicator of treatment status (i.e., 1 if eligible for Pell and 0 otherwise). Huber-White robust standard errors are in parentheses. Columns 1 and 2 are for samples within $\pm\$2,000$ bandwidth, column 3 for $\pm\$1,000$ bandwidth, columns 4 and 5 for $\pm\$4,000$ bandwidth. All specifications control for cohort fixed effects. All columns except column 2 control for covariates-female, Black, Hispanic, Asian, American-Indian, age, income, dependent, dual enrollment, reading, writing, math score prior to entry, and flags on whether they have these test scores-and college fixed effects. All columns except for column 5 use local linear polynomial regression, while column 5 uses quadratic polynomial specification. Rectangular kernel is used in all specifications. ***$p$ ¡ .01. **$p$ ¡ .05. *$p$ ¡ .1.

Table 3.4: RD Estimates of Effect of Pell Eligibility on Academic Outcomes and Student Labor Supply (Loan Schools)

| Outcome | Continous Schools<br>Mean | Non-Continous Schools<br>Mean |
|---|---|---|
| Female (%) | 0.528 | 0.511 |
| Black (%) | 0.223 | 0.285 |
| Hispanic (%) | 0.031 | 0.120 |
| Asian (%) | 0.024 | 0.107 |
| White(%) | 0.717 | 0.481 |
| American Indian (%) | 0.006 | 0.006 |
| Age | 21.616 | 21.601 |
| Dual Enrollment | 0.278 | 0.055 |
| Income | $38,752 | $44,754 |
| Depend | 0.688 | 0.692 |
| Has Remedial Reading (%) | 0.536 | 0.680 |
| Has Remedial Writing (%) | 0.545 | 0.688 |
| Has Remedial Math (%) | 0.387 | 0.614 |
| Remedial Reading placement score | 81.553 | 81.653 |
| Remedial Writing placement score | 69.049 | 71.703 |
| Remedial Math placement score | 34.190 | 36.007 |
| Prior Credits Attempted | 3.864 | 1.249 |
| Prior Credits Earned | 3.507 | 1.040 |
| Prior Year Earnings (Q3-Q4-Q1-Q2) | $2,921 | $2,597 |
| Sample Size | 24,321 | 43,221 |
| | | |
| Local Market | | |
| Avg. Number of nearby 2-year public schools (N) | 0.0 | 0.0 |
| Avg. Distance to nearest 2-year school (miles) | 27.7 | 25.3 |
| Avg. Number of nearby 4-year schools (N) | 0.4 | 1.7 |
| Avg. Distance to the nearest 4-year school (miles) | 20.4 | 3.0 |
| Avg. Number of nearby for-profit schools (N) | 1.8 | 12.7 |
| Avg. Distance to nearest for-profit school (miles) | 18.2 | 2.5 |

*Note.* Source is College Scorecard Data (n.d.). Top panel: We take all samples from 2008-2010 cohorts and average the characteristics by whether student's school is in the non-continuous or continuous group. Bottom panel: We define nearby schools as those located within less than 10 miles from our sample schools. Distance is calculated using latitude and longitude coordinates. All local market variables are averages for schools in the non-continuous or continuous group.

Table 3.5: RD Estimates of Effect of Pell Eligibility on Academic Outcomes and Student Labor Supply (Loan Schools)

| | Continuous Density Schools | | | | Non-Continuous Density Schools | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) Mean Outcomes | (2) | (3) | (4) | (5) Mean Outcomes | (6) | (1) Basic 2000bw. | (7) | (8) |
| Outcome | Just Above Cutoff | Coef. | (S.E.) | | Just Above Cutoff | Coef. | (S.E.) | (S.E.) | |
| Amount of Pell received | $0 | $436 | (25) | *** | $0 | $485 | | (23) | *** |
| Amount of Pell+State grants received | $955 | $400 | (90) | *** | $810 | $697 | | (90) | *** |
| Amount of loans received | $1,500 | -$600 | (160) | *** | $2,263 | -$559 | | (159) | *** |
| Amount of total aid received | $2,664 | -$72 | (183) | | $3,217 | $245 | | (179) | |
| *Year 1 Outcomes* | | | | | | | | | |
| Enrolled full-time, Year 1 Fall | 0.683 | 0.022 | (0.035) | | 0.639 | 0.021 | | (0.032) | |
| Re-enrolled, Year 1 Spring | 0.837 | -0.052 | (0.031) | * | 0.845 | 0.017 | | (0.025) | |
| Enrolled full-time, Year 1 Spring | 0.542 | 0.012 | (0.039) | | 0.505 | 0.076 | | (0.034) | ** |
| Enrolled, Year 1 Summer | 0.284 | -0.066 | (0.035) | * | 0.294 | -0.034 | | (0.032) | |
| Cum. GPA, End of Year | 2.522 | 0.064 | (0.082) | | 2.438 | 0.059 | | (0.078) | |
| Cum. Credits Completed, End of Year | 18.075 | -0.532 | (0.864) | | 17.043 | 1.244 | | (0.728) | * |
| Cum. Year 1 earnings (Q4-Q3) | $4,643 | $38 | (558) | | $5,030 | -$1,269 | | (545) | ** |
| *Year 2 Outcomes* | | | | | | | | | |
| Re-enrolled, Year 2 Fall | 0.584 | -0.004 | (0.040) | | 0.637 | 0.014 | | (0.034) | |
| Enrolled full-time, Year 2 Fall | 0.367 | 0.045 | (0.039) | | 0.373 | 0.094 | | (0.034) | *** |
| Re-enrolled, Year 2 Spring | 0.537 | 0.018 | (0.040) | | 0.609 | -0.001 | | (0.034) | |
| Enrolled full-time, Year 2 Spring | 0.317 | 0.028 | (0.037) | | 0.335 | 0.055 | | (0.034) | |
| Enrolled, Year 2 Summer | 0.189 | 0.008 | (0.031) | | 0.251 | -0.008 | | (0.030) | |
| Cum. GPA, End of Year | 2.464 | 0.081 | (0.077) | | 2.357 | 0.069 | | (0.071) | |
| Cum. Credits Completed, End of Year | 29.140 | 0.148 | (1.608) | | 29.030 | 2.186 | | (1.412) | |
| Cum. Year 2 earnings (Q4-Q3) | $5,270 | $423 | (652) | | $5,359 | -$1,132 | | (607) | * |
| *End of Year 3 Attainment Outcomes* | | | | | | | | | |
| Cum. GPA | 2.454 | 0.076 | (0.077) | | 2.349 | 0.090 | | (0.070) | |
| Cum. credits earned | 34.133 | 1.095 | (1.986) | | 35.937 | 2.512 | | (1.814) | |
| Ever transferred to 4-Yr | 0.243 | -0.022 | (0.033) | | 0.197 | 0.056 | | (0.029) | ** |
| Earned any degree/cert | 0.255 | -0.009 | (0.033) | | 0.173 | 0.018 | | (0.027) | |
| Earned any degree/cert or transferred | 0.378 | -0.011 | (0.037) | | 0.275 | 0.045 | | (0.031) | |
| Sample size | 577 | 2,506 | | | 844 | | 3,247 | | |

*Note.* Samples are restricted to 2008-2010 fall entry cohort students who filed FAFSA, for whom race/ethnicity is not missing, and who are attending loan schools. Columns 1-4 further restrict to the subset of schools that has continuous density by McCrary (2008) test. Columns 5-8 restrict to the subset of schools that fails continuous density test by McCrary (2008). Coefficients indicate beta values for indicator of treatment status (i.e., 1 if eligible for Pell and 0 otherwise). Huber-White robust standard errors are in parentheses. Both regressions are within a ±$2,000 bandwidth except mean outcomes (columns 1 and 5) and control for cohort fixed effects for covariates—female, Black, Hispanic, Asian, American-Indian, age, income, dependent, dual enrollment, reading, writing, math score prior to entry, and flags on whether they have these test scores—and college fixed effects. Local linear polynomial is used with rectangular kernel in all specifications. ***$p$ ¡ .01. **$p$ ¡ .05. *$p$ ¡ .1.

Table 3.6: GRR Bounds on RD Estimates (2008-2010 Cohort, Loan Schools Only)

| | (1) | | (2) | | | (3) | |
|---|---|---|---|---|---|---|---|
| | Original Estimates | | Trim by each outcome | | | Trim by cum. GPA | |
| Outcome | Coef | (S.E.) | [low, | upper] | | [low, | upper] |
| Amount of Pell received | $459 | (17) | - | - | | - | - |
| Amount of Pell+State grants received | $560 | (64) | **[$236,** | **$1,143]** | * | **[$377,** | **$661]** |
| Amount of loans received | -$592 | (113) | [$-1,909, | $847] | | **[$-692,** | **$-442]** |
| Amount of total aid received | $89 | (129) | [$-1,176, | $1,599] | | **[$73,** | **$98]** |
| *Year 1 Outcomes* | | | | | | | |
| Enrolled full-time, Year 1 Fall | 0.020 | (0.024) | [-0.213, | 0.374] | | [-0.017, | 0.021] |
| Re-enrolled, Year 1 Spring | -0.016 | (0.020) | [-0.156, | 0.432] | | [-0.099, | 0.029] |
| Enrolled full-time, Year 1 Spring | 0.048 | (0.026) | [-0.248, | 0.339] | | [-0.061, | 0.085] |
| Enrolled, Year 1 Summer | -0.046 | (0.023) | [-0.546, | 0.041] | | **[-0.125,** | **-0.011]** |
| Cum. GPA, End of Year | 0.061 | (0.056) | [-0.569, | 0.577] | | [-0.516, | 0.511] |
| Cum. Credits Completed, End of Year | 0.480 | (0.559) | [-5.236, | 6.610] | | [-3.885, | 3.048] |
| Cum. Year 1 Earnings (Q4-Q3) | -$312 | (192) | [$-2,749, | $3,630] | | **[$-347,** | **$-121]** |
| *Year 2 Outcomes* | | | | | | | |
| Re-enrolled, Year 2 Fall | 0.003 | (0.026) | [-0.323, | 0.264] | | [-0.074, | 0.062] |
| Enrolled full-time, Year 2 Fall | 0.074 | (0.026) | [-0.343, | 0.244] | | [-0.017, | 0.121] |
| Re-enrolled, Year 2 Spring | 0.005 | (0.026) | [-0.394, | 0.193] | | [-0.091, | 0.067] |
| Enrolled full-time, Year 2 Spring | 0.044 | (0.025) | [-0.424, | 0.163] | | [-0.027, | 0.090] |
| Enrolled, Year 2 Summer | -0.004 | (0.022) | [-0.491, | 0.096] | | [-0.044, | 0.023] |
| Cum. GPA, End of Year | 0.074 | (0.053) | [-0.509, | 0.605] | | [-0.447, | 0.453] |
| Cum. Credits Completed, End of Year | 1.243 | (1.063) | [-9.128, | 13.378] | | [-6.046, | 5.703] |
| Cum. Year 2 Earnings (Q4-Q3) | -$281 | (224) | [$-2,865, | $4,477] | | **[$-381,** | **$-54]** |
| *End of Year 3 Attainment Outcomes* | | | | | | | |
| Cum. GPA | 0.084 | (0.052) | [-0.475, | 0.606] | | [-0.421, | 0.451] |
| Cum. credits earned | 1.741 | (1.342) | [-11.665, | 17.283] | | [-6.900, | 6.815] |
| Ever transferred to 4-Yr | 0.026 | (0.021) | [-0.393, | 0.194] | | [-0.048, | 0.094] |
| Earned any degree/cert | 0.010 | (0.021) | [-0.516, | 0.072] | | [-0.090, | 0.068] |
| Earned any degree/cert or transferred | 0.026 | (0.024) | [-0.406, | 0.181] | | [-0.091, | 0.108] |
| Sample Size | 5,753 | 5,753 | 4,576 | 4,576 | | 4,431 | 4,448 |

*Note.* Samples are restricted to students in 2008-2010 fall entry cohorts who filed FAFSA, for whom race/ethnicity is not missing, and who are attending loan schools. Column 1 is from Table 2 and Table 3. Columns 2 and 3 are bound estimates using GRR bounding exercise. Square brackets indicate lower and upper bounds of treatment effect after adjusting for sample selection bias. Column 2 trims and run a single regression separately for each outcome variable. Column 3 trims using a single variable, cumulative GPA fall semester of 1st year, and runs multiple regressions on different outcomes. All regressions are specified using local linear regression within ±$2,000 bandwidth with rectangular kernel, controls for cohort fixed effects, controls for covariates—female, Black, Hispanic, Asian, American-Indian, age, income, dependent, dual enrollment, reading, writing, math score prior to entry, and flags on whether they have these test scores—and controls for college fixed effects.

# Bibliography

[1] Baum, S., & Ma, J. (2011). *Trends in college pricing 2010.* New York, NY: The College Board.

[2] Benson, J., & Goldrick-Rab, S. (2011). *Putting college first: How social and financial capital impact labor market participation among low-income undergraduates.* Unpublished Manuscript.

[3] Bettinger, E. (2004). How financial aid affects persistence. *In College choices: The economics of where to go, when to go, and how to pay for it* (pp. 207-238). University of Chicago Press.

[4] Bettinger, E. P., Long, B. T., Oreopoulos, P., & Sanbonmatsu, L. (2012). The role of application assistance and information in college decisions: Results from the H&R Block FAFSA experiment. *The Quarterly Journal of Economics*, 127(3), 1205-1242.

[5] Bettinger, E., & Williams, B. (2013). Federal and state financial aid during the great recession. In *How the Financial Crisis and Great Recession Affected Higher Education* (pp. 235-262).

[6] Boadway, R., & Tremblay, J. F. (2012). Reassessment of the Tiebout model. *Journal of public economics*, 96(11), 1063-1078.

[7] Calcagno, J. C., & Long, B. T. (2008). *The impact of postsecondary remediation using a regression discontinuity approach: Addressing endogenous sorting and noncompliance* (No. w14194). National Bureau of Economic Research.

[8] Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2017). *rdro-*

*bust: Software for regression discontinuity designs.* Forthcoming in *Stata Journal.* Retrieved from http://www-personal.umich.edu/~cattaneo/papers/Calonico-Cattaneo-Farrell-Titiunik_2017_Stata.pdf

[9] Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust Nonparametric Confidence Intervals for Regression Discontinuity Designs. *Econometrica*,82(6), 2295-2326.

[10] Cellini, S. R. (2010). Financial aid and for-profit colleges: Does aid encourage entry?. *Journal of Policy Analysis and Management*, 29(3), 526-552.

[11] College Scorecard Data. (n.d.) *Data insights.* Retrieved from https://collegescorecard.ed.gov/data/

[12] Deming, D., & Dynarski, S. (2009). *Into college, out of poverty? Policies to increase the postsecondary attainment of the poor* (No. w15387). Cambridge, MA: National Bureau of Economic Research.

[13] Denning, J. T. (2016). *Born under a lucky star: Financial aid, college completion, labor supply, and credit constraints.* Unpublished manuscript presented at annual Association for Education Finance and Policy conference. Retrieved from https://aefpweb.org/sites/default/files/webform/41/BornUnderALuckyStar.pdf

[14] DesJardins, S. L., & McCall, B. P. (2008). *The impact of the gates millennium scholars program on the retention, college finance-and work-related choices, and future educational aspirations of low-income minority students.* Unpublished Manuscript.

[15] Dynarski, S. M., & Scott-Clayton, J. E. (2006). *The cost of complexity in federal student aid: Lessons from optimal tax theory and behavioral economics* (No. w12227). National Bureau of Economic Research.

[16] Dynarski, S., Scott-Clayton, J., & Wiederspan, M. (2013). Simplifying tax incentives and aid for college: Progress and prospects. *In Tax Policy and the Economy*, Volume 27 (pp. 161-201). Chicago, IL: University of Chicago Press.

[17] Dynarski, S., & Wiederspan, M. (2012). *Student aid simplification: Looking back and*

*looking ahead* (No. w17834). Cambridge, MA: National Bureau of Economic Research.

[18] Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66* (Vol. 66). CRC Press.

[19] Gelman, A., & Imbens, G. (2014). *Why high-order polynomials should not be used in regression discontinuity designs* (NBER Working Paper No. w20405). Cambridge, MA: National Bureau of Economic Research.

[20] Gerard, F., Rokkanen, M., & Rothe, C. (2016) *Identification and Inference in Regression Discontinuity Designs with a Manipulated Running Variable* (CEPR Discussion Paper No. DP11048). Retrieved from SSRN website: `http://ssrn.com/abstract=2717597`

[21] Goldrick-Rab, S., Kelchen, R., Harris, D. N., & Benson, J. (2016). Reducing Income Inequality in Educational Attainment: Experimental Evidence on the Impact of Financial Aid on College Completion 1. *American Journal of Sociology*, 121(6), 1762-1817.

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

[23] Hansen, W. L. (1983). Impact of student financial aid on access. *Proceedings of the Academy of Political Science*, 35(2), 84-96.

[24] Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, 79 (3), 933-959.

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

[26] Johnson, W. R. (1988). Income redistribution in a federal system. *The American Economic Review*, 78(3), 570-573.

[27] Kane, T. J. (1995). *Rising public college tuition and college entry: How well do public subsidies promote access to college?* (NBER Working Paper No. w5164). Cambridge, MA: National Bureau of Economic Research.

[28] Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on

treatment effects. *The Review of Economic Studies*, 76(3), 1071–1102.

[29] Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature*, 48(2), 281-355.

[30] Long, B. T. (2008). *What is known about the Impact of Financial Aid? Implications for Policy*(NCPR Working Paper). New York, NY: National Center for Postsecondary Research.

[31] Ludwig, J., & Miller, D. L. (2005). *Does Head Start improve children's life chances? Evidence from a regression discontinuity design* (NBER Working Paper No. w11702). Cambridge, MA: National Bureau of Economic Research.

[32] Marx, B. M., & Turner, L. J. (2015). *Borrowing trouble? student loans, the cost of borrowing, and implications for the effectiveness of need-based grant aid* (NBER Working Paper No. w20850). Cambridge, MA: National Bureau of Economic Research.

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

[34] McCrary, J., & Royer, H. (2003). *Does Maternal Education Affect Infant Health? A Regression Discontinuity Approach Based on School Age Entry Laws.* Unpublished manuscript.

[35] McPherson, M. S., & Schapiro, M. O. (1991). Does student aid affect college enrollment? New evidence on a persistent controversy. *The American Economic Review*, 81(1), 309-318.

[36] Page, L. C. & Scott-Clayton, J. (2016). Improving college access in the United States: Barriers and policy responses. *Economics of Education Review*, 51, 4–22.

[37] Rizzo, M., & Ehrenberg, R. G. (2004). Resident and nonresident tuition and enrollment at flagship state universities. In C. M. Hoxby (Ed.), *College choices: The economics of where to go, when to go, and how to pay for it* (pp. 303-354). Chicago, IL: University of Chicago Press.

[38] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5), 688.

[39] Schudde, L. (2013). *Heterogeneous Treatment Effects in Higher Education: Exploring Variation in the Effects of College Experiences on Student Success* (Doctoral dissertation). University of Wisconsin-Madison.

[40] Scott-Clayton, J. (2013). Information constraints and financial aid policy. In D. E. Heller & C. Callender (Eds.), *Student financing of higher education: A comparative perspective.* New York, NY: Routledge Publishing.

[41] Seftor, N. S., & Turner, S. E. (2002). Back to school: Federal student aid policy and adult college enrollment. *Journal of human resources*, 37(2), 336-352.

[42] Singell, L. D., & Stone, J. A. (2007). For whom the Pell tolls: The response of university tuition to federal grants-in-aid. *Economics of Education Review*, 26(3), 285-295.

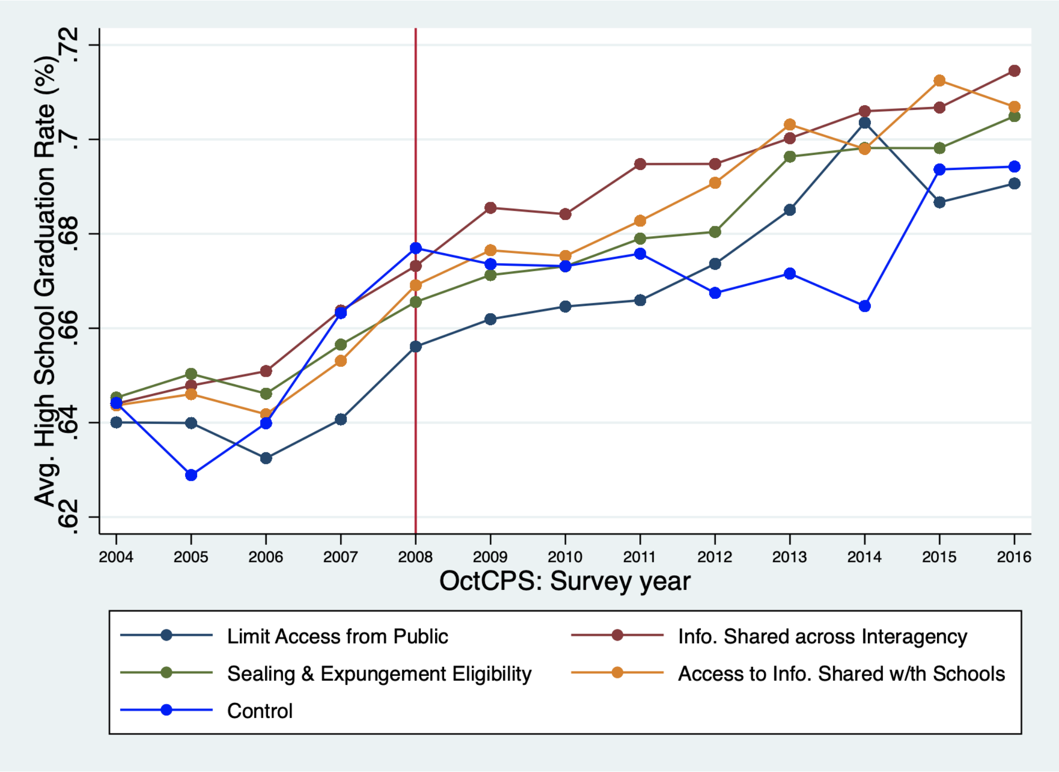
[43] Turner, L. J. (2014). *The Road to Pell is Paved with Good Intentions: The Economic Incidence of Federal Student Grant Aid* (Working Paper). College Park, MD: University of Maryland. Retrieved from `http://econweb.umd.edu/turner/Turner_FedAidIncidence.pdf`.

[44] Wiederspan, M. (2016). Denying loan access: The student-level consequences when community colleges opt out of the Stafford Loan Program. *Economics of Education Review*, 51, 79-96.

# Appendices

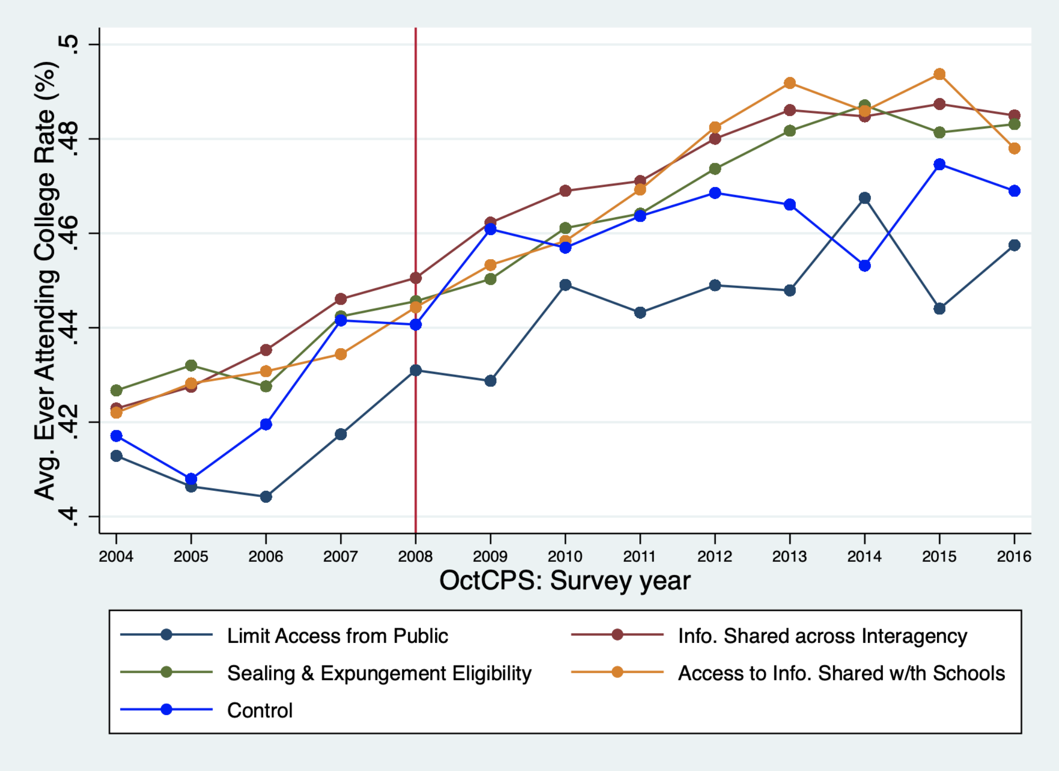
# I The Effects of Revisions to Juvenile Record Laws on Education Outcomes and Employment after the Second Chance Act of 2007

Figure A1.1: Un-Adjusted Outcome Trend: Probability of high school graduation



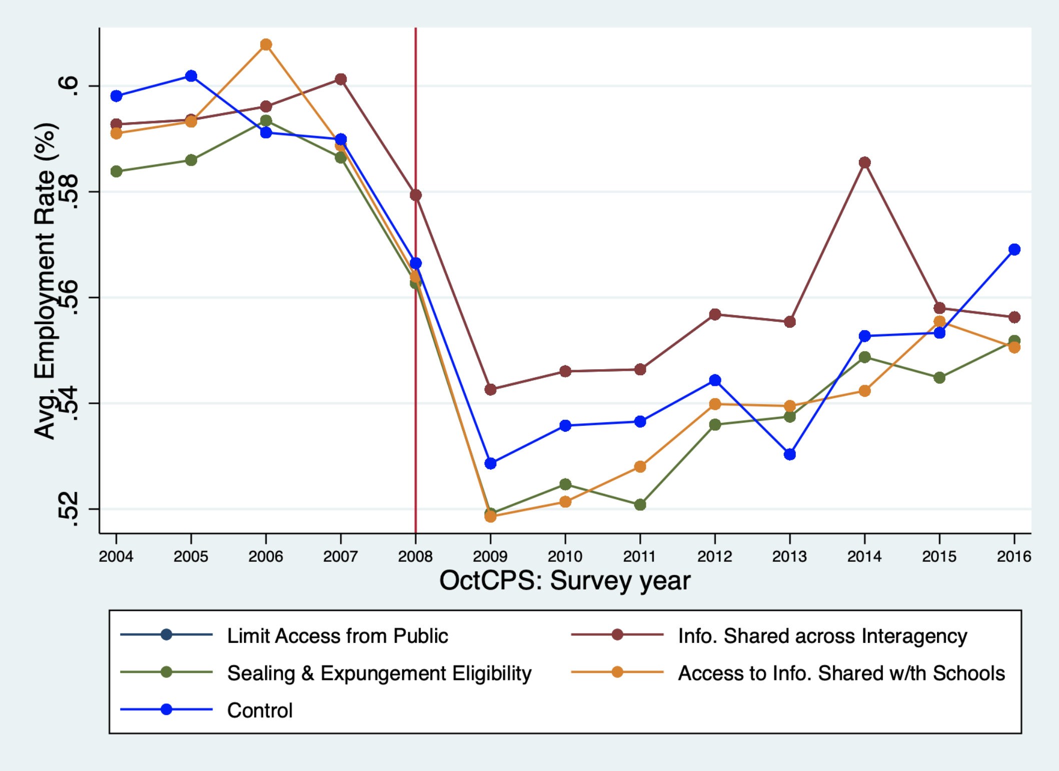
This figure shows un-adjusted trend for high school graduation. Year 2008 is excluded from the regression as a reference year. The red line indicates first year of law changes in treatment. Years prior to the red line indicate pre-treatment periods.

Figure A1.2: Un-Adjusted Outcome Trend: Probability of ever attending college



This figure shows un-adjusted trend for ever attending college. Year 2008 is excluded from the regression as a reference year. The red line indicates first year of law changes in treatment. Years prior to the red line indicate pre-treatment periods.

Figure A1.3: Un-Adjusted Outcome Trend: Employment



This figure shows un-adjusted trend for employment. Year 2008 is excluded from the regression as a reference year. The red line indicates first year of law changes in treatment. Years prior to the red line indicate pre-treatment periods.

# II Understanding Dropouts Among Community College Students: Using Cluster Analysis and Data Mining

# A R Packages

I have used R (Version 3.5.3; R Core Team, 2018) and the R-packages *bestglm* (Version 0.37; McLeod & Xu, 2018), *Biobase* (Version 2.42.0; W. Huber et al., 2015), *BiocGenerics* (Version 0.28.0; Huber et al., 2015), *boot* (Version 1.3.20; Davison & Hinkley, 1997), *caret* (Version 6.0.81; Jed Wing et al., 2018), *circlize* (Version 0.4.5; Z. Gu, Gu, Eils, Schlesner, & Brors, 2014), *cluster* (Version 2.0.7.1; Maechler, Rousseeuw, Struyf, Hubert, & Hornik, 2018), *ComplexHeatmap* (Version 1.99.5; Z. Gu, Eils, & Schlesner, 2016), *dendextend* (Version 1.9.0; Galili, 2015), *devtools* (Version 2.0.1; Wickham, Hester, & Chang, 2018), *dplyr* (Version 0.7.8; Wickham, Francois, Henry, & Muller, 2018), *factoextra* (Version 1.0.5; Kassambara & Mundt, 2017), *fastcluster* (Version 1.1.25; Müllner, 2013), *forcats* (Version 0.3.0; Wickham, 2018a), *foreach* (Version 1.4.4; Microsoft & Weston, 2017), *fpc* (Version 2.1.11.1; Hennig, 2018), *ggplot2* (Version 3.1.0; Wickham, 2016), *glmnet* (Version 2.0.16; Friedman, Hastie, & Tibshirani, 2010; Simon, Friedman, Hastie, & Tibshirani, 2011), *gmodels* (Version 2.18.1; Warnes et al., 2018), *gridExtra* (Version 2.3; Auguie, 2017), *hopach* (Version 2.42.0; van der Laan & Pollard, 2003), *huxtable* (Version 4.3.0; Hugh-Jones, 2018), *kableExtra* (Version 1.0.0; Zhu, 2019), *klaR* (Version 0.6.14; Weihs, Ligges, Luebke, & Raabe, 2005), *lattice* (Version 0.20.38; Sarkar, 2008), *leaps* (Version 3.0; Fortran code by Alan Miller, 2017), *lmtest* (Version 0.9.36; Zeileis & Hothorn, 2002), *MASS* (Version 7.3.51.1; Venables & Ripley, 2002), *Matrix* (Version 1.2.16; Bates & Maechler, 2018), *mvtnorm* (Version 1.0.8; Genz & Bretz, 2009), *NbClust* (Version 3.0; Charrad, Ghazzali, Boiteau, & Niknafs, 2014), *papaja* (Version 0.1.0.9842; Aust & Barth, 2018), *pheatmap* (Version 1.0.12; Kolde, 2019), *purrr* (Version 0.2.5; Henry & Wickham, 2018), *pvclust* (Version 2.0.0; Suzuki & Shimodaira, 2015), *randomForest* (Version 4.6.14; Liaw & Wiener, 2002), *readr* (Version 1.3.1; Wickham, Hester, & Francois, 2018), *rpart* (Version 4.1.13; Therneau & Atkinson, 2018), *stargazer* (Version 5.2.2; Hlavac, 2018), *stringr* (Version 1.4.0; Wickham, 2018b), *tibble* (Version 2.0.1; Müller & Wickham, 2019), *tidyr* (Version 0.8.2; Wickham & Henry, 2018), *tidyverse* (Version

1.2.1; Wickham, 2017), *usethis* (Version 1.4.0; Wickham & Bryan, 2018), and *zoo* (Version 1.8.4; Zeileis & Grothendieck, 2005) for all my analyses.

## B Comparing with CLARA

CLARA algorithm is an extension of K-medoids clustering algorithm allowing efficient way to deal with large number of objects. The k-medoids clustering algorithm (a.k.a PAM), uses classical partitioning technique to cluster objects. In particular, for a given number of cluster k, the algorithm iteratively (1) assigns cluster centers (i.e., medoids) to existing observation that minimizes the total distance to other points within the same cluster and (2)re-assign observations to a cluster that has the minimum distance to the cluster center(Kaufman & Rousseeuw, 1990). This iteration stops when there is no further change in assignments. K-medoids is robust to outliers and variable types (do not require numerical variable).[26] CLARA extends K-medoids algorithm by first, choosing a random sample from the dataset, then applying k-medoids algorithm to find optimal set of medoids, and then assigning cluster to all observations in the sample to its nearest medoids. The process of random sampling, finding cluster medoids, and assigning observations to clusters is repeated for a selected number of times to alleviate sampling bias (Kaufman & Rousseeuw, 1990). For each full-iteration, goodness of clustering is caculated using an average of observation dissimilarity of each cluster. After all iteration, the final choice of clustering assignment is made by the smallest average dissimilarity.

A several set of parameters should be pre-defined in order to run CLARA algorithm. First, the number of cluster (k) needs to be identified, a priori. In order to determine the number of clusters, I use two methods; elbow method and average silhouette method, as suggested in the literature (Kassambara, 2017). Second, the user needs to choose a

[26] One most widely used cluster algorithm is k-means clustering. K-means clustering algorithm differs from that it requires all variables to be numeric and for cluster centers, assigns centroids, mean of the euclidean distances between observations within cluster, rather than actual observations as in K-medoids. K-medoids is more robust to noise and outliers than K-means but, may be more time intensive for large samples (Hastie, et al., 2009). CLARA improves computing time through sampling.

distance matrix that will calculate dissimilarities between objects. Under the {cluster} R package, euclidean, manhattan, and jaccard distances are available, where I choose the default euclidean distance. The dissimilarity matrix for p variables is then computed by:

$$D(x_i, x_j) = \sum_{k=1}^{p} w_k (x_{ik} - x_{jk})^2 \tag{4}$$

$$\text{where, } w_k = 1/(\frac{1}{N^2} \sum_{i=1}^{N} \sum_{i=1}^{N} (x_{ik} - x_{jk})^2)$$

The drawbacks of partitioning algorithm still holds for CLARA. The number of cluster is unknown to a researcher, a priori. Although I use a heuristic search to identify the number of clusters and measure goodness of fit, there is no concrete way to measure how well it aligns with natural(population) number of clusters. Another drawback is, if a data point is positioned at decision boundaries, that point can be randomly assigned to any cluster in the decision boundary, which may create space for potential mis-classification errors.

### Number of Clusters

As mentioned earlier, unlike HCA, Clara requires the user to specify the number of clusters, a priori. In order to identify the number of clusters (k), I use two common methods in the literature to determine optimal number of clusters; elbow method and average silhouette method (Kassambara, 2017). The elbow method allows the user to choose a reasonable number of cluster where, the total within-cluster sum of square(wss) do not improve much. Figure A2.1 plots results from the elbow method, where y-axis is the wss and x-axis is the number of clusters (k). As the name suggests, the optimal number of cluster is at bend of the plot. As shown in top panel of Figure A2.1 (a), the optimal number of cluster is 3. The silhouette coefficient contrasts the average distance of observations within cluster to the average distance of observations in other clusters. A high silhouette width means that objects aligns well with its cluster and a low value means that there may be decent number of outliers. The optimal number of cluster is determined when the average silhouette width is

Table A2.1: CLARA Clustering Distribution

| CLARA Clustering | % |
|---|---|
| clara.cluster | proportions |
| 1 (Early Dropouts) | 0.434 |
| 2 (Late Dropouts) | 0.162 |
| 3 (Trials) | 0.404 |

*Note.* Samples are restricted to dropouts among 2002-2004 fall entry cohort students. This table shows distribution of cluster assignment from CLARA. First column indicates cluster assignments and second column indicates sample proportions of each clusters.

maximized (Kaufman & Rousseeuw, 1990). The bottom panel of Figure A2.1 (b) shows that average silhouette width is equally optimized for either two or three clusters. Combining the two results from elbow method and average silhouette method, I perform Clara with three number of clusters.

### Clara Results

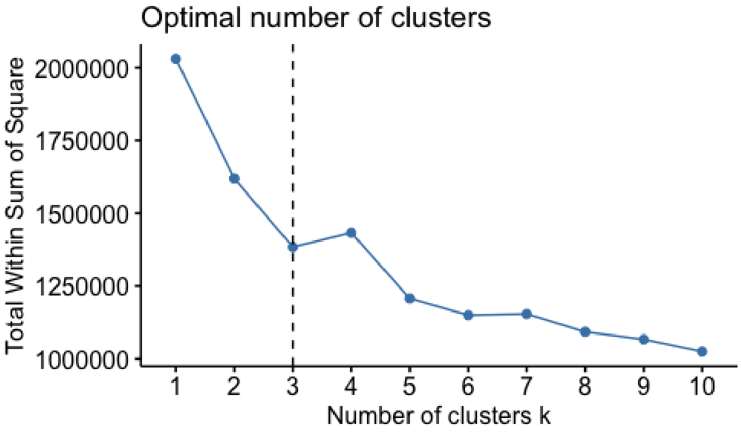
Table A2.1 presents distribution of the three clara clusters. Cluster 1 consist more than 40%, Cluster 2 consists about 16%, and Cluster 3 consists about 40% of the dropout sample. One way to examine goodness of fit is to see silhouette plot. Figure A2.2 present goodness of Clara clustering from silhouette plot. As mentioned above, silhouette width calculates mean similarity of students within cluster subtracted by mean similarity of students in the next similar cluster. Cluster 3 has .44 silhouette width, which means that Cluster 3 has a decent goodness of fit. Close to zero silhouette width for Cluster 1 and Cluster 2 suggests that observations may be somewhat similar to objects in another cluster, as I explore further below.

Visualizing clara clusters in high dimension can be difficult.[27] For the purpose to compare

[27] In data mining literature, one way to draw a cluster plot after performing principal component analysis with first two principal component vectors as x-axis and y-axis (reference). Unfortunately, only 15.7 % of variability was explained using the first two principal components, which makes it hard to visually detect

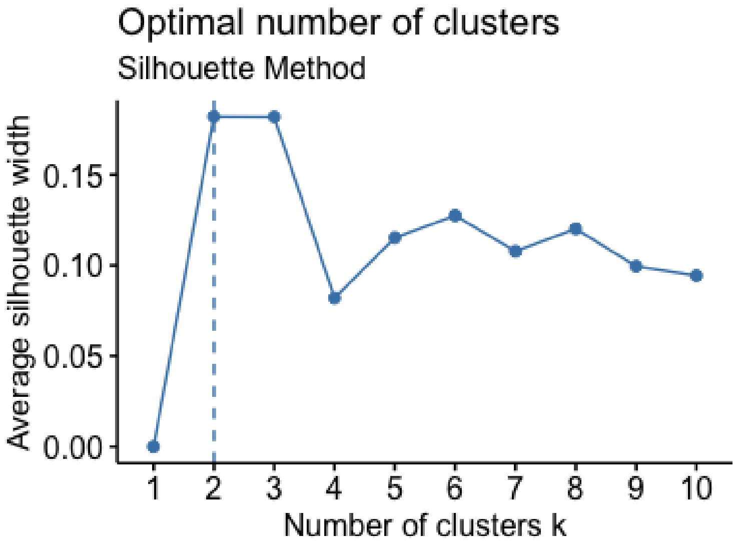
Figure A2.1: Number of Clusters using Elbow Method (top) and Silhouette Width (bottom)

(a) Elbow Method



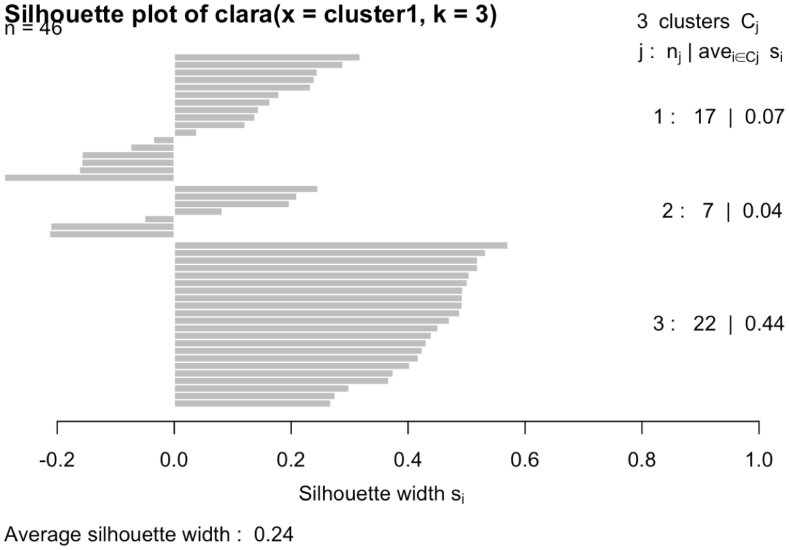
Samples are restricted to dropouts among 2002-2004 fall entry cohorts. This figure shows total within-cluster sum of square (WSS) for a range of number of clusters. The vertical dash line indicates optimal number of clusters.

(b) Average Silhouette Method



Samples are restricted to dropouts among 2002-2004 fall entry cohorts. This figure shows average silhouette width for a range of number of clusters. The vertical dash line indicates optimal number of clusters.

Figure A2.2: Clara Goodness of Fit: Silhouette Plot



Samples are restricted to dropouts among 2002-2004 fall entry cohorts. This figure shows silhouette plot of the three clusters identified by CLARA. Silhouette width range between -1 to 1 and values close to zero indicate good fit of data to cluster assignments.

with HCA, I use heatmap to represent my data and compare clara cluster assignment with HCA cluster assignment. Figure A2.3 top panel shows clara cluster assignment from the original HCA heatmap. By specifying, a priori, three numbers of clusters, CLARA grouped the last high-credit and low-credit early dropout groups into a single cluster. Figure A2.3 bottom panel re-arrange students by clara cluster assignments. Similar to HCA, the grouping seems to reflect timing of when students decide to dropout. The top group resembles late dropouts who leave school beyond third to fourth year. The middle group resembles early dropouts who leaves school beyond their first year. Lastly, the third group resembles trial group, who drops out right after their first semester. Note that the distinction of Clara clusters is somewhat less distinctive than HCA. Combining the two results from HCA and Clara clustering, two components distinguish the dropout groups: timing of when student dropouts (either after first semester, beyond first year, or beyond second to third year) and credit-side academic performances (high or low credits attempted and college-level credits earned). Interestingly, none of the two clusters weight much on GPA when distinguishing

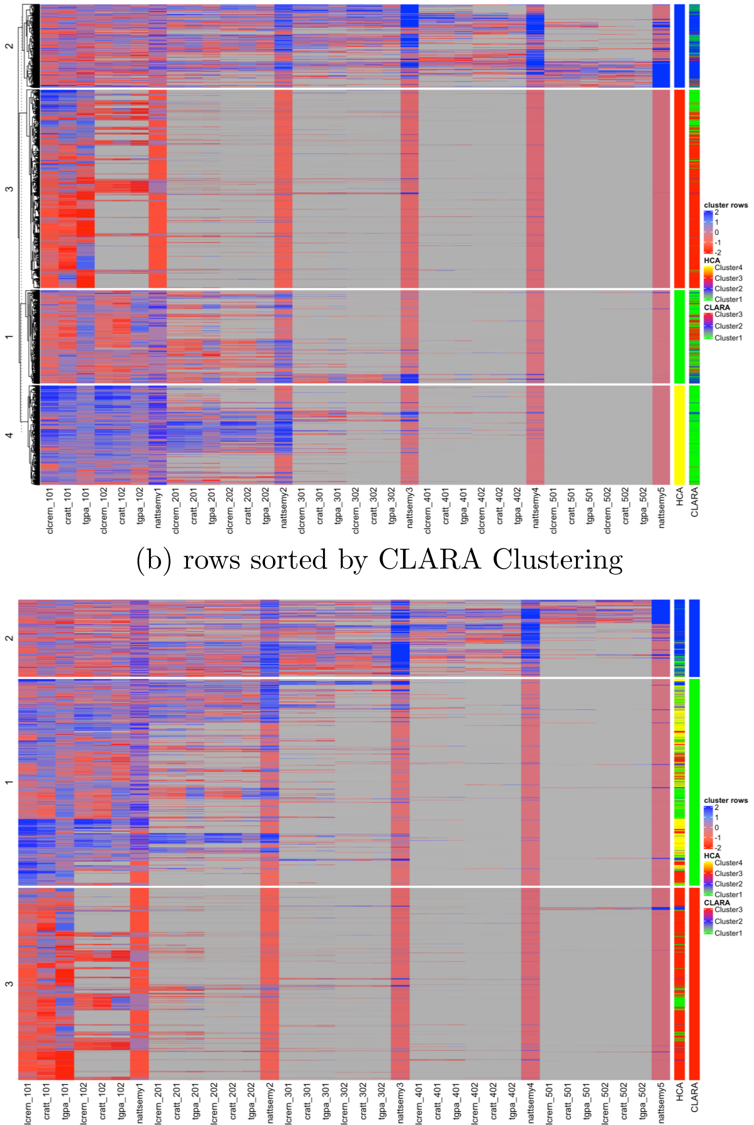
the three clusters. Request to the author to see the plots.

the groups.

To summarize, both HCA and CLARA cluster commonly shows distinctive three groups of dropouts. A trial group, who tries out college for first semester and decides early to leave college. This population tends to be older, non-white population, and are getting less financial aid. The early dropouts, as a whole, does not have distinctive characteristics except that they stay longer in school than trials group (leaves school after first year) and that they are a group with relatively high financial aid. HCA does a better job in explaining this group by dividing into two sub-groups: low-credits and high-credit early dropouts. The low-credit early dropouts resembles what we would traditionally characterize as a community college population; older, racially diverse, and are at work when first entering to school. The high-credit early dropout is an interesting group as their characteristics resemble closely to completers except that these students take more classes in their first year. Lastly, late dropout group aligns well with remedial population. These students slowly progress up the credit hours and elongate their time at school.

Figure A2.3: Comparison between HCA cluster (top) and CLARA cluster (bottom)

(a) rows sorted by HCA clustering



(b) rows sorted by CLARA Clustering

This figure compares HCA clustering assignments (4 clusters) to CLARA clustering assignments (3 clusters) using heatmap visual representation. Sample is restricted to dropouts among 2002-2004 fall entry cohorts. Each student represent rows. Each columns are listed in the order from left to right from year one to five: college-level credits earn (fall semester), total credits attempted (fall semester), term GPA (fall semester), college-level credits earn (spring semester), total credits attempted (spring semester), term GPA (spring semester),and number of attended semesters (out of total three semesters). Column values are standardized and represented in four quartiles, where red to blue represent lowest to highest quartiles. The ordering is represented by row clusters only by HCA cluster (top) and by CLARA cluster (bottom)

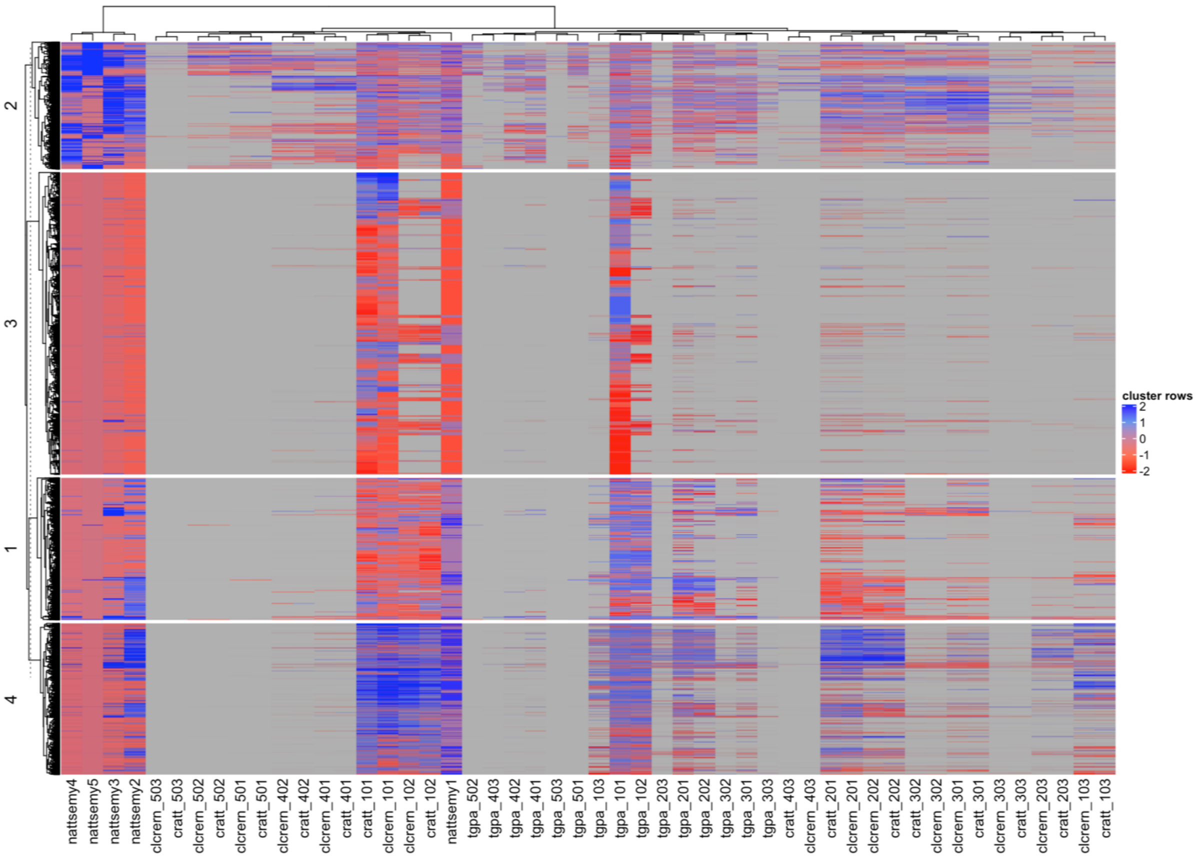
## Appendix Tables and Figures

Table A2.2: Logistic Regression with Elastic Net Regularization: Grid Search Results

|  |  | AUC |  |  |
|---|---|---|---|---|
| alpha | Low-credit Medium | Late Dropouts | Trials | High-credit Medium |
| 0.10 | 0.746703 | 0.844118 | 0.676706 | 0.813589 |
| 0.20 | 0.746772 | 0.844195 | 0.676632 | 0.814258 |
| 0.30 | 0.746801 | 0.844787 | 0.676642 | 0.814341 |
| 0.40 | 0.746844 | 0.84498 | 0.676578 | 0.814441 |
| 0.50 | 0.746819 | 0.845212 | 0.676595 | 0.81457 |
| 0.60 | 0.746811 | 0.845489 | 0.676594 | 0.814472 |
| 0.70 | **0.746852** | 0.845885 | 0.676609 | 0.814637 |
| 0.80 | 0.746783 | 0.84592 | 0.676674 | 0.814705 |
| 0.90 | 0.746819 | **0.846153** | **0.67671** | **0.814825** |

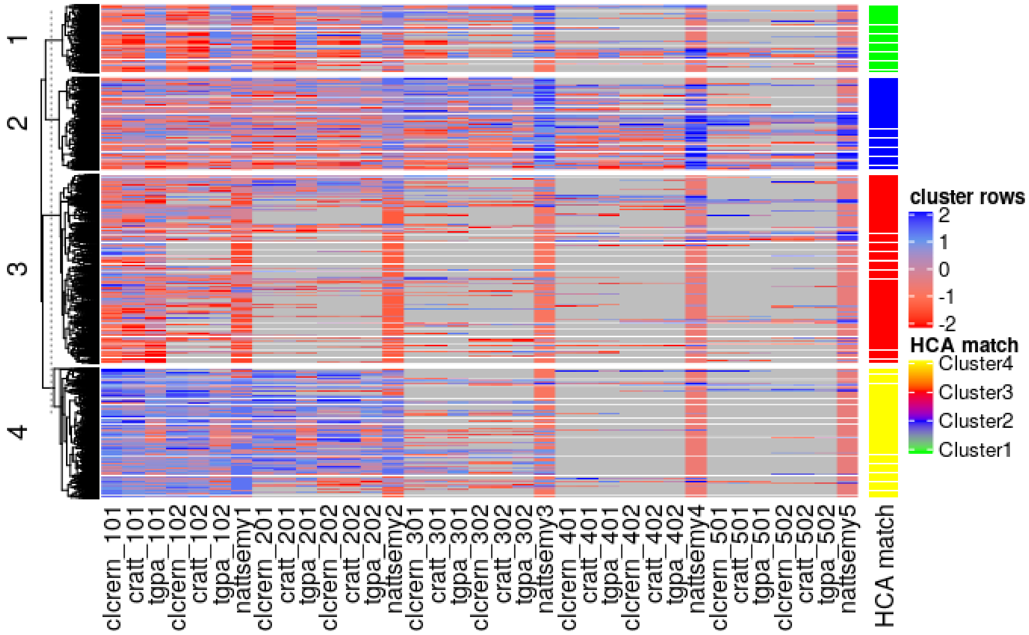
*Note.* Samples are restricted to dropouts and matched students who complete. The coefficients show AUC values for a Cartesian grid search on varying hyper parameter $\alpha$, which range between 0.1 and 0.9, by an interval of 0.1. The optimal value of $\alpha$ (highest AUC value) is indicated in bold.

Figure A2.4: HCA Heatmap (Cluster Row and Columns Only)



This figure is a visual representation of hierarchical clustering analysis result using a heatmap. Sample is restricted to dropouts among 2002-2004 fall entry cohorts. Columns exclude summer semesters. Each student represent rows. Each columns are listed in the order from left to right from year one to five: college-level credits earn (fall semester), total credits attempted (fall semester), term GPA (fall semester), college-level credits earn (spring semester), total credits attempted (spring semester), term GPA (spring semester),and number of attended semesters (out of total three semesters). Column values are standardized and represented in four quartiles, where red to blue represent lowest to highest quartiles. The ordering is represented by row clusters only.

Figure A2.5: Matched Completers using Euclidean Distance



This figure is a heatmap representation of matched sample of completing students assigned to a cluster number with similar enrollment patterns, where the match used euclidean distance. Sample is restricted to completers among 2002-2004 fall entry cohorts. Each student represent rows. Each columns are listed in the order from left to right from year one to five: college-level credits earn (fall semester), total credits attempted (fall semester), term GPA (fall semester), college-level credits earn (spring semester), total credits attempted (spring semester), term GPA (spring semester),and number of attended semesters (out of total three semesters). Column values are standardized and represented in four quartiles, where red to blue represent lowest to highest quartiles. The ordering is represented by row clusters only.

# III The Impact of Pell Grant Eligibility on Community College Students' Financial Aid Packages, Labor Supply, and Academic Outcomes

Appendix Tables and Figures are available online:

Park, R. SE., & Scott-Clayton, J. (2018). The Impact of Pell Grant Eligibility on Community College Students' Financial Aid Packages, Labor Supply, and Academic Outcomes. Educational Evaluation and Policy Analysis, 40(4), 557–585. https://doi.org/10.3102/0162373718783868