Three Essays in Economics: Recidivism, Economic Decision Making, and Biases in Beliefs

by

Catalina Franco Buitrago

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in the University of Michigan
2017

Doctoral Committee:

Professor Tanya Rosenblat, Chair
Professor Martha J. Bailey
Professor Hoyt Bleakley
Professor Dean Yang

Catalina Franco Buitrago

cfrancob@umich.edu

ORCID iD: 0000-0002-4399-5891

DEDICATION

This doctoral dissertation is dedicated to my family, friends, and mentors who have been by my side throughout this journey.

ACKNOWLEDGEMENTS

I thank Martha Bailey, Hoyt Bleakley, John Bound, Shawn Bushway, Charlie Brown, Yan Chen, Austin Davis, Laura Flak, Formarte SAS, Leonardo Garzón, David J. Harding, Amelia Hawkins, Nicolas Idrobo, Gaurav Khanna, Meera Mahadevan, Yusufcan Masatlioglu, Nicolas Morales, Jeff Morenoff, Mike Mueller-Smith, Lauren Pulay, Tanya Rosenblat, Andrea Sanchez, Jeff Smith, Neslihan Uler, Gonzalo Vazquez-Bare, Melanie Wasserman, and Dean Yang.

Chapter 1 is coauthored with Shawn Bushway, David Harding and Jeff Morenoff. Chapter 2 is coauthored with Meera Mahavedan. Chapter 1 is part of a larger project funded by NSF, NICHD, PSC, and UM Office of Research. Chapters 2 and 3 received funding from the Michigan Institute for Teaching and Research in Economics (MITRE), the Rackham Graduate Student Research Grant, and the George Katona Economic Behavior Research Award from the Institute for Social Research (ISR). I acknowledge receiving partial funding of my doctoral studies from Fulbright Colombia and Colciencias under the Francisco José de Caldas scholarship.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

My dissertation covers topics in the economics of crime and the interesction between behavioral and development economics. The first chapter provides causal evidence that sentencing low-level offenders in the State of Michigan to prison rather than probation lowers their future criminal behavior but only through incapacitation, that is, during the time they spend in prison. We identify two sources of incapacitation: primary, from the original sentence, and secondary, from higher rates of future imprisonment among those who were initially sentenced to prison. The second chapter studies how economic decision making changes along the transition from college to the labor market. By collecting panel data from students in a university in Colombia, we are able to track changes occurring after students who are in their last semester of college receive and accept a job offer, and after they receive a paycheck relative to a comparison group of students who remain in college. We find evidence that students who transition to the labor market are less present-biased, more generous, and report having lower stress about finances and higher access to resources after the job offer. After starting to work and receiving a paycheck, they perform worse on cognitive tasks and report being more worried and frustrated than students in the comparison group. This suggests that there may be greater cognitive load associated with becoming more independent and earning money. We also highlight the role of incorporating phychological measures in experimentally-elicited preference tasks. Even though it seems that last-semester students become less risk averse when receiving and accepting a job offer, this result vanishes when controlling for psychological factors. In the third chapter, we study gender differences in beliefs regarding performance and in the updating process in the developing country context. Students in the sample are enrolled in a test-preparation course to take a high-stakes college entrance exam. They are randomized into receiving or not receiving feedback about their relative ability in the five areas covered by the exam. The findings suggest that there are substantial biases in assessing own ability. Across all areas of the test, between 50 and 70 percent of the students fail to correctly predict the quartile in which their score will be. Moreover, women are more biased and more likely to underestimate their performance in math and overestimate in text analysis relative to men. I show evidence that feedback may help close the gender in gap in confidence as women report being more positive about their chances of admission to this university while the men seem less sure of this outcome.

CHAPTER I

Estimating the effects of imprisonment on recidivism: Evidence from a regression discontinuity design

1.1 Introduction

The dramatic increase in the number of people incarcerated in the United States over the last three decades (Western, 2006; West et al., 2010) has generated a discussion among policy makers, criminal justice officials, researchers, and citizens about the causes and consequences of mass incarceration and ways of reducing the size of the nation's prison population without compromising public safety (e.g., Raphael & Stoll, 2009; Travis, 2005; Alexander, 2012; National Research Council, 2008). One of the central questions in this discussion is whether sentencing a convicted felon to prison - at considerably higher cost than alternative sentences such as probation - will reduce the likelihood that the person will reoffend in the future.[1]

Despite the centrality of these questions to scholarly and policy debates, studies on the economic and social consequences of incarceration often base their inferences on nebulous counterfactual comparisons and usually fail to adequately rule out competing explanations for the putative effects of incarceration they estimate, leading some observers to conclude that "existing research is not nearly sufficient for making firm evidence-based conclusions for either science or public policy" (Nagin, Cullen, & Jonson, 2009). In recent years, a new wave of studies has used quasi-experimental designs that leverage the random assignment of judges to felony cases to estimate

[1] In the state of Michigan it is estimated that the annual cost of a bed in prison is about $34,000 while the cost of probation supervision is around $3,000.

the effects of incarceration on measures of subsequent recidivism, employment, and earnings (Abrams, 2009; Nagin & Snodgrass, 2013; Berube & Green, 2007; Green & Winik, 2010; Kling, 2006; Loeffler, 2013; Aizer & Doyle, 2015; Mueller-Smith, 2016).

The current paper contributes to this literature by implementing a quasi-experimental design to estimate the effect of being sentenced to prison compared to probation on the probability of prospectively being (a) convicted of a new felony offense, (b) severity of the new offense, (c) imprisonment due to technical violations of parole or probation,[2] and (d) imprisonment due to new felony convictions. We use data on a sample of convicted felony offenders in the state of Michigan sentenced between 2003 and 2006. The research design emerges from the structure of the sentencing guidelines and capitalizes on discontinuities in the probability of being sentenced to prison based on the formal system that is used for scoring and classifying convicted offenders in pre-sentence investigation reports, as dictated by the Michigan Sentencing Guidelines.

Empirically, the discontinuities in the probabilities of receiving a prison sentence can be analyzed under a fuzzy regression discontinuity (RD) design framework. Given the concordance between the fuzzy RD and the instrumental variables (IV) estimators, we propose an IV approach to provide the causal effect of imprisonment on recidivism. Rather than pooling all cutoffs together to construct a single cutoff as in other RD papers, we use the individual cutoffs as multiple instruments in the IV regression and obtain tighter standard errors as a result. Our setup is different than other RD applications in the sense that it contains a series of complexities such as a discrete and very rugged discrete running variable. We employ and adapt the recent methodologies in the RD literature to overcome these challenges.

Our identification strategy allows us to conclude that, among low-level offenders who are sentenced to prison, the recidivism rates for all post sentence periods analyzed and some of the post-release periods are lower than among those sentenced to probation. We present evidence that lower recidivism is fundamentally a conse-

[2] Technical violations are violations to the conditions of the original sentence by an offender under supervision (parole or probation). Examples of technical violations include missing a curfew, failure to report to office visits, or testing positive for alcohol or drugs.

quence of incapacitation. First, as previously documented in the recidivism literature, we observe incapacitation associated with the original prison sentence. Second, there is incapacitation resulting from higher future imprisonment rates among those originally sentenced to prison. We distinguish these two types of incapacitation as primary and secondary, respectively. Our results from decomposing future imprisonment into the part due to new sentences and the part due to technical violations of parole indicate that the higher rate of re-imprisonment among those sentenced to prison is primarily explained by technical violations. Furthermore, our results suggest that rehabilitation is not a channel explaining lower recidivism rates among those sentenced to prison. When analyzing the type of felonies of those who are convicted of a new felony, we find that those who were initially sentenced to prison are more likely to engage in high-severity crime.

To our knowledge, this is the first paper to causally show that re-imprisonment is a causal effect of imprisonment itself. We call this effect “secondary” incapacitation to distinguish it from the standard (“primary”) incapacitation effect reported extensively in the literature. We also identify that the mechanism for this secondary incapacitation effect is violations of parole conditions rather than engaging in criminal activity that leads to new sentences. In addition, this is one of the first papers in the recidivism literature to use a natural experiment leading to a regression discontinuity design as an identification strategy. Finally, from a methodological point of view we extend the widespread analysis of pooling multiple cutoffs together to using the individual variation of each cutoff, which increases the precision of our estimates.

The paper proceeds as follows: Section 1.2 provides a brief theoretical motivation and discussion of previous studies, section 1.3 presents the details of the Michigan Sentencing Guidelines on which our research design in based, and section 1.4 discusses the data sources and presents descriptive statistics for our analytical sample. Section 1.6 presents the results. The last two sections discuss robustness checks and conclude.

1.2 Theoretical motivation and prior research

Contemporary criminological accounts emphasize three general mechanisms through which incarceration can reduce the likelihood that a person will reoffend in the future: (1) incapacitation, (2) rehabilitation, and (3) specific deterrence. Although being incarcerated has a clear mechanical effect on suppressing crime for the duration of one's custodial sentence, the effectiveness and efficiency of incapacitation as a crime-control strategy are open to question. One limitation on its effectiveness is that by removing offenders from the community, incarceration may create criminal opportunities for new offenders through so-called "replacement" effects (Miles & Ludwig, 2007). Incapacitation is also a financially costly way to control crime, with a bed in a state prison or local jail costing an average of roughly $26,000 per year, compared to average expenditures of $2,800 per parolee and $1,300 per probationer (Schmitt et al., 2010). Moreover, it is very difficult to disentangle the incapacitative effects of prison from its behavioral effects, which could operate through rehabilitation, specific deterrence, or other mechanisms. Finally, the magnitude of any incapacitation effects depends on the criminal behavior of the comparison group, those who are not sentenced to prison but rather remain in the community.

After the decline in the support for rehabilitation as the guiding philosophy of the American penitentiary system (Bushway & Paternoster, 2009; Cullen & Jonson, 2011), there has been a resurgence of interest in and support for rehabilitation in recent decades brought about by new research on corrections programs (Cullen, 2005; Cullen & Jonson, 2011). A general conclusion of this research is that there are successful programs that curb recidivism, but their effectiveness hinges on the way they are matched to the needs of individual offenders and the extent to which they maintain program integrity (Bushway & Paternoster, 2009; Cullen & Jonson, 2011). Some scholars also argue that exposure to programs and interventions that are inappropriately matched to an offender's needs - especially those that violate the "risk principle" by exposing low-risk offenders to excessive interventions – can have criminogenic effects (Nagin et al., 2009).

Specific deterrence is another theoretical framework used to motivate studies of the effects of incarceration on reoffending. It refers to the possibility that an offender will be less likely to engage in future criminal activity after being punished for a previous crime (Bushway & Paternoster, 2009; Nagin et al., 2009). Its focus on deterring people who have already been punished for previous crimes distinguishes it from the notion of general deterrence, which refers to the broader deterrent effects that punishments may have on members of society, regardless of their prior experience with crime and punishment.

Despite the emphasis placed on the three crime-suppressive mechanisms outlined above, other theoretical perspectives suggest that incarceration may increase criminal behavior, in part through the potential effects of incarceration on employment (Western, 2006) and the subsequent effects of employment on crime (e.g., Sampson & Laub, 1995). First, prisons and jails can have "labeling" effects that can operate through stigma (and social reaction to the label) or through transformation of one's identity (the internalization of the label), and both of these can be reinforced through interactions inside and outside of prison (Nagin et al., 2009). Labeling is often evoked as one of the main reasons that former prisoners have trouble finding jobs (Pager, 2008). Also, insufficient opportunities for education and job training in prisons, along with the atrophy of job skills one brings to prison and lost job experience can all be viewed from a human capital perspective as reasons why returning prisoners may have more difficulty (re)connecting with the labor market than probationers (Kling, 2006; Loeffler, 2013; Tyler, Kling, et al., 2007). Prisons and jails are viewed by social learning theorists as "schools of crime" where pro-criminal attitudes, values, skills, and roles can be transmitted through informal interactions (Jaman, Dickover, & Bennett, 1972). To the extent that prisoners acquire pro-criminal skills and experience human capital deficits that make it harder for them to find jobs in the formal labor market, they may face more strain and differential opportunities that make crime more accessible and profitable than legal forms of work. Finally, incarceration - especially imprisonment - can deplete the social capital that one can access after prison (Loeffler, 2013). The combination of time and distance away from home can make it difficult to stay connected to relatives and friends, especially "weak ties" that can be especially useful for finding jobs (Rees, 1966; Granovetter, 1973).

It may also be the case that imprisonment increases the probability of future incarceration without increasing criminal behavior by subjecting the offender to greater surveillance and monitoring. According to the Bureau of Justice Statistics, nationwide nearly 80 percent of released prisoners are released onto parole supervision (Hughes & Wilson, 2003). These individuals can be re-incarcerated for technical violations of parole that are not crimes, such as curfew violations, failure to report, or consuming alcohol, or that are minor crimes that would not ordinarily result in imprisonment, such as drug use, petty theft, or fighting. Although individuals sentenced to probation also face surveillance and monitoring, it is generally less intensive than parole supervision, involving larger caseloads and fewer restrictions (Petersilia, 2011). Criminologists have long argued that greater surveillance will lead to greater detection of technical violations (e.g., Austin & Krisberg, 1981; Palumbo, Clifford, & Snyder-Joy, 1992), which account for almost 30 percent of all prison admissions nationwide (Carson & Golinelli, 2013). Of course, imprisonment for technical violations may prevent crime through incapacitation. Our analysis differentiates between various forms of recidivism, including differentiating between new felony convictions from imprisonment, between more and less serious new felony convictions, and between imprisonment for new convictions and technical violations of parole or probation.

In terms of prior research, a small set of studies have utilized quasi-experimental or experimental designs to study the effects of incarceration on recidivism and employment. A pair of studies using data from the Superior Court of the District of Columbia (Berube & Green, 2007; Green & Winik, 2010) use randomly assigned judges as instruments and find no statistically significant relationship between incarceration and reoffending among drug offenders. Abrams (2009) also uses an instrumental variables design in a comparison of recidivism among prisoners and probationers in Clark County, Nevada, but in his case, the instruments come from the random assignment of public defenders and his treatment is sentence length (rather than sentence type). He finds that there is a relationship between sentence length and recidivism but it is complex and non-monotonic - negative for both the shortest/weakest and longest sentences, and positive for mid-range sentences. Nagin and Snodgrass (2013) use the random assignment of judges to felony defendants in Pennsylvania who were sentenced during 1999 to estimate the effects of incarceration (compared to non-custodial sanctions) on recidivism. Aizer and Doyle (2015) use the random assignment of judges to defendants in a juvenile court in Chicago, Illinois between 1991 and 2006 to study the effects of juvenile incarceration on high school completion and incarceration in adult facilities later in life. Finally, Mueller-Smith (2016) uses the random assignment of "courtrooms" (combinations of judges and prosecutors) to misdemeanor and felony defendants sentenced in Harris County, Texas between 1980 and 2009, to estimate the effects of sentence type and length on recidivism, employment, wages, take-up of food stamps, marriage, and divorce.

In a study of sentences in Washington juvenile courts, Hjalmarsson (2009) used a regression discontinuity design that capitalizes on large discrepancies between neighboring cells of the sentencing grid in the probability of being incarcerated in a state detention facility (for 15-36 weeks) vs. a “local sanction” (which could include combinations of time served at a local detention center, community supervision, and community service) and found that incarceration reduced future offending by 35 percent. Although similar in spirit to our regression discontinuity design, this study had a very different substantive focus - the juvenile justice system where the treatments are qualitatively different.

Kuziemko (2013) also uses a regression discontinuity framework, but focuses on the effects of time served on recidivism among individuals serving around two years in prison. Given the endogeneity of time served, she exploits the sentencing guidelines in the state of Georgia as an exogenous source of variation in the number of months served. Her findings indicate that an additional month in prison reduces the 3-year recidivism rate by about 1.3 percentage points. Mueller-Smith and Schnepel (2016) take advantage of discontinuities in conviction status and type of sentence generated by the transitions between harsh to lenient regimes at two points in time in Harris County, Texas. They find that first-time drug offenders on the lenient side of the cutoff are less likely to reoffend compared to those in the harsh side of the cutoff.

As a whole, this small group of studies using quasi-experimental designs to analyze the impacts of incarceration on employment and recidivism yield several general conclusions. First, most of the quasi-experimental studies of the adult criminal justice system found no significant effects of either sentence type (e.g., incarceration vs. a non-custodial sanction) or length on recidivism outcomes. The only exception was the Mueller-Smith (2016) study, one of the few to separate the effects of incapacitation (comparing individuals currently in prison or jail to those who received non-custodial sanctions) from the longer-run effects of incarceration after the incarcerated group has been released back to the community. This study found that (a) incarceration was negatively associated with recidivism when currently incarcerated individuals were compared to those released to the community on non-custodial sanctions (i.e., an incapacitation effect), but (b) incarceration was positively associated with recidivism when both groups were compared post-release. The two studies of the juvenile justice system produced discrepant results. Using the identification strategy based on random judge assignment, Aizer and Doyle (2015) found that juvenile incarceration increased the likelihood of recidivism, defined as future incarceration as an adult. Hjalmarsson (2009), however, found that juvenile incarceration was associated with lower probability of future incarceration as a juvenile.

1.3 Michigan Sentencing Guidelines

The sentencing guidelines manual contains recommendations for the type of sentence and the sentence length that judges impose. With the exception of offenses for which there is no sentencing discretion,[3] the sentencing guidelines describe in detail the recommended sentences and sentence lengths for an offender based on the current offense, prior criminal history, and type of crime. [4]

The guidelines are indeterminate in that they (a) provide a range of minimum sentences within each cell from which judges choose, and (b) present recommended rather than mandatory minimum sentences (Deming, 2000).[5] Because the sentencing guidelines are only recommendations, judges are free to "depart" from the recommended range,[6] but departures are relatively rare, occurring in less than 2 percent of the cases analyzed in this sample.

The guidelines divide offenses into nine classes based on their severity as defined by the maximum term of imprisonment set by statute for the offense (classes A-H,

[3] Examples of felonies excluded from the guidelines are first degree murder, which carries a mandatory life sentence, or felony firearm, which carries a mandatory two-year "flat" sentence (sentence to prison for a minimum of 2 years and maximum of 2 years).

[4] The version of the Michigan sentencing guidelines for our sample applies to felonies committed on or after January 1, 1999. The current version of the guidelines can be found online: https://mjieducation.mi.gov/documents/sgm-files/94-sgm/file. The links to all prior manuals can be found here: https://mjieducation.mi.gov/felony-sentencing-online-resources.

[5] Maximum sentences are set by statute in Michigan.

[6] Judges must justify any departure in writing and are precluded from basing departures on any information already taken into account in the guidelines or on race, gender, ethnicity, nationality, religion, employment, or similar factors.

with A being the most severe, H the least severe, and class M reserved for second-degree murder). Each class has its own sentencing grid, with cells divided according to scores on two measures, the offender prior record (PR) and offense severity (OS), which are each computed as sums of scores on component measures. There are seven components to the PR score and 20 components to the OS score.[7] The total PR scores are divided into seven intervals to generate the prior record variable (PRV) level. The PRV cut-points are the same for all grids. The OS scores are also divided in intervals which determine the offense severity variable (OV) level. The number of OV levels and the cut points defining them are not the same across grids. Each cell defined by the intersection of PRV and OV levels contains a range of possible minimum sentences in months. In the example grid (see Appendix 1.9.1) the lowest minimum sentence (in months) is the large number on the left of the cell while the four numbers on the right of the cell are the highest minimum sentence lengths in months. These four subdivisions correspond to the offender's "habitual" status for offenders with prior felony records (Michigan Judicial Institute, 2016), and their function is basically to increase the upper limit of the minimum sentence of the appropriate cell by a fixed percentage.

Judges are responsible for guideline score calculations, but in practice this work is done as part of the pre-sentence investigation and sentencing information report that is provided to the judge by the Michigan Department of Corrections (MDOC) and typically prepared by an MDOC probation officer.[8] The officer relies on police reports, interviews with victims, and criminal history searches to calculate the prior record (PR) and offense severity (OS) scores and to determine the offender's habitual status. The probation officer is also the person who typically places the offender in a cell on the relevant grid based on the calculated guidelines scores. Our conversations with probation officers suggest that judges rarely request that scores be recalculated.

For our purposes, a key aspect of the sentencing guidelines is that cells on most

[7] Our understanding is that many other states have a more discrete sentencing guidelines system for classifying offenders based on prior record. Our use of the regression discontinuity design in this study depends on the fairly continuous nature of the prior record variables in Michigan.

[8] Michigan is somewhat unique compared to other states in that the Department of Corrections handles probation supervision of all offenders sentenced to felony probation. Offenders sentenced to jail or jail followed by probation for a felony also appear in MDOC records because MDOC conducts all pre-sentence investigations for all circuit courts throughout the state.

grids (classes B-G) are divided into three categories based on the types of sentences recommended: (1) “Intermediate” cells, including jail, probation and other (rarely used) sentences like fines, drug treatment, or house arrest; (2) “straddle” cells, in which any type of sentence is possible, and (3) “prison” only cells.[9] In the example grid in the appendix, intermediate cells are marked with asterisks, straddle cells are shaded, and prison cells are unmarked. As we will explain in section 1.5, our research design exploits the discontinuous jump in the probability of going to prison when crossing from an intermediate cell to a straddle cell. While four sentence types are possible in the ranges of the prior record score we study (prison, probation, jail, and jail with probation), we focus on the comparison between prison and probation. These two sentences constitute the two most extreme sentence types for offenders who are near the cutoffs and presumably have similar baseline characteristics.

1.3.1 Manipulation

There is the possibility of manipulation in assigning points to the components of the scores, but we consider this far more likely for offense severity scores than for prior record scores. Offense severity scores include potentially subjective aspects of the crime, such as whether there was psychological injury to a victim or a victim’s family member or whether a firearm was discharged in the direction of a victim, whereas prior record scores include objective characteristics of the offender’s prior criminal history, such as whether the offender was on parole or probation at the time of the offense and how many prior misdemeanors, low severity felonies, or high severity felonies the offender had been convicted of in the past (with severity defined by the exact crime of the prior conviction). For this reason, we focus on variation in sentence type generated by prior record scores, as described below.

Another potential source of manipulation is the plea bargaining process, as prosecutors and defense attorneys are well aware of the details of the sentencing guidelines

[9] Grids M and A contain only prison cells. Grid H contains intermediate and straddle cells but no prison cells. Intermediate cells have ranges in which the upper recommended limit for the minimum sentence is 18 months or less. When offenders in intermediate cells are sentenced to jail, their jail term can be 0-12 months (or zero to the statutory maximum if the statutory maximum is less than 12 months). Straddle cells have ranges in which the lower limit of the range of the minimum sentence is 5 to 12 months and the upper limit is at least 19 months. When offenders in straddle cells are sentenced to jail, their jail terms can be 0-12 months.

system. In our analytic sample, 97 percent of convictions occurred through a plea bargain (as opposed to a bench or jury trial). If a prosecutor were to base plea agreements on the exact grid cell that the individual would be placed in and on her expectations of the probability of recidivism from the likely sentence in that cell, then such manipulation would be a threat to the validity of our regression discontinuity design. However, our conversation with the probation officers who prepare pre-sentence investigations and sentencing information reports for judges lead us to doubt that such extreme and intentional manipulation is occurring. First, the cases in our analytical sample are typical cases that are processed very quickly, leaving little time and attention for such careful calculation or concern. Second, we believe that most plea bargaining occurs over the exact crime the offender will plead guilty to, and therefore which crime severity grid will govern his or her sentencing. Our analysis only makes comparisons within sentencing grids.

1.4 Data

We draw primarily on administrative data from the Michigan Department of Corrections (MDOC), which provided information on all individuals convicted of a felony between 2003 and 2006. The pre-sentence investigation records, called the "Basic Information Report" (BIR), contain the individual sentencing guidelines scores and components, identifiers for the sentencing grid and cell for each case, legal codes for the offense charged and convicted, habitual offender status, type of conviction (plea, bench trial, jury trial, etc.), offense date, conviction date, sentence date, days spent in jail ("jail credits"), sentence(s) imposed, and IDs for judges, defense attorneys, counties, and circuits. Additionally, the BIR records offender demographics, prior convictions and arrests, and substance abuse history.[10]

The main outcome of interest we analyze in this study is recidivism. Recidivism is measured in three ways: new felony convictions, severity of the new felony, and future imprisonment due to new sentences and technical violations. Data on

[10] Demographic and economic characteristics used in the analysis include age, race, gender, marital status at arrest, years of schooling, and age at first arrest. A few characteristics in the PSI are crudely measured (i.e., whether or not the offender has a history of mental illness, drug abuse, or alcohol abuse) but were nonetheless retained in the analysis as they serve as important pre-sentence variables.

new felony convictions (convictions recorded after the original sentence and for offenses occurring after the original sentence) are drawn from the BIR from MDOC. Severity of the felony is coded according to the statutory maximum sentence: a conviction with 0 to 48 months is low severity, 49 months or more includes medium or high severity and 73 or more months is high severity. Supervision records from MDOC document subsequent incarceration in prison, for a technical violation or a new sentence.[11] Conviction and imprisonment records are available through 2013. We analyze recidivism outcomes 1, 3 and 5 years after sentence and after release. In this paper we do not analyze more minor forms of recidivism that might be captured by misdemeanor convictions.[12]

An important distinction that we make concerns the start of the risk period for the outcome. One approach taken by many previous studies is starting the risk period at release. This means that for a probationer, the risk period starts at sentence but for the other three sentence types it starts once the period of incarceration in prison or jail ends.[13] An alternative approach is to define the risk period as beginning at the date of sentencing for offenders in all sentence types.

We view both approaches as having strengths and weaknesses and therefore present estimates from both approaches. Starting the risk period at release allows for comparisons with prior research and removes any incapacitation effects during

[11] Our access to data on multiple forms of recidivism and to MDOC data on the supervision of all parolees and probationers allows us to capture moves to prison for parole and probation violations that are not recorded in arrest records, a potentially important form of censoring that is not addressed in many studies.

[12] We also do not consider arrests as an outcome. We are unable to construct a comparable arrest measure for prisoners and probationers. Individuals on parole might be taken into custody by a parole officer instead of being arrested so they will not appear in the arrests data. For probationers, their “held in custody” events are not recorded in the data. Since the measurement of arrests and held in custody events are likely not the same for prisoners and probationers, we do not use these variables.

[13] For those sentenced to jail or jail followed by probation, we must estimate the date of release from jail based on the jail credits at sentencing and the sentence length because MDOC does not run the jails or track jail inmates who are not also under MDOC supervision or custody (e.g., parolees or probationers serving jail time, prisoners temporarily housed in local jails for court appearances). In an unknown number of cases these release dates are overestimates due to early release from jails, which is at the discretion of the local jail and often due to overcrowding. Given the short length of most jail sentences and because we are not concerned with estimating the effects of jail or jail followed by probation sentences, we do not see this as a problem for the present analysis.

incarceration. However, estimates under this approach may confound the effects of incarceration with period and aging effects. We overcome this problem by residualizing the after-release outcomes on calendar year and age at the moment of measurement. Moreover, starting the risk period at release has the potential to allow some endogeneity to creep back into our estimates, as release dates are potentially a function of behavior in prison. In addition, those sentenced to particularly long minimum sentences will not have post-release outcomes, especially for time periods furthest from release, creating potential for sample selection bias. Measuring outcomes starting at sentence avoids these problems, but produces estimates that may be dominated by incapacitation effects. Starting the risk period at sentence may also have more policy relevance because legislators and judges surely consider incapacitation effects in making decisions or policies related to sentencing or release from prison. In what follows, we report results from both approaches.

The analytic sample excludes re-sentences, "flat" or mandatory sentences (including life sentences), community service and fines sentences, as well as records from specialty courts (e.g., drug and family courts).[14] We retained only the "carrying offense" (the offense that determines the type of sentence, usually the most serious offense) and associated sentencing outcome when the offender was convicted of multiple offenses (around 77 percent of all cases). We perform all analyses using records for non-habitual offenders only as this category contains the vast majority of observations. The analytic sample for the RD analysis consists of around 18,000 individual records from 83 counties in Michigan whose PRV score (the running variable) is within 16 points of the relevant cutoff.[15]

Table 1.1 shows basic descriptive statistics by sentence type around a narrow window from the cutoff. Among all offenders in the sample, about 30 percent are

[14] "Flat" sentences are those for which the minimum and maximum are the same and the minimum sentence is also set by statute. In Michigan, these are primarily sentences for "felony firearms" offenses, in which a firearm is used in the process of committing another crime, either a felony or misdemeanor. Re-sentences refer to individuals previously sentenced to probation who are sentenced again due to technical violations of the terms of probation. In Michigan, probation violators must be sentenced again by a judge. The re-sentences can be for prison, jail, or longer probation. We note that re-sentences are not included in the initial selection into the analytic sample, but probationers resentenced to prison who are already in the sample are included in our measure of imprisonment for a technical violation as a recidivism outcome.

[15] Iin section 1.5 we explain how we choose the bandwidth of 16 points.

Table 1.1: Descriptive statistics of offenders sentenced to prison or probation

	Sentence type	
	Probation	**Prison**
% of observations in sample	0.30	0.10
% of women	0.19	0.09
Age at sentence	31.00	32.99
% white	0.48	0.58
% married	0.14	0.13
% with less than high school	0.45	0.43
Age at first arrest	20.44	19.47
On parole at sentence	0.01	0.15
Total number of arrests before sentence	6.40	9.89
% with mental illness	0.18	0.20
% with drug addiction	0.49	0.53
% with alcohol addiction	0.32	0.49
Months employed within a year before sentence	4.20	3.60
Months employed within 2 years before sentence	8.64	7.92
Minimum sentence length (months)	26.87	17.54
Time served (months in prison)		22.13

Notes: All figures correspond to means of the variables within 16 points from the cutoff. Less than high school does not include GED.

sentenced to probation and 10 percent to prison. The rest is sentenced to either jail or jail with probation (see Appendix Table 1.9 for descriptive statistics of all sentence types). The table shows means of the baseline covariates and average sentence length and time served in prison. The sample of offenders is primarily male, white, and non-married. Irrespective of sentence type, almost half of the individuals have very low education, 20 percent have a mental illness, and around 50 percent have an addiction to drugs and alcohol. On average, at the time of sentence the offenders were in their early thirties, and were first arrested when they were 20 years old. Finally, employment was very low, with the average offender working in the formal labor market only about a third of the time before sentence.[16] The average minimum

[16] Pre-sentence employment data come from matched records from the Michigan unemployment

sentence length is 27 months for a probationer and 18 months for a prisoner. The actual average time served in prison is 22 months on average for offenders in this sample.

Descriptive statistics for the outcomes are presented in Table 1.2 (and for all sentence types in Appendix Table 1.10). As mentioned above, we analyze both after sentence and after release risk periods for individuals within a small window around the cutoff. Panel A shows average felony recidivism for prisoners and probationers. Within 1 year after sentence, the incidence of new felonies is below 10 percent for all offenders (this reflects in part the fact that we measure new felonies at the conviction date, not the offense date, because offense dates were more frequently missing and in some cases unreliable). It increases monotonically with time for both sentence types, even though the levels are always higher for probation sentences. An incapacitation effect seems to be present for offenders sentenced to prison particularly in years 1 and 3 after sentence in which recidivism rates are substantially below those of offenders sentenced to probation. Five years after sentence around 25 percent of those originally sentenced to prison and 32 percent of those in probation have committed a new felony.

The incidence of recidivism is also low within 1 year after release for both groups. However, in contrast to the after sentence statistics, the increase in the recidivism rate is similar in both sentence types, reaching around 25 percent and 33 percent on average within 3 and 5 years after release, respectively. For prison sentences, recidivism rates start slightly smaller 1 year after release and end up slightly higher 5 years after release when compared to probation.

Panel B of Table 1.2 describes the severity of the new felony. The upper part of Panel B shows the medium- or high-severity crime rates of those originally sentenced to probation or prison. In this table and hereafter, a value of one is given when the offender commited a medium or high severity felony and zero if the new felony is low-severity or there is no new felony. Similarly, the dummy for high severity (lower part of Panel B) takes the value of one when the new felony is high severity and zero if there was no felony or the new felony is classified as low- or medium-severity. After sentence, there is a higher proportion of probationers engaging in medium and high

insurance system, which records only formal employment.

Table 1.2: Descriptive statistics of outcomes of interest

	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. Any new felony						
Probation	0.07	0.23	0.32	0.07	0.23	0.32
Prison	0.00	0.12	0.25	0.06	0.24	0.36
Panel B. Severity of new felony						
Medium and high severity						
Probation	0.05	0.14	0.19	0.05	0.14	0.19
Prison	0.00	0.08	0.16	0.04	0.15	0.22
High severity						
Probation	0.02	0.07	0.09	0.02	0.07	0.09
Prison	0.00	0.04	0.07	0.02	0.07	0.10
Panel B. Future imprisonment						
Overall						
Probation	0.03	0.11	0.16	0.03	0.11	0.16
Prison	0.01	0.16	0.28	0.11	0.28	0.34
Due to new sentence						
Probation	0.02	0.08	0.12	0.02	0.08	0.12
Prison	0.00	0.06	0.14	0.03	0.13	0.19
Due to technical violation						
Probation	0.01	0.04	0.05	0.01	0.04	0.05
Prison	0.01	0.10	0.17	0.08	0.17	0.21

Notes: Robust standard errors. The outcomes are defined as the variables in italics in the time frame specified in the headings of columns 2 to 7 (e.g. any new felony within 1 year after sentence). The figures represent the means for probationers and prisoners for each outcome.

severity crime and the average rates are fairly similar between the two sentence types after release. In the case of high-severity felonies only, we see that the averages for prisoners and probationers are virtually the same after release and slightly smaller for prisoners in the after-sentence period.

Panel C of Table 1.2 describes the rates at which offenders sentenced to each sentence type are imprisoned 1, 3 and 5 years after the original sentence and after release. In all periods after release and 3 and 5 years after sentence overall imprisonment rates of offenders originally sentenced to prison are higher than the rates of those sentenced to probation. One year after sentence, only 31.5 percent of prisoners have been released to the community which makes the future imprisonment figure smaller for prisoners than for probationers.[17] Decomposing future incarceration into the parts due to a new sentence and due to technical violations of parole or probation we see that the average of prisoners and probationers charged with a new sentence is similar in the after sentence outcomes but substantially higher for prisoners when we look at technical violations. After release, the averages in Panel B are higher for those originally receiving a prison sentence but the difference between prisoners and probationers is substantially larger in the case of technical violations relative to new sentences.

1.5 Empirics and first stages

A simple OLS analysis in this setting is likely to confound potentially omitted variables with sentence type assignment. For example, factors unobserved by the researchers but observed by the judges may lead them to assign prison sentences to individuals who are more likely to recidivate. Hence, a regression of the recidivism outcomes on a treatment indicator for prison does not represent the causal effect of receiving a prison sentence on recidivism but rather a combination of causal effects and omitted factors. For this reason, we exploit the quasi-random variation provided by the Michigan Sentencing Guidelines.

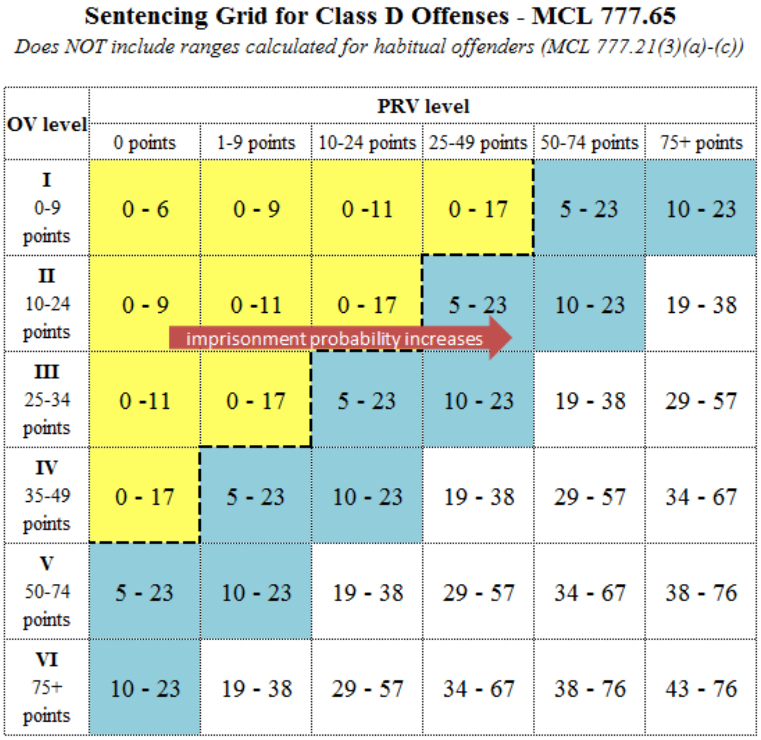
Our analysis exploits the exogenous change in the probability of being sentenced to prison arising from the marginal increase in prior record (PRV) scores that moves an offender from an intermediate cell (where the presumptive sentence is something other than prison) to a straddle cell (where recommended sentence types include prison). In other words, offenders with similar PRV scores face different probabilities of going to prison depending on whether their PRV score lies to the left or right of a cutoff that determines the boundary between an intermediate and a straddle cell. This setting naturally leads to a fuzzy regression discontinuity design (RDD) because the increase in the probability of going to prison from crossing the boundary between an intermediate and a straddle cell is less than one (see Figure 1.3 in subsection 1.5.3).

Figure 1.1 shows a simplified version of the exogenous variation from the sentenc-

[17] The proportion of prisoners released 1, 3, and 5 years after sentence is 31.5, 83.4, and 94.4 percent respectively.

ing guidelines used in our identification strategy. The grid shows the cells created by the interesection of PRV and OV scores and within each cell the recommended minimum sentence length for the judge to choose from. Intermediate, straddle and prison cells are in yellow, blue and white, respectively. As the arrow in the figure shows, crossing the cutoffs between intermediate and straddle cells increases the probability of receiving a prison sentence. The full version of the grid is in Appendix 1.9.1.

Figure 1.1: Simplified version of SGL grid - basis for identification strategy



Sentencing Grid for Class D Offenses - MCL 777.65

Does NOT include ranges calculated for habitual offenders (MCL 777.21(3)(a)-(c))

OV level	PRV level					
	0 points	1-9 points	10-24 points	25-49 points	50-74 points	75+ points
I 0-9 points	0 - 6	0 - 9	0 -11	0 - 17	5 - 23	10 - 23
II 10-24 points	0 - 9	0 -11	0 - 17	5 - 23	10 - 23	19 - 38
III 25-34 points	0 -11	0 - 17	5 - 23	10 - 23	19 - 38	29 - 57
IV 35-49 points	0 - 17	5 - 23	10 - 23	19 - 38	29 - 57	34 - 67
V 50-74 points	5 - 23	10 - 23	19 - 38	29 - 57	34 - 67	38 - 76
VI 75+ points	10 - 23	19 - 38	29 - 57	34 - 67	38 - 76	43 - 76

For a single cutoff, the fuzzy RD can be described by the following two-equation system:

$$y_i = \beta_0 + \tau D_i + \beta_1(PRV_i - c_i) + \beta_2(PRV_i - c_i) \cdot D_i + X\gamma + \varepsilon_i \tag{1.1}$$

$$D_i = \alpha_0 + \eta T_i + \alpha_1(PRV_i - c_i) + \alpha_2(PRV_i - c_i) \cdot T_i + X\theta + \nu_i \tag{1.2}$$

Where (1.2) is the first stage equation relating the treatment dummy (D_i) with an indicator for crossing the cutoff (T_i), and (1.1) is the structural equation relating the outcome y_i with the treatment dummy. The parameter of interest is τ, the effect of being sentenced to prison on the outcome. The running variable (PRV score) is centered at zero by subtracting the value of the cutoff relevant to each individual (c_i) so that equation (1.1) estimates the treatment effect at the cutoff. The matrix X includes the treatment dummies that are not being instrumented as well as controls for age and predetermined covariates.[18] The outcomes we consider are various measures of recidivism over different time frames as described in the data section.

We also conduct a reduced-form analysis that estimates the effects of crossing the cutoffs on the outcomes. From the equations above, the reduced-form equation is obtained by substituting equation (1.2) on equation (1.1) as follows:

$$y_i = \gamma_0 + \tau_R T_i + \gamma_1(PRV_i - c_i) + \gamma_2(PRV_i - c_i) \cdot T_i + X\lambda + \epsilon_i \tag{1.3}$$

In this case, the coefficient τ_R is the intent-to-treat effect, that is, the effect of being eligible for a prison sentence (by crossing the boundary between an intermediate and a straddle cell).

The fuzzy RD estimation is mathematically equivalent to instrumental variables (IV) estimation. In this sense, the indicator for crossing the cutoff (T_i) can be used as an instrument for the treatment dummy (D_i), and the two equations above can be estimated by 2SLS. This method provides the causal effect of the treatment on the outcomes of interest for those who are affected by the instrument (crossing the cutoff) provided that the instrument only affects the outcome through its effect on the probability of going to prison (the exclusion restriction), and that crossing the cutoff only makes offenders more likely to go to prison (monotonicity).[19]

[18] Age is mechanically correlated with the running variable, a composite of the offender's prior record. Older offenders will on average have higher prior record scores since they have had more time to commit and be apprehended for crimes. As a result, we understand the RD design to be valid in this setting only once we have conditioned on age.

[19] One possible threat to the exclusion restriction is that crossing the cutoff could affect the outcome through changes in sentence lengths as well as sentence types. However, even though we do find some variation in prison length for those sentenced to prison to the right of the cutoff, we do not find this effect to be large enough to be important for our estimation strategy (see robustness checks section).

1.5.1 Bandwidth choice

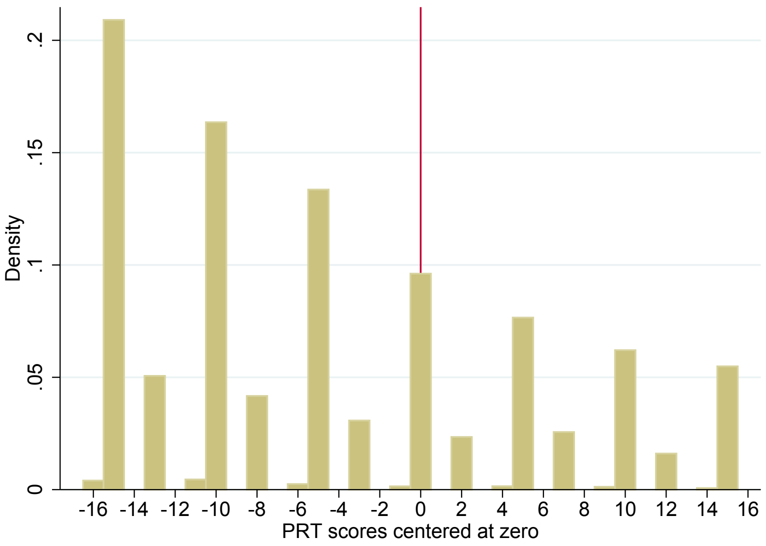
One estimation method for RD designs involves choosing a high polynomial order to estimate flexible parametric regressions of the running variable on either side of the cutoff using all available observations. We avoid this approach because it produces biased point-estimates and standard errors if done incorrectly.[20] In general, in RD it is generally recommended to choose a narrow bandwidth close to the cutoff and perform the estimation using only observations within that bandwidth. The basic trade-off in bandwidth selection in RD is between bias and variance. A very small window around the cutoff will have low bias but high variance due to the small number of observations. Alternatively, a bigger bandwidth will give estimates with higher bias but low variance. There are procedures like those of Calonico, Cattaneo, and Titiunik (2014) or Imbens and Kalyanaraman (2011) and cross-validation (Ludwig and Miller, 2007) that compute the optimal bandwidth. These procedures are based on continuity of the running variable which we do not have in this application.

All estimates in the next section are obtained using observations within a bandwidth of 16 points from the cutoff. To arrive at this window around the cutoff we do not perform McCrary-type tests (McCrary, 2008) due to the nature of the PRV scores (running variable). As discussed in section 1.3, the PRV scores are constructed from 7 different prior record variables. The majority of these variables are coded in multiples of 5. While values of 1 and 2 are also possible, they are far less common than the multiples of 5. Hence, it is impossible or very unlikely to observe certain values of the score (see Figure 1.2). In this sense, the running variable is not only not continuous but very rugged, which makes the McCrary test and the version of this test for discrete variables (Frandsen, 2014) non-informative, as the tests will appear to detect evidence of manipulation where there are merely mathematically impossible values of the running variable. Hence, we rely on balance tests of the

[20] As we explored the possibility of a strategy involving polynomial functional forms for the running variable and all observations, we conducted AIC tests to obtain the optimal polynomial order for each outcome / cutoff combination when using the whole range of observations in Grids D, E and F. The optimal orders varied considerably across cutoffs, so to avoid imposing the same functional form for all cutoffs, we opted for an alternative approach in which the linear approximation is valid for all outcomes / cutoffs once we restrict the observations to a narrow window around each cutoff. The global polynomial approach is also not recommended on the basis of the high weight that observations far from the cutoff receive, the sensitivity of the estimates to different polynomial fits, and the narrow confidence intervals this method estimates (Gelman & Imbens, 2014).

covariates to establish the validity of our RD design.

Figure 1.2: Running variable within a 16-point bandwidth



The approach we take in this paper is based on balance tests of predetermined covariates, similar to Cattaneo, Frandsen, and Titiunik (2015) who propose randomization inference for RD designs. We choose a bandwidth in which the joint test of the hypothesis that the covariates in Table 1.1 are balanced is not rejected. Simultaneously, we require that the candidate bandwidth is not too far from the cutoff so a linear fit of the running variable is appropriate. We implement this test by running seemingly unrelated regressions (SUR) where the outcome of each regression is a covariate in a specification like the one given in equation (2). From this approach, we conclude that an appropriate window for the analysis is within 16 points from the cutoff because we do not reject the null hypothesis that the covariates are jointly balanced at the cutoff.[21] See Appendix 1.9.2 for the results from the SUR tests for different bandwidths. Graphs of the covariates and of covariates residualized by age are in Appendix 1.9.3.

[21] As mentioned earlier, in the SUR regressions we control for a quadratic on age at sentence because age is mechanically correlated with the running variable.

1.5.2 Empirical considerations

One complexity of our setting is that our treatment, sentence type, is not dichotomous but rather has four categories. The sentence types contemplated in the Michigan sentencing guidelines can be many as discussed in section 1.3. In particular, the vast majority of offenders in our sample are sentenced to prison, jail, jail with probation, or probation only. We do not consider other possible sentences such as fines or community service because we believe individuals who receive such sentences have zero probability of receiving a prison sentence and are therefore not appropriate comparisons for those sentenced to prison. Our main theoretical question relates to the prison vs. probation comparison but since other sentence types are possible in intermediate and straddle cells, we must incorporate them in the analysis.

We construct indicator variables for sentence type in which the reference category is probation. The sentence type variables are defined as prison vs. everything else, jail vs. everything else, and jail with probation vs. everything else. However, our treatment of interest is sentence to prison as compared to sentence to probation. To make this comparison explicit in the regression we include the indicators for jail and jail with probation as controls in the regression without giving a causal interpretation to their coefficients as this will require additional exogenous variation.[22]

Another complexity in our application of fuzzy RD is that the Michigan sentencing guidelines contain many potential discontinuities. Depending on the grid and OV level where an offender is placed, the offender will be affected by a different cutoff. For example, in the grid shown in the appendix, an offender in OV level I would be affected by the PRV score cutoff of 50 points while the relevant cutoff for an offender in OV level II is 25 points. Therefore, only one cutoff is relevant for each individual. This differs from the setup in other RD designs with multiple discontinuities in which the same individual is affected by all the cutoffs (see for example Van der Klaauw, 2002).[23]

[22] See the section on robustness checks for further discussion.

[23] A common strategy to analyze the data from multiple cutoffs in a regression discontinuity setup is to normalize all cutoffs to zero and run a pooled regression on all observations to obtain the treatment effect based on a single-cutoff estimator. Cattaneo et al. (2016) show that normalization of the cutoffs estimates an average of local average treatment effects that is weighted by the relative density of observations around each cutoff.

The multiple discontinuities in our setting provides a richer framework to analyze the data relative to a setting in which all cutoffs are pooled together. First, we can account for heterogeneity in the first stage across multiple cutoffs in our data by reformulating the IV setup so that we can use the 11 cutoffs as 11 separate instrumental variables. This relaxes an assumption implicit in an analysis that pools all cutoffs together, that the effect of crossing each cutoff on treatment assignment is the same for all cutoffs. Such an assumption may be unreasonable here because the various cutoffs are for different crime classes (severities), for offenders with different criminal histories, and for offenses with different characteristics. Second, we expect that sampling variability will be lower when using all discontinuities as separate IVs than when pooling because the first stages more accurately capture the change in treatment probability that comes from crossing each cutoff, leading to a stronger first stage. In 2SLS, a stronger first stage leads to smaller standard errors for the treatment effect coefficient(s) (Wooldridge, 2010).

Empirically, we construct 11 instrumental variables in the form of 11 indicator variables equal to one when the PRV score is greater than or equal to each individual cutoff value. We also include an interaction between the treatment dummy and the running variable to allow for different slopes in either side of the cutoffs, which is instrumented by interactions between each of the IVs and the PRV scores. We make sure that we compare only observations affected by the relevant cutoff by including cutoff indicators and their interactions with the running variable. We center the running variable at zero for each cutoff, and include controls for the non-relevant sentence types, age and baseline covariates, and estimate the model by 2SLS.

In terms of interpretation of the parameters of interest, we estimate an efficient linear combination of the instrument-specific LATEs with weights given by the relative strength of each instrument in the first stage. In the words of Angrist and Pischke (2008), “2SLS is a weighted average of causal effects for instrument-specific compliant subpopulations.”

One futher consideration in our setting is the discreteness of the running variable. In the case of a discrete running variable, the confidence intervals based on

Eicker-Huber-White standard errors do not have good coverage because with few values of the running variable the bandwidth has to be too large. This means that the asymptotic bias of the polynomial approximation may not be negligible as undersmoothing requires (Kolesár & Rothe, 2016). Lee and Card (2008) recommend using standard errors clustered at the value of the running variable level when it is discrete. However, Kolesár and Rothe (2016) show that clustering with a small number of support points biases the standard errors downward and is sensitive to misspecification. This issue is analogous to the small number of clusters problem in other settings. In response to this conundrum and because there are no validated methods to solve either problem, we resort to Eicker-Huber-White standard errors, which seem to perform better than clustered errors in simulations by Kolesar and Rothe (2016).[24]

1.5.3 First stages

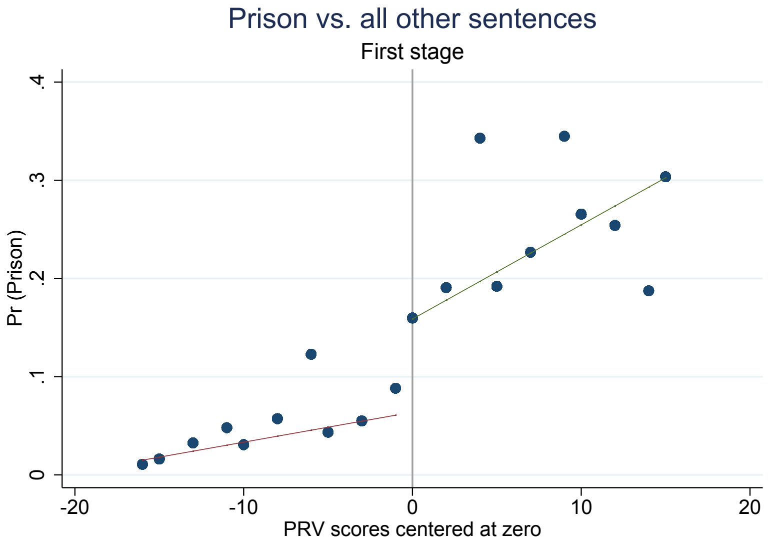
Figure 1.3 shows the basic relationship between the probability of going to prison and the dummy for crossing the cutoff when all cutoffs are pooled together and centered at zero. The y-axis shows the probability of going to prison relative to the other three types of sentences, and the x axis shows the PRV score. Scores greater or equal to zero indicate that the individual's PRV score is at or to the right of the cutoff relevant for that individual. Each dot represents the average probability of going to prison for each value of the PRV scores, and the lines are the fitted values from a regression of the prison indicator on a dummy for crossing the cutoff, the running variable and an interaction between the two.[25]

There is a clear discontinuity in the probability of being sentenced to prison for individuals with PRV scores at and to the right of the cutoff, i.e., offenders whose PRV score places them in a straddle cell. Visually, the jump in the probability of prison at the cutoff is around 10 percentage points.

[24] These authors propose two new estimators to correct the small number of clusters problem. As of now it is not possible to implement either of the two estimators because the paper is not yet published and the authors do not provide any code to make implementation feasible.

[25] Note that although all dots are drawn the same size, they may represent very different numbers of observations. For example, the first dot to the left of the cutoff contains 34 observations while the dot at the cutoff contains 2,234. This is the case with all graphs we present in this paper.

Figure 1.3: First stage: Probability of going to prison (pooled cutoffs)



The first stage graph in Figure 1.3 illustrates the raw discontinuity, that is, without controlling for age, baseline covariates and cutoff fixed effects. Table 1.3 presents the first stage regressions for the pooled sample and each cutoff individually. In the pooled case, the regression controls for a quadratic on age, baseline covariates, cutoff fixed effects and their interactions with the running variable. Crossing the cutoff increases the probability of receiving a prison sentence by 7.8 percentage points.

We estimate the first stage for each of the cutoffs in Grids D, E and F. The magnitude of the first stage oscillates between 1.2 percentage points in Grid E, OV III to 36.6 percentage points in Grid F, OV IV. There is variation in the size of the jump across cutoffs, as well as in the sample sizes. The small sample sizes in each cutoff suggest that estimating separate models for each cutoff would lead to extremely underpowered estimates.[26]

[26] The cutoffs that seem to provide stronger first stages in terms of the size of the jump and the statistical significance are those in Grid D. This makes sense based on the classification of the crime classes provided by the Michigan Sentencing Guidelines; high severity crimes are in grids letters closer to A. Offenders in Grid D have committed more serious crimes, and are therefore more likely to be sentenced to prison if their scores place them in a straddle cell. One of the cutoffs (Grid F, OV I) is not statistically significant.

Table 1.3: First stages: pooled and by cutoff

	$1\{PRV_i \geq \text{cutoff}\}$	S.E.	Observations	R-squared
All cutoffs pooled	0.078***	0.008	18,479	0.147
Grid D				
OV I	0.123*	0.053	895	0.172
OV II	0.087**	0.034	1,530	0.148
OV III	0.257***	0.045	553	0.129
Grid E				
OV I	0.037**	0.012	5,065	0.122
OV II	0.088***	0.015	5,257	0.145
OV III	0.120**	0.043	887	0.198
OV IV	0.093*	0.047	698	0.129
Grid F				
OV I	0.022	0.043	716	0.078
OV II	0.038*	0.018	2,130	0.078
OV III	0.135**	0.050	558	0.111
OV IV	0.362***	0.097	190	0.174

Notes: Robust standard errors. All models regress the prison dummy on a dummy for crossing the cutoff, the PRV scores, the interaction between the two, cutoff fixed effects, their interaction with the PRV scores and a quadratic on age at sentence.
*** p<0.01, ** p<0.05, * p<0.1

1.6 Results

This section provides evidence that offenders sentenced to prison are less likely to recidivate than probationers. This result is mainly driven by incapacitation of two types: primary incapacitation (from the original sentence) and secondary incapacitation (due to higher rates of future imprisonment). We present results for the new-felony recidivism, severity of new felony and future incarceration outcomes. All reduced-form and IV models are estimated parametrically with a linear functional form of the running variable allowing for a different slope on either side of the cutoff. Consistent with current best practice in RD designs (see for example Lee & Lemieux, 2010), we use local linear regressions (LLR) around each cutoff using only observations that are within a 16-point bandwidth from the cutoff. Standard errors are Eicker-Huber-White.

1.6.1 Reduced-form estimates

We first analyze the intent-to-treat effect, that is, the change in the outcomes when a prison sentence is more likely as a result of crossing the cutoff. The basic specification in these regressions is in equation (1.3).

The analysis of reduced forms is directly relevant for policy because moving cutoffs slightly to the right or to the left of the current cutoffs will possibly translate into changes in offenders' recidivism outcomes. The thought experiment consists in marginally shifting the cutoffs in either direction and inferring how recidivism outcomes would change. As an example, policy makers could marginally move the cutoffs to the left, that is, offenders with slighly lower PRV scores relative to the current cutoffs have a higher chance or facing a prison sentence. We are interested in inferring what would happen to the recidivism of those offenders in such scenario.

Overall, we find that offenders to the right of the cutoff are incapacitated 1 and 3 years after sentence but there is no statistical difference in new-felony recidivism rates 5 years after sentence and in all post-release periods. Furthermore, offenders to the right of the cutoffs have higher rates of future imprisonment 1 and 3 years after release than offenders to the left. Those who are to the right are much more likely to be imprisoned in the future as a result of a technical violation of their parole or probation conditions.

Table 1.4 presents the results from the pooled reduced-form regressions for all outcomes. The coefficients in the table are the point estimates of the indicator for crossing the cutoff. The outcomes are defined by two components: the variable of interest (column 1 in the table) and the time frame in which it is measured (within 1 year, 3 years, and 5 years after sentence and after release in subsequent columns). The after-sentence estimates include any incapacitation effect due to offenders still being in prison at the time the outcome is measured. Recall from section 3.3 that, in this sample, about 30 and 85 percent of prisoners have been released to the community 1 and 3 years after sentence, respectively. We also report, for each regression, the mean of the outcome for individuals below the cutoff, i.e. offenders with a prior record score between -16 and 0, not inclusive.

Table 1.4: Reduced-form regressions (pooled cutoffs)

	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. Any new felony						
Any new felony	-0.018**	-0.029*	-0.012	-0.012	-0.015	-0.007
	(0.007)	(0.012)	(0.014)	(0.008)	(0.013)	(0.014)
Mean below cutoff	0.048	0.165	0.234	0.073	0.215	0.296
Observations	18479	18479	18479	18383	18348	18233
Panel B. Severity of new felony						
Medium and high severity of new felony	-0.007	-0.008	-0.001	0.001	0.003	0.002
	(0.006)	(0.010)	(0.012)	(0.007)	(0.011)	(0.012)
Mean below cutoff	0.029	0.095	0.133	0.046	0.130	0.177
High severity of new felony	-0.002	0.012	0.005	0.003	0.014	0.003
	(0.004)	(0.007)	(0.008)	(0.004)	(0.007)	(0.009)
Mean below cutoff	0.013	0.047	0.065	0.020	0.057	0.076
Observations	18479	18479	18479	18383	18348	18233
Panel B. Future Imprisonment						
Overall	0.011	0.019	0.020	0.021**	0.022*	0.017
	(0.006)	(0.010)	(0.012)	(0.007)	(0.011)	(0.012)
Mean below cutoff	0.017	0.071	0.102	0.041	0.118	0.157
Due to new sentence	-0.003	-0.010	-0.012	-0.002	-0.012	-0.013
	(0.004)	(0.008)	(0.010)	(0.005)	(0.008)	(0.010)
Mean below cutoff	0.009	0.045	0.071	0.019	0.070	0.105
Due to techical violation	0.013**	0.031***	0.035***	0.023***	0.037***	0.037***
	(0.004)	(0.008)	(0.009)	(0.006)	(0.008)	(0.009)
Mean below cutoff	0.008	0.029	0.038	0.022	0.052	0.066
Observations	18440	18440	18440	18383	18365	18306

Notes: Robust standard errors. The outcomes are defined as the variables in italics in the time frame specified in the headings of columns 2 to 7 (e.g. any new felony within 1 year after sentence). All models regress the outcome on a dummy for crossing the cutoff, the PRV scores, the interaction between the two, cutoff fixed effects, their interaction with the PRV scores and a quadratic on age at sentence. The coefficients in the table are the point estimates of the dummy for crossing the cutoff. Means of after release variables are from the non-residualized variables.*** p<0.01, ** p<0.05, * p<0.1

We first present evidence of incapacitation effects for offenders whose prior record score is located at or to the right of the cutoff. Panel A of Table 1.4 shows that, after sentence, offenders with PRV scores at or above the cutoff are less likely to be convicted of a new felony than offenders with scores below the cutoff. The difference in the average probability of committing a new felony varies between -1.2 and -2.9 percentage points between somebody at or to the right of the cutoff and somebody to the left 1 and 3 years after sentence, respectively. We do not find statistically significant differences between those above and below the cutoff five years after sentence or for any of the felony outcomes after release.

Furthermore, the mean new-felony recidivism rates of those below the cutoff increases over time for both risk periods. Within one year after sentence, only 5 percent of individuals to the left of the cutoff commit a new felony. Within five years of sentence, this mean is up to 23.4 percent. The means are slightly higher for the after release outcomes but the patterns remain the same.

Individuals at or to the right of the cutoff are more likely to commit a high severity felony in all post-release periods except within one year of release (Panel B of Table 1.4). Even though the effects are not statistically significant, we will see in the next section that the IV analysis finds a significant increase in high severity felonies for those sentenced to prison relative to probationers unconditional on commiting a new felony.

Another incapacitation effect relates to future imprisonment. In this case, individuals are re-imprisoned if their original sentence was prison or imprisoned for the first time if they initially received a sentence other than prison. We find that offenders at or to the right of the cutoff are more likely to be imprisoned in the future than offenders to the left of the cutoff (Panel C of Table 1.4). Overall future imprisonment differs by one or two percentage points between individuals above and below the cutoff. What differs substantially is the mean of those below the cutoff. Average imprisonment rates are very low (below 2 percent) one year after sentence and increase up to 10 percent within 5 years of the original sentence for those below the cutoff. After release, we find that offenders above the cutoff are 2 percentage points more likely to be imprisoned within 1 and 3 years after release. Notice, in addition, that the mean of those below the cutoff increases from 4 to 12 percent between years 1 and 3 after release, respectively. Within 5 years of release the future imprisonment rates of both groups are statistically equal at about 16 percent.

We separate future imprisonment into imprisonment for a new sentence and imprisonment for a technical violation. The reduced-form analysis points out that the higher rate of future imprisonment does not seem to be a result of new sentences but rather of technical violations. Recall that technical violations are violations of the conditions of sentence during parole or probation such as missing a curfew. Across all risk periods after sentence and after release, offenders with a prior record score at or to the right of the cutoff are significantly more likely to be imprisoned as a result of a technical violation. Even though future incarceration due to technical violations rates are low (below 7 percent), individuals at or to the right of the cutoff are incarcerated at rates that are between 1.3 and 3.7 percentage points higher.

In sum, reduced-form evidence suggests that in a counterfactual scenario where the cutoffs were moved marginally to the left, we would expect higher incapacitation among individuals who would now be at or to the right of the cutoffs. Incapacitation would arise because they would be more likely to receive a prison sentence initially and because they would be more likely to be sentenced to prison in the future than someone to the left of the cutoff.

1.6.2 OLS and 2SLS estimates

In this section we exploit the fact that our RD is a fuzzy design because the probability of going to prison does not increase from zero to one, on average, when crossing the cutoff. As discussed in section 1.5, the estimation strategy for a fuzzy RD design is equivalent to an IV setup in which the first stage is given by the change in the probability of receiving the treatment when crossing the cutoff. In this sense, 2SLS is an appropriate estimation method to obtain a consistent estimate of the treatment effect for observations near the cutoff and for whom crossing the cutoff changes the treatment assignment. We implement parametric local linear regressions within a bandwidth of 16 points around the cutoff. As we have done throughout this paper, we differentiate between the effects after sentence and after release in the results that follow.

The results tables show the point estimate for the treatment of interest, i.e. the indicator for whether the individual was sentenced to prison. Each table contains two panels indicating the estimator that was used to obtain the coefficient: OLS and 2SLS with multiple instruments. We report the coefficient associated with receiving a prison sentence. This coefficient comes from a regression of the outcome on a linear specification of the PRV scores, interaction of the treatment dummy with the PRV scores, three indicators for sentence type (excluding probation) and their interactions with the PRV scores, baseline covariates, and cutoff fixed effects and their interactions with the PRV scores. Since the omitted treatment dummy is probation, the coefficient on the dummy for prison is measuring the difference, at the cutoff, in the outcome of interest between individuals sentenced to prison relative to those sentenced to probation.

The results for the new felony outcome are presented in Table 1.5. Recall that our research question seeks to find out how offenders sentenced to prison rather than probation differ in terms of future criminal behavior. The OLS estimates show that the probability of committing a new crime is negatively correlated with receiving a prison sentence within 1, 3 and 5 years after sentence. For the periods after release, there does not seem to be any difference in recidivism rates between individuals sentenced to prison or probation.

To deal with the potential endogeneity of OLS estimates, Panel B shows the 2SLS regressions using the 11 cutoffs as instrumental variables.[27] The after sentence results shows strong incapacitation effects in new-felony recidivism. After sentence, prisoners are less likely to recidivate than probationers by 18.9, 36.9, and 31.9 percentage points within 1, 3, and 5 years of being sentenced, respectively. As discussed above, these estimates are primarily capturing an incapacitation effect partially explained by prisoners serving their original prison sentence. An interesting point is that, even though almost 95 percent of those sentenced to prison have been released by the time 5 years after sentence have passed (column 4), we still see important incapacitation effects. As we will discuss below, this seems to be related to a secondary incapacitation phenomenon in which individuals sentenced to prison are significantly more likely to be re-imprisoned within a few years of the original sentence. After release we find a significant negative coefficient of 24.3 percentage points 3 years after release. This may coincide with the time in which some of the offenders sentenced to prison are sent back to prison and hence this coefficient would be capturing the secondary

[27] Since this is a Wald-type estimator, we expect that the coefficients obtained with this approach are larger in magnitude than the reduced-form coefficients because 2SLS inflates the reduced-form estimates by the size of the first stage which is less than one in this case (fuzzy RD).

incapacitation effect.

Table 1.5: LLR results (new felonies)

Outcome: Any new felony

	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
1{Sentence = prison}	-0.069***	-0.118***	-0.093***	-0.016	-0.017	-0.000
	(0.005)	(0.012)	(0.015)	(0.009)	(0.014)	(0.016)
Panel B. 2SLS - Multiple cutoffs						
1{Sentence = prison}	-0.189***	-0.369***	-0.319***	-0.025	-0.243**	-0.079
	(0.050)	(0.087)	(0.096)	(0.062)	(0.092)	(0.103)
Observations	18479	18479	18479	18383	18348	18233

Notes: Robust standard errors. The outcomes are defined as the variables in the leftmost column on the table in the time frame specified in the headings of columns 2 to 7 (e.g. any new felony within 1 year after sentence). OLS results are from regressions of each outcome on three treatment dummies which indicate the sentence type: prison, jail, and jail with probation. Regressions include a linear term on the PRV scores and the interaction between the treatment dummies and the PRV scores. In Panel B, 2SLS regressions run the same OLS regression instrumenting the dummy for prison with a dummy indicating whether the PRV score is to the right of the pooled cutoff. In Panel B, the dummy for prison is instrumented with 11 dummies indicating whether the PRV score is to the right of each of 11 cutoffs. All models include cutoff fixed effects and their interactions with the PRV scores, as well as a quadratic on age at sentence. The coefficients in the table are the point estimates an indicator variable equal to one if the offender was sentenced to prison.*** p<0.01, ** p<0.05, * p<0.1

Another outcome of interest is whether those sentenced to prison differ in terms of the severity of the new felony relative to probationers. The results from medium / high- and high-severity felony recidivism are in Table 1.6. Within each panel we show separate results for medium and high severity, and high severity only. The indicators for severity are constructed such that the high severity indicator is a subset of the medium and high severity indicator so it is possible to distinguish between the effect of the prison sentence on the new felonies in these two severity levels. Recall that the way these indicators are constructed do not condition on commiting a new felony. A value of one in this variable indicates that the offender has commited a felony in the severity level indicated, while zero includes felonies in all other severity categories as well as no new felony.

For medium / high severity felonies, Table 1.6 shows that OLS coefficients are negative post-sentence and close to zero after release. Receiving a prison sentence is associated with lower medium / high and high severity felonies after sentence but there is no association after release.

Table 1.6: LLR results (severity of new felony)

Outcome: Severity of new felony						
	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
Medium and high severity						
1{Sentence = prison}	-0.042***	-0.062***	-0.049***	-0.008	-0.001	0.009
	(0.004)	(0.010)	(0.012)	(0.007)	(0.012)	(0.014)
High severity						
1{Sentence = prison}	-0.019***	-0.033***	-0.031***	-0.006	-0.014	-0.014
	(0.003)	(0.007)	(0.009)	(0.005)	(0.008)	(0.010)
Panel B. 2SLS - Multiple cutoffs						
Medium and high severity						
1{Sentence = prison}	-0.112**	-0.086	-0.033	0.028	0.016	0.142
	(0.039)	(0.072)	(0.082)	(0.051)	(0.077)	(0.090)
High severity						
1{Sentence = prison}	0.023	0.154**	0.265***	0.142***	0.202***	0.337***
	(0.028)	(0.055)	(0.066)	(0.039)	(0.059)	(0.073)
Observations	18479	18479	18479	18383	18348	18233

Notes: Robust standard errors. The outcomes are defined as the variables in the leftmost column on the table in the time frame specified in the headings of columns 2 to 7 (e.g. comitted a high severity felony within 1 year after sentence). OLS results are from regressions of each outcome on three treatment dummies which indicate the sentence type: prison, jail, and jail with probation. Regressions include a linear term on the PRV scores and the interaction between the treatment dummies and the PRV scores. In Panel B, 2SLS regressions run the same OLS regression instrumenting the dummy for prison with a dummy indicating whether the PRV score is to the right of the pooled cutoff. In Panel B, the dummy for prison is instrumented with 11 dummies indicating whether the PRV score is to the right of each of 11 cutoffs. All models include cutoff fixed effects and their interactions with the PRV scores, as well as a quadratic on age at sentence. The coefficients in the table are the point estimates an indicator variable equa to one if the offender was sentenced to prison.*** p<0.01, ** p<0.05, * p<0.1

The multiple IV analysis provides radically different results relative to OLS. In Panel B of Table 1.6, most coefficients for medium / high severity felonies are close to zero and not statistically distinguishable from zero. That is, there is no causal effect of receiving a prison sentence vs. a probation sentence on new medium / high severity felonies. However, when we isolate high-severity felonies we see that offenders originally sentenced to prison are much more likely to be convicted of a high-severity felony across all after sentence and after release periods except one year after sentence. This finding of positive effects of the prison sentence on high-severity crimes goes against the idea of the rehabilitation or specific deterrence effects of prison as a future crime suppressor. On the contrary, it suggests that being sentenced to prison may have criminogenic effects, encouraging crimes that are more serious in nature than the original crime for which the offender was sentenced. These results are consistent with Mueller-Smith's (2016) finding that incarceration encourages criminal activity to become more serious.

Based on how we defined the offense severity indicators we can conclude that the majority of new felonies the prison sentence may have prevented during the incapacitation period would be classified as low- or medium-severity. This is because we find a lower overall felony recidivism rate among offenders sentenced to prison rather than probation in Table 1.5 but we find the opposite result when we break new felonies into high-severity vs. everything else. Hence, the negative effect of prison on recidivism is concentrated on low- or medium-severity recidivism which makes sense given the characteristics of the offenders in our sample. However, from a public safety perspective, it is worrisome that low-level offenders receiving a prison sentence are more likely to engage in future high-severity crimes than offenders who are very similar to them ex-ante but receive a probation sentence.

We now examine imprisonment in a future period in Table 1.7. Future imprisonment is a rarely-studied outcome that may have policy-relevant implications. For example, if prison sentences causally increase the likelihood of receiving a new prison sentence in the future as we find in this paper, there is a hidden-cost multiplier of this type of sentence that is likely ignored by criminal justice policy-makers. Future imprisonment can be divided into imprisonment due to a new sentence or due to a technical violation while on supervision. Hence, within each panel we present the overall imprisonment measure and disaggregate it into imprisonment due to new sentences and due to technical violations (see the headings in column 1 of each panel).

OLS results suggest that there is a negative correlation between being sentenced to prison and future incarceration within 1 year after sentence. Three years after sentence there is no association, and 5 years after sentence the correlation is positive. The negative correlation within 1 year after sentence is likely a result of incapacitation effects, as both re-imprisonment due to new sentences and technical violations are lower for individuals sentenced to prison and most offenders sentenced to prison are still incarcerated within a year of their sentence. Naturally, the after-release estimates provide a clearer picture of future imprisonment. In columns 5 to

7, OLS reports positive correlations between receiving a prison sentence and future imprisonment in all time frames analyzed. While the coefficients for new sentence are positive in this case, they are close to zero. The variable driving the positive correlation between prison and overall future imprisonment is imprisonment due to a technical violation.

The IV results are in line with the OLS pattern except that we do not see negative coefficients 1 year after sentence. The coefficients for overall future incarceration in the IV strategy are positive and statistically significant across all time frames analyzed. This finding means that, on average, individuals sentenced to prison are more likely to be imprisoned in the future than probationers across all time periods analyzed except 1 year after sentence in which there is no difference between prisoners and probationers.

We find strong evidence of greater imprisonment among those originally sentenced to prison which we call secondary incapacitation. Relative to probationers, receiving a prison sentence increases the probability of future imprisonment by 28.9 percentage points within 3 years after sentence and by 38.2 percentage points 5 years after sentence. These effects are mostly explained by higher imprisonment due to technical violations, as the coefficients on imprisonment for a new sentence are insignificant at the 5 percent level. After release, the overall effects on imprisonment are even larger (24.6 percentage points within 1 year, 39.2 percentage points within 3 years, and 43.3 percentage points 5 years after release). Even though there is an increase in imprisonment due to new sentences that is significant at the 10 percent level, most of the future imprisonment is due to violations of parole conditions. Individuals originally sentenced to prison are between 16 and 31 percentage points more likely than probationers to be imprisoned in the future due to technical violations across all post-release periods (see header “Due to technical violation” in Panel B).

The differential rates of future imprisonment due to technical violations for parolees (released prisoners) and probationers probably result from differences in the intensity of supervision.[28] Even though these individuals seem, in general, not to be commit-

[28] According to our conversations with MDOC staff, probation supervision is typically less intense than parole supervision in Michigan. This is consistent with the research literature on probation

Table 1.7: LLR results (Future Imprisonment)

Outcome: Imprisonment

	After sentence			After release		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
Overall						
1{Sentence = prison}	-0.038***	-0.007	0.043**	0.037***	0.069***	0.083***
	(0.005)	(0.012)	(0.014)	(0.010)	(0.014)	(0.015)
Due to new sentence						
1{Sentence = prison}	-0.022***	-0.026**	-0.016	0.005	0.011	0.019
	(0.003)	(0.009)	(0.011)	(0.006)	(0.011)	(0.013)
Due to technical violation						
1{Sentence = prison}	-0.017***	0.020*	0.064***	0.034***	0.069***	0.089***
	(0.004)	(0.010)	(0.012)	(0.008)	(0.012)	(0.013)
Panel B. 2SLS - Multiple cutoffs						
Overall						
1{Sentence = prison}	0.063	0.289***	0.382***	0.246***	0.392***	0.433***
	(0.041)	(0.077)	(0.087)	(0.057)	(0.082)	(0.092)
Due to new sentence						
1{Sentence = prison}	0.018	0.052	0.152*	0.090*	0.131*	0.200*
	(0.026)	(0.059)	(0.073)	(0.038)	(0.064)	(0.078)
Due to technical violation						
1{Sentence = prison}	0.047	0.263***	0.304***	0.163***	0.298***	0.314***
	(0.031)	(0.058)	(0.065)	(0.044)	(0.061)	(0.068)
Observations	18440	18440	18440	18383	18365	18306

Notes: Robust standard errors. The outcomes are defined as the variables in column 1 on the table in the time frame specified in the headings of columns 2 to 7. OLS results are from regressions of each outcome on three treatment dummies which indicate the sentence type: prison, jail, and jail with probation. Regressions include a linear term on the PRV scores and the interaction between the treatment dummies and the PRV scores. In Panel B, 2SLS regressions run the same OLS regression instrumenting the dummy for prison with a dummy indicating whether the PRV score is to the right of the pooled cutoff. In Panel B, the dummy for prison is instrumented with 11 dummies indicating whether the PRV score is to the right of each of 11 cutoffs. All models include cutoff fixed effects and their interactions with the PRV scores, as well as baseline covariates. The coefficients in the table are the point estimates an indicator variable equal to one if the offender was sentenced to prison.
*** p<0.01, ** p<0.05, * p<0.1

ting new crimes that lead to a new sentence, they are back in prison relatively soon after their release from their original sentence.[29] We believe we are the first to causally identify the secondary incapacitation effect. This effect would seem to be of high policy relevance given the high costs of a prison sentence. Due to secondary incapacitation, costs of future re-imprisonment are added to the imprisonment costs of the original sentence, creating a hidden-cost multiplier for prison sentences.

and parole discussed above.

[29] A caveat in this interpretation is that prosecutors are less likely to charge parolees with minor crimes if they can be re-imprisoned on a technical violation. Hence, some of what we observe may reflect prosecutor discretion rather than differences in offender behavior.

Overall, the general pattern across all our results is consistent with primary and secondary incapacitation effects of receiving a prison sentence. Primary incapacitation arises because a significant fraction of offenders sentenced to prison are still in prison when measuring the outcomes close to the sentence date. Conversely, at dates farther out from the original sentence, secondary incapacitation is a result of the higher future imprisonment rates for offenders originally sentenced to prison as compared to probationers. Secondary incapacitation is primarily due to imprisonment for violations of parole rather than for new sentences although imprisonment due to a technical violation could be an expedited way to send an offender back to prison when he or she in fact commited a new felony. Primary and secondary incapacitation lead to overall lower rates of new felony recidivism among those sentenced to prison. The exception is for high-severity new felonies, where receiving a prison sentence increases the probability of conviction for a high severity felony post-release.

In the next section we discuss the various robustness checks we have performed, including those related to the possible endogeneity of other sentences types (jail and jail with probation) and of sentence length, sensitivity to different bandwidths, and heaping of the running variable.

1.7 Robustness checks

One possible criticism of our approach is that controlling for the jail and jail with probation sentences invalidates the instrument because these are outcomes of the same decision process and are, hence, endogenous. We assess this possibility by examining the effects of a prison sentence compared to all intermediate sentences combined (probation, jail, and jail with probation). In this approach we no longer need to control for the jail and jail with probation sentences. Appendix Tables 1.11 to 1.13 show results when we do not control for other sentence types. In this case, the point estimate shown in the tables measures the difference between a prison sentence and other intermediate sentences (jail, jail with probation, and probation).[30] Our results

[30] Despite the fact that jail and jail with probation sentences involve a period of incarceration, this period is not generally long (see Table 1.1) and the conditions of the two forms of incarceration seem substantively different.

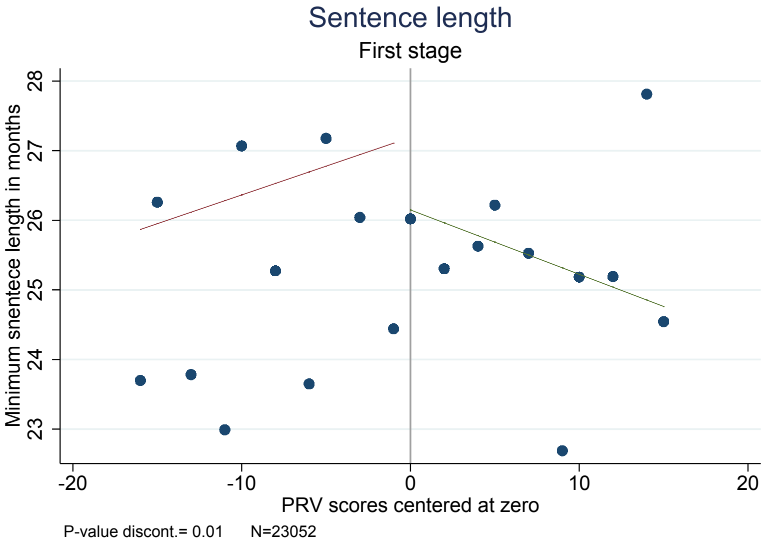
are qualitatively and quantitatively the same as when we control for jail and jail with probation. Overall, as before, being sentenced to prison relative to an intermediate sentence (including jail and jail with probation), results in primary and secondary incapacitation where the latter is explained by a higher rate of technical violations among prisoners. The only discernible difference relative to our main results is that 5 years after release both future incarceration coefficients, due to new sentence and due to technical violations, are highly significant.

One of the requirements for using the IV strategy is that the exclusion restriction holds, that is, that the instrument affects the outcome only through its effect on the endogenous variable (the probability of being sentenced to prison in our case). In fact, in the example grid in Figure 1.1 and Appendix 1.9.1 it is clear that the minimum sentence ranges are higher in straddle cells (shaded) relative to intermediate cells (marked with an asterisk). Figure 1.4 shows the first stage for sentence length within the 16-point bandwidth. There is a statistically significant reduction of about one month in sentence length at the cutoff. This observation could imply a violation to the exclusion restriction in our setting if a one month difference in time imprisoned were to affect recidivism. Given that a single month is a small fraction of the typical prison term in our sample, we do not expect this to be the case, but it is nevertheless important to examine whether this is an issue.

Without an additional source of exogenous variation in sentencing, we are not able to separately instrument sentence length. We conduct two expercises to check the sensitivity of our results to this potential violation of the exclusion restriction. First, we control for sentence length in Appendix Tables 1.14 to 1.16. There is no evidence that the results change when controlling for sentence length except that for the new felony outcome, the coefficients for 3 and 5 years after release are now statistically significant. This could be further evidence of secondary incapacitation.

Second, we exclude individuals with offenses in the highest offense level for each grid (level III for grid D, and level IV for grids E and F). Excluding these observations results in an insignificant first stage for sentence length while we only lose about 1,000 observations. The results are in Appendix Tables 1.17 to 1.19. The incapacitation effect is now significant only 1 and 3 years after sentence. While the 5-year

Figure 1.4: First stage: Sentence length (pooled cutoffs)



coefficient is still substantial in magnitude, it is no longer significant (see Panel B of Table 1.17). The rest of the results are very similar, with a slight increase in the point estimates for overall future incarceration (Panel B of Table 1.19). We now observe that 5 years after sentence and in all periods after release, offenders whose original sentence was prison are more likely to be re-imprisoned due to a new sentence (and not primarily due to technical violations of parole) relative to our main results. This finding suggests that excluding the individuals with the most severe offenses in our sample increases the importance of secondary incapacitation due to new sentences, now on par with technical violations.

Another robustness check usually conducted in the RD literature is testing sensitivity of the estimates to alternative bandwidth choices. Given the large number of outcomes we have, we perform tests for two bandwidths near our 16-point window. Appendix Tables 1.20 to 1.22 show results for a bandwidth of 15 points around the cutoff and Appendix Tables 1.23 to 1.25 for a bandwidth of 18 points around the cutoff. We do not see any unexpected behavior in the point estimates or their statistical significance.

Finally, when the running variable exhibits heaping, as is the case here, Barreca et al. (2016) recommend estimating the model using only observations at the heaps. We do so using the heaps at multiples of 5 within the 16-point bandwidth. These are the values of the running variable where most of our observations are concentrated. Appendix Tables 1.26 to 1.28 show that the results are virtually the same as in our main specification.

1.8 Conclusion

This paper leverages discontinuities in the probability of being sentenced to prison arising from the structure of the Michigan Sentencing Guidelines to provide new estimates of the effects of imprisonment on future recidivism. The sentencing guidelines provide a framework in which low-level offenders (classified in low-severity crime classes) may receive a prison sentence if their prior record score is at or above a certain cutoff determined by the specific grid and the offense severity level. This setup leads to a fuzzy RD design in which the running variable is the prior record score that measures offenders' criminal history.

To estimate the causal effect of receiving a prison sentence relative to probation, we perform reduced-form and instrumental-variable analyses based on the logic of instrumental variables applied to the fuzzy RD setting. We take advantage of the fact that each individual in our setting is only affected by one cutoff. We use the 11 cutoffs in Grids D, E, and F of the Michigan Sentencing Guidelines as instrumental variables that take the value of one if the prior record score of the offender is to the right of the cutoff. Our estimates provide the LATE for the population of individuals whose sentences are affected by crossing the boundary between cells in the guideline grid, weighted by the strength of the first stage of each of the cutoffs.

When comparing offenders sentenced to prison vs. probation, we identify significant incapacitation effects from the original prison sentence as well as those resulting from future imprisonment, which our results show are primarily due to technical violations of parole. We call these incapacitation effects primary and secondary, respectively. While primary incapacitation is well documented in the literature, we believe we are the first to causally identify secondary incapacitation as an effect of the original prison sentence, that is, to estimate the causal effect of a prison sentence on future imprisonment. Receiving a prison sentence reduces the likelihood of committing a new felony by at least 19 percentage points within the first year after sentence and by between 32 and 37 percentage points within 3 and 5 years after sentence, respectively, relative to an offender sentenced to probation. After release, our evidence suggests that offenders sentenced to prison are less likely to be convicted of a new felony, particularly within 3 years after release. Importantly, among those who in fact commit a new felony, those whose original sentence was prison are significantly more likely to commit a high severity felony. In terms of future imprisonment, we find that those sentenced to prison are more likely to be incarcerated in the future relatively soon after their original sentence date. For most of our specifications, the channel for this secondary incapacitation is re-imprisonment resulting from technical violations of parole. Hence, our results suggest that, at the cutoff, the main effect of sentencing offenders to prison operates through incapacitation as offenders sentenced to prison serve time not only from their original sentence but also from returning to prison due to technical violations. If this finding were to generalize to other states, this would be an important for criminal justice policy, as it suggests that sentencing offenders on the margin between prison and probation to prison primarily reduces their average future offending during the time they spend in prison.

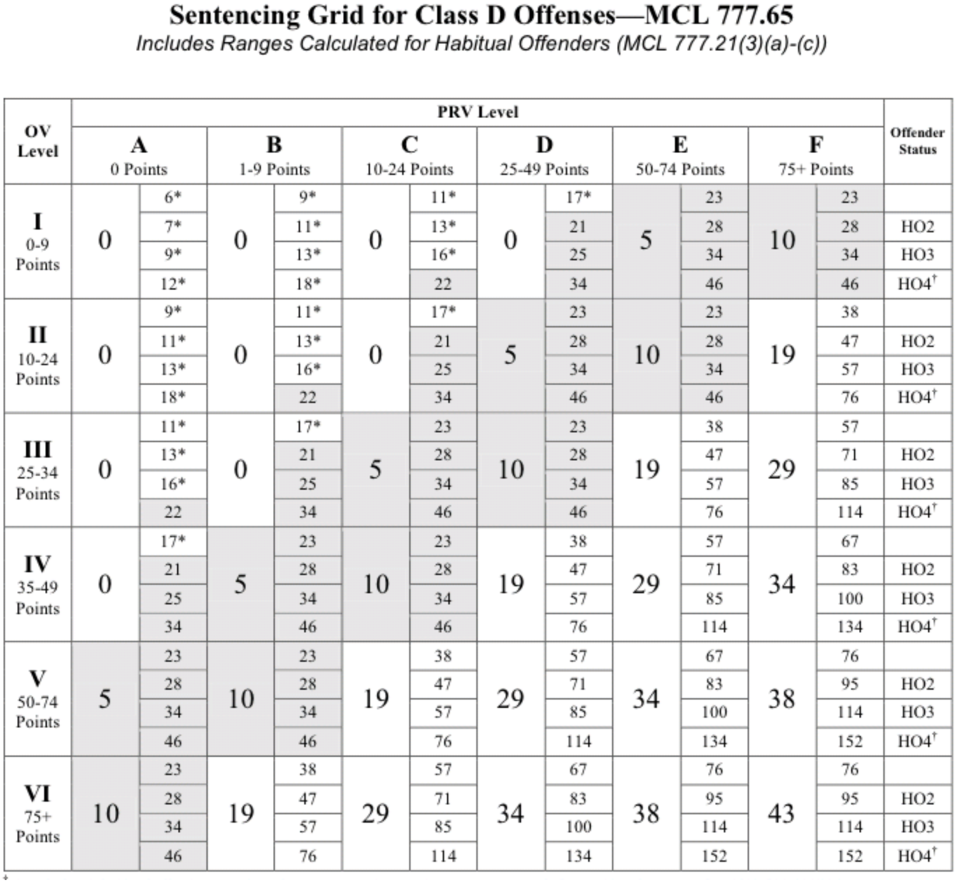
The specific policy implications of our findings depend on how one interprets the effects of a prison sentence on the probability of future imprisonment in conjunction with its effects on new felonies after release. One interpretation is that secondary incapacitation is preventing new felony convictions among those originally sentenced to prison. A second interpretation is that future imprisonment resulting from technical violations is due to greater surveillance of individuals on parole relative to those on probation. The difference between the two depends on what one believes re-imprisoned parolees would have done had they not been re-imprisoned on technical violations. The first interpretation suggests they would indeed have committed felonies and been prosecuted for them, implying that secondary incapacitation is crime preventative. The second suggests they would commit minor crimes and parole violations, implying that re-imprisonment is creating significant incarceration costs while preventing little serious crime. We cannot adjudicate between these two different counterfactuals with our data, as we do not have an identification strategy for estimating the causal effect of re-imprisonment for a technical violation on future crime. Our findings suggest that addressing this question in future research is imperative for any cost-benefit analysis of imprisonment and for providing definitive policy recommendations.

Finally, we note several limitations of this study. First, our analysis is focused on offenders whose sentence type is affected by a marginal increase in their prior record score. In that sense, our results are local to narrow window around the cutoffs determining sentence type. An additional limitation is that we can only assess re-offending based on offending known to law enforcement. Furthermore, our analysis is limited to a single state, and social and economic conditions as well as criminal justice policies vary considerably from state to state. In terms of criminal justice policies and practices, we note that Michigan's rates of incarceration and parole are close to the national averages. Michigan also accounts for a nontrivial share of the nation's prisoner population. However, our findings may be sensitive to state-specific resources and policies related to prison administration, and probation or parole supervision and revocation. Future work should examine whether results generalize to other contexts and other identification strategies.

1.9 Appendices

1.9.1 SGL grid example

Figure 1.5: Grid D from Michigan Sentencing Guidelines



Sentencing Grid for Class D Offenses—MCL 777.65

Includes Ranges Calculated for Habitual Offenders (MCL 777.21(3)(a)-(c))

OV Level	PRV Level												Offender Status
	A 0 Points		B 1-9 Points		C 10-24 Points		D 25-49 Points		E 50-74 Points		F 75+ Points		
I 0-9 Points	0	6*	0	9*	0	11*	0	17*	5	23	10	23	
		7*		11*		13*		21		28		28	HO2
		9*		13*		16*		25		34		34	HO3
		12*		18*		22		34		46		46	HO4†
II 10-24 Points	0	9*	0	11*	0	17*	5	23	10	23	19	38	
		11*		13*		21		28		28		47	HO2
		13*		16*		25		34		34		57	HO3
		18*		22		34		46		46		76	HO4†
III 25-34 Points	0	11*	0	17*	5	23	10	23	19	38	29	57	
		13*		21		28		28		47		71	HO2
		16*		25		34		34		57		85	HO3
		22		34		46		46		76		114	HO4†
IV 35-49 Points	0	17*	5	23	10	23	19	38	29	57	34	67	
		21		28		28		47		71		83	HO2
		25		34		34		57		85		100	HO3
		34		46		46		76		114		134	HO4†
V 50-74 Points	5	23	10	23	19	38	29	57	34	67	38	76	
		28		28		47		71		83		95	HO2
		34		34		57		85		100		114	HO3
		46		46		76		114		134		152	HO4†
VI 75+ Points	10	23	19	38	29	57	34	67	38	76	43	76	
		28		47		71		83		95		95	HO2
		34		57		85		100		114		114	HO3
		46		76		114		134		152		152	HO4†

†

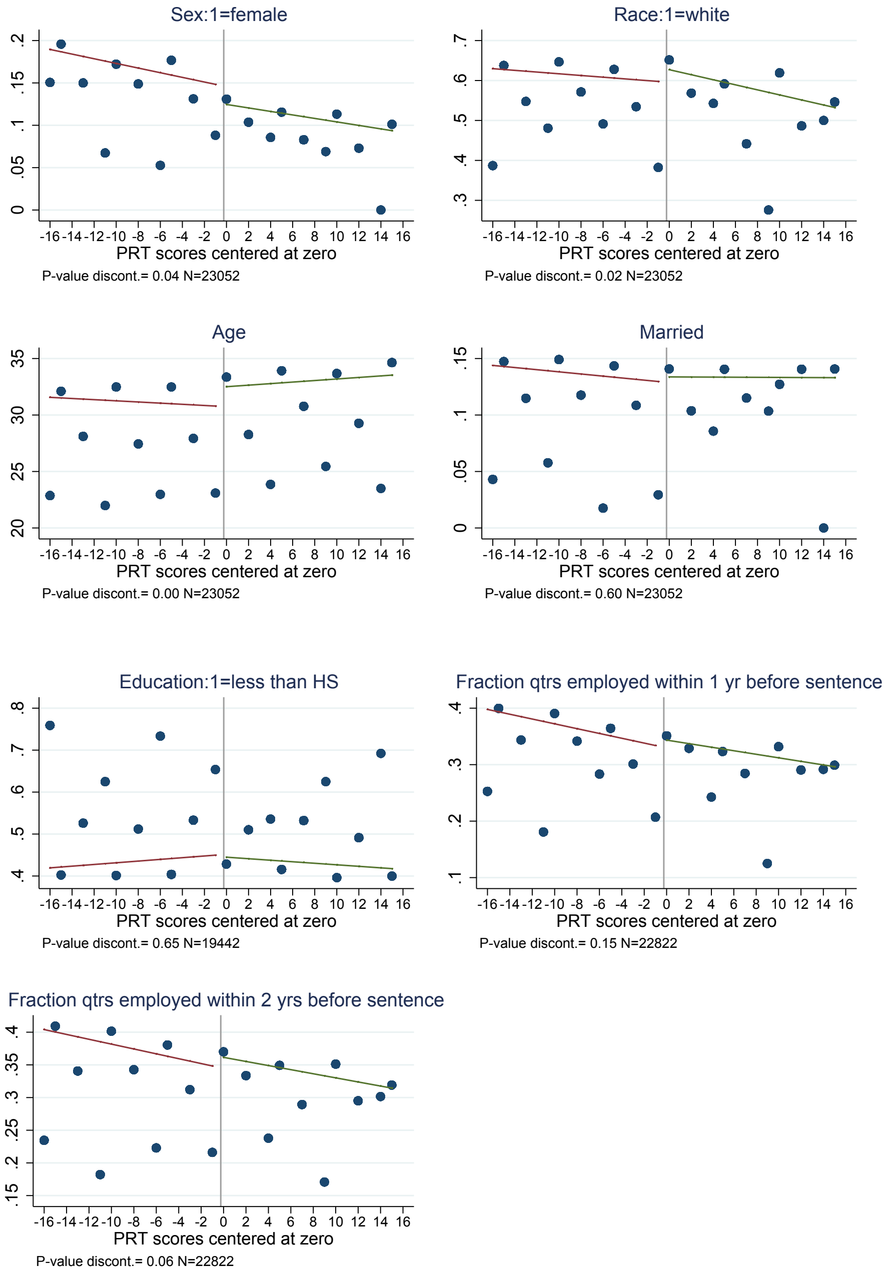
1.9.2 Joint test of covariates at the cutoff (SUR tests)

Table 1.8: Joint test of covariates at the cutoff (SUR tests)

Bandwidth	**F(12,∼12·N)**	**p-value**	**N**	**Bandwidth**	**F(12,∼12·N)**	**p-value**	**N**
+/- 1	1.355	0.180	1,793	+/- 13	3.628	0.000	13,431
+/- 2	1.355	0.180	1,793	+/- 14	3.628	0.000	13,431
+/- 3	0.767	0.686	2,779	+/- 15	1.171	0.298	17,471
+/- 4	0.767	0.686	2,779	+/- 16	1.197	0.278	18,479
+/- 5	4.518	0.000	5,301	+/- 17	1.197	0.278	18,479
+/- 6	3.457	0.000	6,674	+/- 18	1.168	0.300	20,358
+/- 7	3.457	0.000	6,674	+/- 19	1.168	0.300	20,358
+/- 8	1.638	0.074	7,910	+/- 20	1.714	0.057	23,713
+/- 9	1.638	0.074	7,910	+/- 21	1.772	0.047	24,580
+/- 10	4.176	0.000	11,024	+/- 22	1.772	0.047	24,580
+/- 11	3.864	0.000	12,201	+/- 23	1.937	0.026	28,017
+/- 12	3.864	0.000	12,201	+/- 24	1.937	0.026	28,017

1.9.3 Graphs of covariates

Figure 1.6: Balance of covariates (1)



Sex:1=female

P-value discont.= 0.04 N=23052

Race:1=white

P-value discont.= 0.02 N=23052

Age

P-value discont.= 0.00 N=23052

Married

P-value discont.= 0.60 N=23052

Education:1=less than HS

P-value discont.= 0.65 N=19442

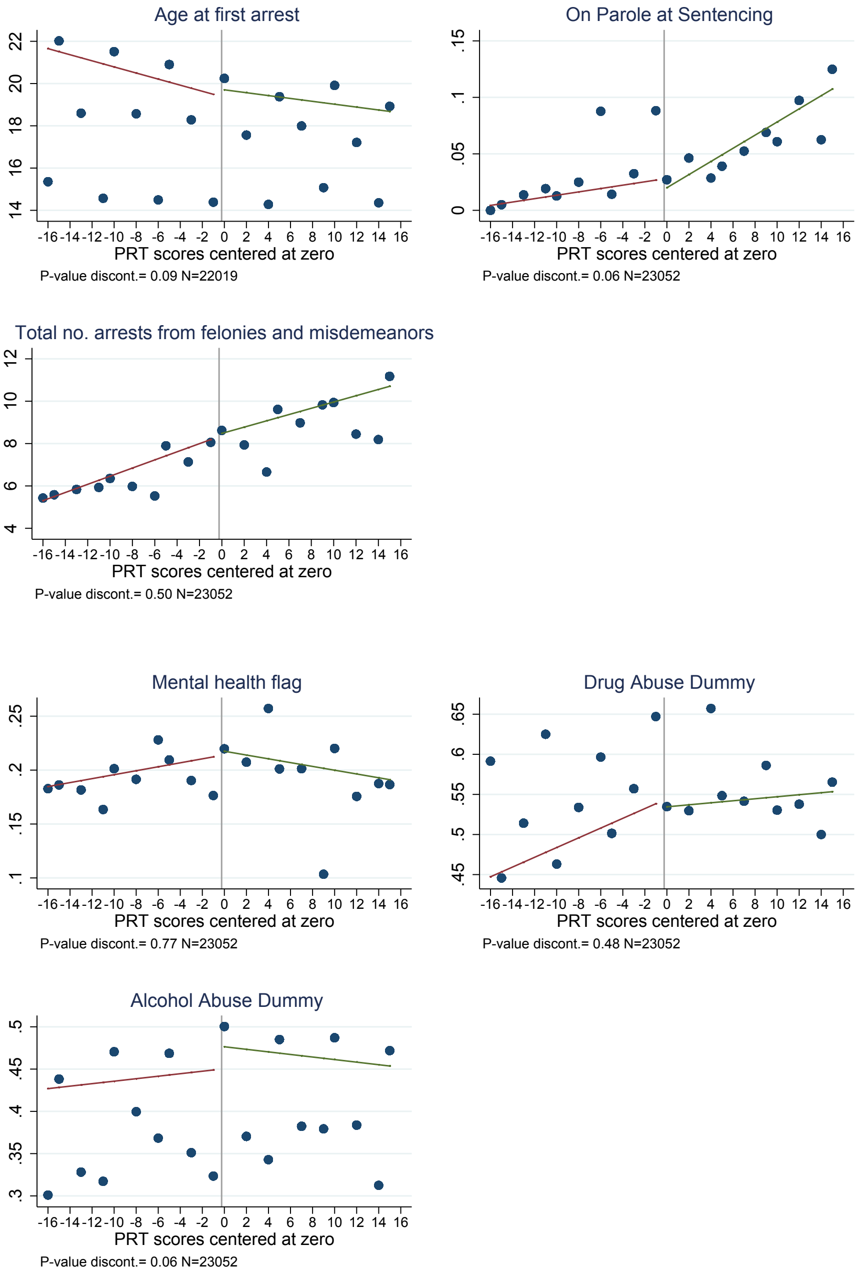
Fraction qtrs employed within 1 yr before sentence

P-value discont.= 0.15 N=22822

Fraction qtrs employed within 2 yrs before sentence

P-value discont.= 0.06 N=22822

Figure 1.7: Balance of covariates (2)


Age at first arrest

P-value discont.= 0.09 N=22019

On Parole at Sentencing

P-value discont.= 0.06 N=23052

Total no. arrests from felonies and misdemeanors

P-value discont.= 0.50 N=23052

Mental health flag

P-value discont.= 0.77 N=23052

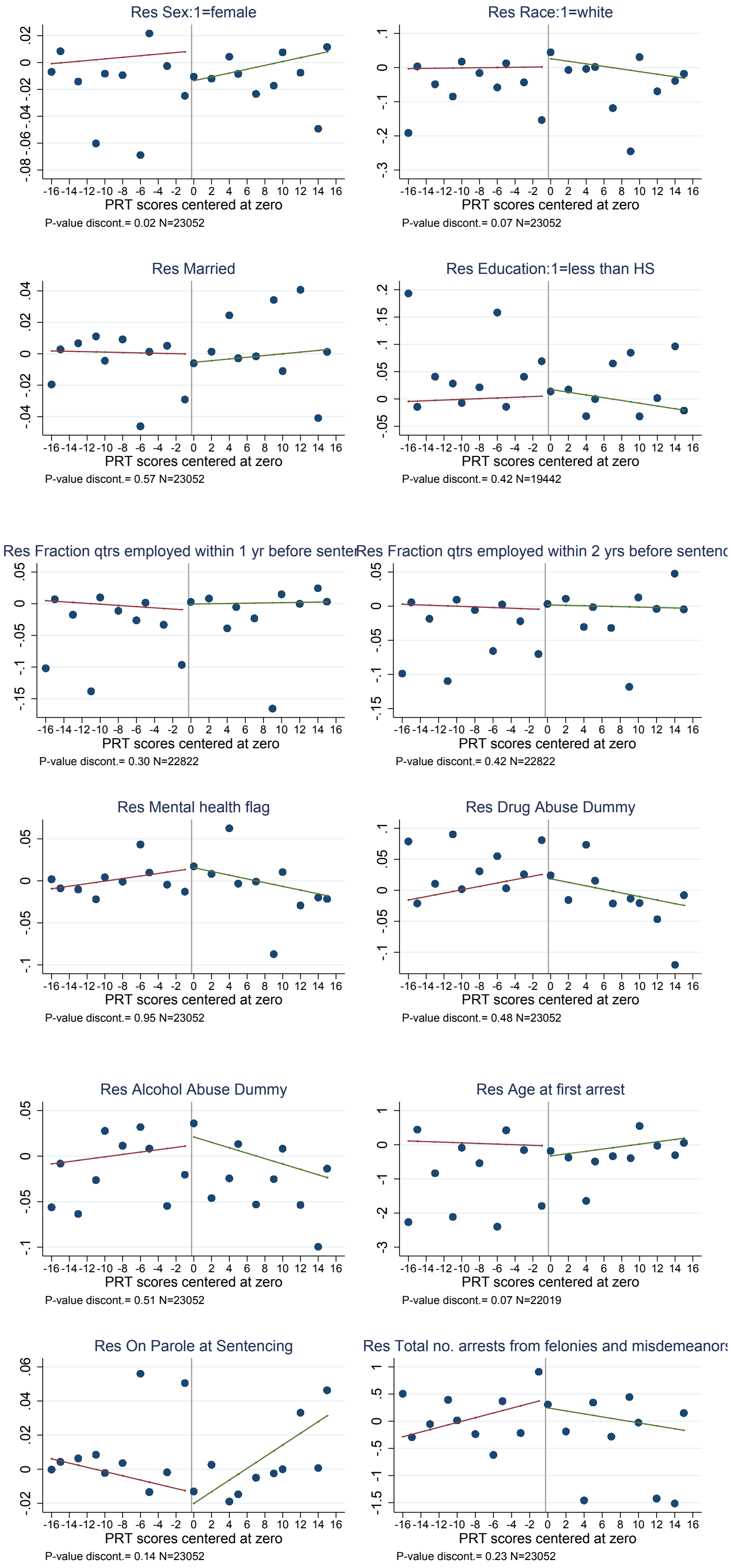
Drug Abuse Dummy

P-value discont.= 0.48 N=23052

Alcohol Abuse Dummy

P-value discont.= 0.06 N=23052

Figure 1.8: Balance of covariates residualized on age



1.9.4 Descriptive statistics of all sentence types

Table 1.9: Descriptive statistics - all sentence types

	Sentence type			
	Probation	**Prison**	**Jail**	**Jail with probation**
% of observations in sample	0.30	0.10	0.08	0.52
% of women	0.19	0.09	0.12	0.15
Age at sentence	31.00	32.99	32.87	31.99
% white	0.48	0.58	0.58	0.69
% married	0.14	0.13	0.12	0.14
% with less than high school	0.45	0.43	0.47	0.42
Age at first arrest	20.44	19.47	19.67	20.44
On parole at sentence	0.01	0.15	0.06	0.01
Total number of arrests before sentence	6.40	9.89	8.23	7.51
% with mental illness	0.18	0.20	0.20	0.21
% with drug addiction	0.49	0.53	0.52	0.50
% with alcohol addiction	0.32	0.49	0.42	0.52
Months employed within a year before sentence	4.20	3.60	3.24	4.56
Months employed within 2 years before sentence	8.64	7.92	6.96	9.36
Average sentence length (months)	26.87	17.54	6.77	30.35
Time served (months in prison)		22.13		

Notes: All figures correspond to means of the variables within 16 points from the cutoff. Less than high school does not include GED.

Table 1.10: Descriptive statistics of outcomes of interest - all sentence types

	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. Any new felony						
Probation	0.07	0.23	0.32	0.07	0.23	0.32
Prison	0.00	0.12	0.25	0.06	0.24	0.36
Jail	0.06	0.24	0.32	0.10	0.26	0.33
Jail with probation	0.05	0.21	0.30	0.08	0.23	0.31
Panel B. Medium and high severity felonies						
Probation	0.05	0.14	0.19	0.05	0.14	0.19
Prison	0.00	0.08	0.16	0.04	0.15	0.22
Jail	0.04	0.14	0.18	0.06	0.15	0.19
Jail with probation	0.03	0.13	0.19	0.05	0.14	0.19
Panel B. High severity felonies						
Probation	0.02	0.07	0.09	0.02	0.07	0.09
Prison	0.00	0.04	0.07	0.02	0.07	0.10
Jail	0.02	0.08	0.10	0.03	0.08	0.10
Jail with probation	0.01	0.06	0.07	0.02	0.06	0.08
Panel D. Future incarceration						
Probation	0.03	0.11	0.16	0.03	0.11	0.16
Prison	0.01	0.16	0.28	0.11	0.28	0.34
Jail	0.03	0.12	0.17	0.05	0.13	0.18
Jail with probation	0.04	0.15	0.20	0.07	0.16	0.20
Panel E. Future incarceration due to new sentences						
Probation	0.02	0.08	0.12	0.02	0.08	0.12
Prison	0.00	0.06	0.14	0.03	0.13	0.19
Jail	0.02	0.09	0.13	0.04	0.10	0.14
Jail with probation	0.01	0.07	0.11	0.02	0.08	0.12
Panel F. Future incarceration due to technical violations						
Probation	0.01	0.04	0.05	0.01	0.04	0.05
Prison	0.01	0.10	0.17	0.08	0.17	0.21
Jail	0.01	0.04	0.05	0.02	0.04	0.05
Jail with probation	0.03	0.08	0.11	0.04	0.09	0.11

Notes: All figures correspond to observations within 16 points from the cutoff. Sample sizes are around 18,000 observations depending on the variable.

1.9.5 Robustness checks

Table 1.11: Robustness check: Prison vs. intermediate sentences (Any new felony)

	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
Dummy prison	-0.058***	-0.115***	-0.094***	-0.028***	-0.025*	-0.009
	(0.003)	(0.010)	(0.013)	(0.008)	(0.013)	(0.015)
Panel B. 2SLS - Multiple cutoffs						
Dummy prison	-0.166**	-0.325***	-0.258**	0.012	-0.168	0.012
	(0.051)	(0.089)	(0.098)	(0.065)	(0.095)	(0.109)
Observations	18479	18479	18479	18383	18348	18233

Notes: Robust standard errors. *** p<0.01, ** p<0.05, * p<0.1

Table 1.12: Robustness check: Prison vs. intermediate sentences (Severity of new felony)

	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
Medium and high severity						
Dummy prison	-0.035***	-0.065***	-0.053***	-0.016*	-0.011	0.001
	(0.003)	(0.009)	(0.011)	(0.006)	(0.011)	(0.013)
High severity						
Dummy prison	-0.015***	-0.036***	-0.034***	-0.010*	-0.020**	-0.019*
	(0.002)	(0.006)	(0.007)	(0.004)	(0.007)	(0.009)
Panel B. 2SLS - Multiple cutoffs						
Medium and high severity						
Dummy prison	-0.113**	-0.088	-0.057	0.044	0.028	0.149
	(0.040)	(0.074)	(0.085)	(0.054)	(0.080)	(0.096)
High severity						
Dummy prison	0.031	0.180**	0.311***	0.172***	0.250***	0.411***
	(0.029)	(0.059)	(0.072)	(0.044)	(0.065)	(0.083)
Observations	18479	18479	18479	18383	18348	18233

Notes: Robust standard errors. *** p<0.01, ** p<0.05, * p<0.1

Table 1.13: Robustness check: Prison vs. intermediate sentences (Future incarceration)

	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
Overall						
Dummy prison	-0.039***	-0.027*	0.019	0.016	0.044***	0.057***
	(0.004)	(0.011)	(0.013)	(0.009)	(0.013)	(0.014)
Due to new sentence						
Dummy prison	-0.017***	-0.027***	-0.016	0.001	0.007	0.015
	(0.002)	(0.008)	(0.010)	(0.005)	(0.010)	(0.011)
Due to technical violation						
Dummy prison	-0.022***	0.001	0.039***	0.016*	0.046***	0.064***
	(0.004)	(0.009)	(0.011)	(0.008)	(0.011)	(0.012)
Panel B. 2SLS - Multiple cutoffs						
Overall						
Dummy prison	0.078	0.299***	0.412***	0.281***	0.423***	0.479***
	(0.043)	(0.081)	(0.093)	(0.062)	(0.088)	(0.100)
Due to new sentence						
Dummy prison	0.021	0.056	0.177*	0.106*	0.161*	0.243**
	(0.029)	(0.061)	(0.077)	(0.042)	(0.068)	(0.084)
Due to technical violation						
Dummy prison	0.058	0.267***	0.318***	0.178***	0.302***	0.327***
	(0.033)	(0.061)	(0.069)	(0.046)	(0.065)	(0.073)
Observations	18440	18440	18440	18383	18365	18306

Notes: Robust standard errors. The outcomes are defined as the variables in the leftmost column on the table in the time frame specified in the headings of columns 2 to 7 (e.g. returned to prison within 1 year after sentence). OLS results are from regressions of each outcome on three treatment dummies which indicate the sentence type: prison, jail, and jail with probation. Regressions include a linear term on the PRV scores and the interaction between the treatment dummies and the PRV scores. In Panel B, 2SLS regressions run the same OLS regression instrumenting the dummy for prison with a dummy indicating whether the PRV score is to the right of the pooled cutoff. In Panel B, the dummy for prison is instrumented with 11 dummies indicating whether the PRV score is to the right of each of 11 cutoffs. All models include cutoff fixed effects and their interactions with the PRV scores, as well as a quadratic on age at sentence. The coefficients in the table are the point estimates an indicator variable equal to one if the offender was sentenced to prison.*** p<0.01, ** p<0.05, * p<0.1

Table 1.14: Robustness check: Controlling for sentence length (Any new felony)

	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
Dummy prison	-0.076***	-0.132***	-0.106***	-0.025**	-0.028	-0.012
	(0.006)	(0.012)	(0.015)	(0.009)	(0.015)	(0.017)
Panel B. 2SLS - Multiple cutoffs						
Dummy prison	-0.200***	-0.416***	-0.403***	-0.060	-0.315***	-0.200*
	(0.047)	(0.081)	(0.089)	(0.058)	(0.086)	(0.097)
Observations	18479	18479	18479	18383	18348	18233

Notes: Robust standard errors. *** p<0.01, ** p<0.05, * p<0.1

Table 1.15: Robustness check: Controlling for sentence length (Severity of new felony)

	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
Medium and high severity						
Dummy prison	-0.046***	-0.071***	-0.056***	-0.012	-0.008	0.004
	(0.005)	(0.010)	(0.013)	(0.007)	(0.012)	(0.015)
High severity						
Dummy prison	-0.019***	-0.037***	-0.034***	-0.007	-0.017*	-0.018
	(0.003)	(0.007)	(0.009)	(0.005)	(0.008)	(0.010)
Panel B. 2SLS - Multiple cutoffs						
Medium and high severity						
Dummy prison	-0.135***	-0.158*	-0.143	-0.017	-0.077	0.003
	(0.037)	(0.067)	(0.077)	(0.048)	(0.072)	(0.084)
High severity						
Dummy prison	0.010	0.115*	0.208***	0.113**	0.156**	0.277***
	(0.025)	(0.051)	(0.060)	(0.035)	(0.055)	(0.067)
Observations	18479	18479	18479	18383	18348	18233

Notes: Robust standard errors. *** p<0.01, ** p<0.05, * p<0.1

Table 1.16: Robustness check: Controlling for sentence length (Future incarceration)

	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
Overall						
Dummy prison	-0.032***	0.004	0.060***	0.047***	0.082***	0.103***
	(0.005)	(0.012)	(0.015)	(0.010)	(0.014)	(0.016)
Due to new sentence						
Dummy prison	-0.025***	-0.036***	-0.027*	0.001	0.000	0.010
	(0.003)	(0.009)	(0.011)	(0.006)	(0.011)	(0.013)
Due to technical violation						
Dummy prison	-0.007	0.041***	0.091***	0.047***	0.091***	0.118***
	(0.004)	(0.010)	(0.012)	(0.008)	(0.012)	(0.013)
Panel B. 2SLS - Multiple cutoffs						
Overall						
Dummy prison	0.042	0.221**	0.286***	0.208***	0.309***	0.325***
	(0.037)	(0.070)	(0.079)	(0.052)	(0.075)	(0.084)
Due to new sentence						
Dummy prison	0.008	0.014	0.086	0.067	0.075	0.122
	(0.024)	(0.054)	(0.067)	(0.035)	(0.059)	(0.071)
Due to technical violation						
Dummy prison	0.036	0.230***	0.261***	0.147***	0.265***	0.272***
	(0.029)	(0.052)	(0.059)	(0.040)	(0.056)	(0.062)
Observations	18440	18440	18440	18383	18365	18306

Notes: Robust standard errors. The outcomes are defined as the variables in the leftmost column on the table in the time frame specified in the headings of columns 2 to 7 (e.g. returned to prison within 1 year after sentence). OLS results are from regressions of each outcome on three treatment dummies which indicate the sentence type: prison, jail, and jail with probation. Regressions include a linear term on the PRV scores and the interaction between the treatment dummies and the PRV scores. In Panel B, 2SLS regressions run the same OLS regression instrumenting the dummy for prison with a dummy indicating whether the PRV score is to the right of the pooled cutoff. In Panel B, the dummy for prison is instrumented with 11 dummies indicating whether the PRV score is to the right of each of 11 cutoffs. All models include cutoff fixed effects and their interactions with the PRV scores, as well as a quadratic on age at sentence. The coefficients in the table are the point estimates an indicator variable equal to one if the offender was sentenced to prison.*** p<0.01, ** p<0.05, * p<0.1

Table 1.17: Robustness check: Excluding highest OV level (Any new felony)

	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
Dummy prison	-0.074***	-0.128***	-0.100***	-0.017	-0.023	-0.006
	(0.006)	(0.013)	(0.016)	(0.009)	(0.015)	(0.017)
Panel B. 2SLS - Multiple cutoffs						
Dummy prison	-0.166**	-0.299***	-0.187	0.024	-0.158	-0.006
	(0.052)	(0.090)	(0.099)	(0.066)	(0.095)	(0.105)
Observations	17038	17038	17038	16953	16927	16842

Notes: Robust standard errors. *** p<0.01, ** p<0.05, * p<0.1

Table 1.18: Robustness check: Excluding highest OV level (Severity of new felony)

	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
Medium and high severity						
Dummy prison	-0.046***	-0.068***	-0.054***	-0.010	-0.007	0.003
	(0.005)	(0.011)	(0.013)	(0.008)	(0.013)	(0.015)
High severity						
Dummy prison	-0.020***	-0.036***	-0.034***	-0.008	-0.016	-0.017
	(0.003)	(0.007)	(0.009)	(0.005)	(0.009)	(0.010)
Panel B. 2SLS - Multiple cutoffs						
Medium and high severity						
Dummy prison	-0.092*	-0.019	0.072	0.067	0.092	0.200*
	(0.041)	(0.075)	(0.087)	(0.054)	(0.080)	(0.092)
High severity						
Dummy prison	0.029	0.180**	0.312***	0.156***	0.232***	0.359***
	(0.029)	(0.058)	(0.070)	(0.041)	(0.063)	(0.075)
Observations	17038	17038	17038	16953	16927	16842

Notes: Robust standard errors. *** p<0.01, ** p<0.05, * p<0.1

Table 1.19: Robustness check: Excluding highest OV level (Future incarceration)

Outcome: Returned to prison	**After sentence**			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
Overall						
Dummy prison	-0.042***	-0.009	0.045**	0.041***	0.074***	0.087***
	(0.006)	(0.013)	(0.015)	(0.011)	(0.015)	(0.016)
Due to new sentence						
Dummy prison	-0.024***	-0.029**	-0.017	0.006	0.012	0.019
	(0.003)	(0.010)	(0.012)	(0.007)	(0.012)	(0.014)
Due to technical violation						
Dummy prison	-0.018***	0.021*	0.068***	0.035***	0.074***	0.095***
	(0.005)	(0.011)	(0.013)	(0.009)	(0.013)	(0.014)
Panel B. 2SLS - Multiple cutoffs						
Overall						
Dummy prison	0.068	0.354***	0.471***	0.276***	0.469***	0.515***
	(0.043)	(0.082)	(0.093)	(0.061)	(0.087)	(0.097)
Due to new sentence						
Dummy prison	0.027	0.105	0.231**	0.113**	0.191**	0.269**
	(0.028)	(0.062)	(0.078)	(0.041)	(0.068)	(0.082)
Due to technical violation						
Dummy prison	0.045	0.277***	0.330***	0.170***	0.322***	0.342***
	(0.033)	(0.061)	(0.069)	(0.046)	(0.065)	(0.071)
Observations	17001	17001	17001	16953	16941	16896

Notes: Robust standard errors. The outcomes are defined as the variables in the leftmost column on the table in the time frame specified in the headings of columns 2 to 7 (e.g. returned to prison within 1 year after sentence). OLS results are from regressions of each outcome on three treatment dummies which indicate the sentence type: prison, jail, and jail with probation. Regressions include a linear term on the PRV scores and the interaction between the treatment dummies and the PRV scores. In Panel B, 2SLS regressions run the same OLS regression instrumenting the dummy for prison with a dummy indicating whether the PRV score is to the right of the pooled cutoff. In Panel B, the dummy for prison is instrumented with 11 dummies indicating whether the PRV score is to the right of each of 11 cutoffs. All models include cutoff fixed effects and their interactions with the PRV scores, as well as a quadratic on age at sentence. The coefficients in the table are the point estimates an indicator variable equal to one if the offender was sentenced to prison.*** p<0.01, ** p<0.05, * p<0.1

Table 1.20: Robustness check: Bandwidth=+/-15 (Any new felony)

	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
Dummy prison	-0.070***	-0.120***	-0.099***	-0.016	-0.022	-0.007
	(0.006)	(0.013)	(0.015)	(0.009)	(0.015)	(0.017)
Panel B. 2SLS - Multiple cutoffs						
Dummy prison	-0.178***	-0.322***	-0.255**	-0.015	-0.192*	-0.044
	(0.049)	(0.086)	(0.095)	(0.062)	(0.091)	(0.103)
Observations	17471	17471	17471	17392	17363	17264

Notes: Robust standard errors. *** p<0.01, ** p<0.05, * p<0.1

Table 1.21: Robustness check: Bandwidth=+/-15 (Severity of new felony)

	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
Medium and high severity						
Dummy prison	-0.043***	-0.063***	-0.052***	-0.009	-0.003	0.006
	(0.005)	(0.011)	(0.013)	(0.008)	(0.013)	(0.015)
High severity						
Dummy prison	-0.020***	-0.036***	-0.035***	-0.007	-0.016	-0.018
	(0.003)	(0.007)	(0.009)	(0.005)	(0.009)	(0.010)
Panel B. 2SLS - Multiple cutoffs						
Medium and high severity						
Dummy prison	-0.118**	-0.081	-0.028	0.023	0.017	0.125
	(0.038)	(0.072)	(0.082)	(0.051)	(0.077)	(0.090)
High severity						
Dummy prison	0.014	0.162**	0.255***	0.128***	0.195**	0.329***
	(0.027)	(0.056)	(0.066)	(0.039)	(0.060)	(0.073)
Observations	17471	17471	17471	17392	17363	17264

Notes: Robust standard errors. *** p<0.01, ** p<0.05, * p<0.1

Table 1.22: Robustness check: Bandwidth=+/-15 (Future incarceration)

Outcome: Returned to prison						
	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
Overall						
Dummy prison	-0.041***	-0.011	0.038*	0.030**	0.063***	0.077***
	(0.006)	(0.013)	(0.015)	(0.010)	(0.015)	(0.016)
Due to new sentence						
Dummy prison	-0.023***	-0.027**	-0.020	0.004	0.008	0.015
	(0.003)	(0.009)	(0.012)	(0.006)	(0.011)	(0.013)
Due to technical violation						
Dummy prison	-0.018***	0.017	0.059***	0.028***	0.063***	0.085***
	(0.005)	(0.010)	(0.012)	(0.008)	(0.012)	(0.013)
Panel B. 2SLS - Multiple cutoffs						
Overall						
Dummy prison	0.054	0.317***	0.393***	0.282***	0.381***	0.434***
	(0.040)	(0.078)	(0.088)	(0.057)	(0.083)	(0.093)
Due to new sentence						
Dummy prison	0.013	0.049	0.171*	0.081*	0.119	0.219**
	(0.026)	(0.059)	(0.073)	(0.038)	(0.064)	(0.078)
Due to technical violation						
Dummy prison	0.043	0.291***	0.322***	0.210***	0.306***	0.326***
	(0.030)	(0.058)	(0.066)	(0.043)	(0.062)	(0.069)
Observations	17439	17439	17439	17392	17378	17325

Notes: Robust standard errors. The outcomes are defined as the variables in the leftmost column on the table in the time frame specified in the headings of columns 2 to 7 (e.g. returned to prison within 1 year after sentence). OLS results are from regressions of each outcome on three treatment dummies which indicate the sentence type: prison, jail, and jail with probation. Regressions include a linear term on the PRV scores and the interaction between the treatment dummies and the PRV scores. In Panel B, 2SLS regressions run the same OLS regression instrumenting the dummy for prison with a dummy indicating whether the PRV score is to the right of the pooled cutoff. In Panel B, the dummy for prison is instrumented with 11 dummies indicating whether the PRV score is to the right of each of 11 cutoffs. All models include cutoff fixed effects and their interactions with the PRV scores, as well as a quadratic on age at sentence. The coefficients in the table are the point estimates an indicator variable equal to one if the offender was sentenced to prison.*** p<0.01, ** p<0.05, * p<0.1

Table 1.23: Robustness check: Bandwidth=+/-18 (Any new felony)

	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
Dummy prison	-0.071***	-0.124***	-0.100***	-0.018*	-0.024	-0.008
	(0.005)	(0.012)	(0.014)	(0.008)	(0.014)	(0.016)
Panel B. 2SLS - Multiple cutoffs						
Dummy prison	-0.174***	-0.331***	-0.271**	-0.013	-0.212*	-0.020
	(0.048)	(0.084)	(0.093)	(0.061)	(0.090)	(0.101)
Observations	20358	20358	20358	20254	20216	20089

Notes: Robust standard errors. *** p<0.01, ** p<0.05, * p<0.1

Table 1.24: Robustness check: Bandwidth=+/-18 (Severity of new felony)

	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
Medium and high severity						
Dummy prison	-0.044***	-0.066***	-0.055***	-0.009	-0.006	0.003
	(0.004)	(0.010)	(0.012)	(0.007)	(0.012)	(0.014)
High severity						
Dummy prison	-0.021***	-0.034***	-0.032***	-0.008	-0.015	-0.016
	(0.003)	(0.007)	(0.008)	(0.005)	(0.008)	(0.009)
Panel B. 2SLS - Multiple cutoffs						
Medium and high severity						
Dummy prison	-0.102**	-0.053	0.005	0.033	0.046	0.191*
	(0.038)	(0.070)	(0.081)	(0.050)	(0.076)	(0.090)
High severity						
Dummy prison	0.038	0.170**	0.286***	0.159***	0.222***	0.377***
	(0.027)	(0.055)	(0.065)	(0.039)	(0.060)	(0.074)
Observations	20358	20358	20358	20254	20216	20089

Notes: Robust standard errors. *** p<0.01, ** p<0.05, * p<0.1

Table 1.25: Robustness check: Bandwidth=+/-18 (Future incarceration)

Outcome: Returned to prison	**After sentence**			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
Overall						
Dummy prison	-0.039***	-0.009	0.038**	0.038***	0.066***	0.079***
	(0.005)	(0.012)	(0.014)	(0.009)	(0.014)	(0.015)
Due to new sentence						
Dummy prison	-0.022***	-0.029***	-0.020	0.004	0.007	0.014
	(0.003)	(0.009)	(0.011)	(0.006)	(0.010)	(0.012)
Due to technical violation						
Dummy prison	-0.017***	0.021*	0.063***	0.034***	0.068***	0.089***
	(0.004)	(0.010)	(0.011)	(0.008)	(0.011)	(0.013)
Panel B. 2SLS - Multiple cutoffs						
Overall						
Dummy prison	0.062	0.302***	0.389***	0.236***	0.403***	0.469***
	(0.039)	(0.076)	(0.086)	(0.056)	(0.081)	(0.091)
Due to new sentence						
Dummy prison	0.020	0.074	0.173*	0.079*	0.147*	0.235**
	(0.025)	(0.058)	(0.072)	(0.037)	(0.064)	(0.077)
Due to technical violation						
Dummy prison	0.044	0.249***	0.296***	0.163***	0.289***	0.325***
	(0.030)	(0.057)	(0.064)	(0.043)	(0.061)	(0.068)
Observations	20316	20316	20316	20254	20235	20169

Notes: Robust standard errors. The outcomes are defined as the variables in the leftmost column on the table in the time frame specified in the headings of columns 2 to 7 (e.g. returned to prison within 1 year after sentence). OLS results are from regressions of each outcome on three treatment dummies which indicate the sentence type: prison, jail, and jail with probation. Regressions include a linear term on the PRV scores and the interaction between the treatment dummies and the PRV scores. In Panel B, 2SLS regressions run the same OLS regression instrumenting the dummy for prison with a dummy indicating whether the PRV score is to the right of the pooled cutoff. In Panel B, the dummy for prison is instrumented with 11 dummies indicating whether the PRV score is to the right of each of 11 cutoffs. All models include cutoff fixed effects and their interactions with the PRV scores, as well as a quadratic on age at sentence. The coefficients in the table are the point estimates an indicator variable equal to one if the offender was sentenced to prison.*** p<0.01, ** p<0.05, * p<0.1

Table 1.26: Robustness check: Heaping (Any new felony)

	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
Dummy prison	-0.068***	-0.116***	-0.091***	-0.019*	-0.017	-0.000
	(0.006)	(0.013)	(0.016)	(0.010)	(0.016)	(0.018)
Panel B. 2SLS - Multiple cutoffs						
Dummy prison	-0.193***	-0.429***	-0.441***	-0.056	-0.346**	-0.228
	(0.056)	(0.099)	(0.110)	(0.070)	(0.105)	(0.120)
Observations	14741	14741	14741	14663	14633	14546

Notes: Robust standard errors. *** p<0.01, ** p<0.05, * p<0.1

Table 1.27: Robustness check: Heaping (Severity of new felony)

	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
Medium and high severity						
Dummy prison	-0.040***	-0.055***	-0.038**	-0.006	0.008	0.019
	(0.005)	(0.011)	(0.014)	(0.008)	(0.014)	(0.016)
High severity						
Dummy prison	-0.016***	-0.025***	-0.021*	-0.003	-0.007	-0.006
	(0.003)	(0.008)	(0.010)	(0.005)	(0.009)	(0.011)
Panel B. 2SLS - Multiple cutoffs						
Medium and high severity						
Dummy prison	-0.107*	-0.115	-0.089	0.030	-0.047	0.080
	(0.044)	(0.081)	(0.092)	(0.058)	(0.086)	(0.103)
High severity						
Dummy prison	0.019	0.087	0.186**	0.126**	0.112	0.239**
	(0.031)	(0.060)	(0.071)	(0.043)	(0.064)	(0.080)
Observations	14741	14741	14741	14663	14633	14546

Notes: Robust standard errors. *** p<0.01, ** p<0.05, * p<0.1

Table 1.28: Robustness check: Heaping (Future incarceration)

	After sentence			**After release**		
	1 year	*3 years*	*5 years*	*1 year*	*3 years*	*5 years*
Panel A. OLS						
Overall						
Dummy prison	-0.040***	-0.003	0.054***	0.034**	0.071***	0.086***
	(0.006)	(0.014)	(0.016)	(0.011)	(0.016)	(0.017)
Due to new sentence						
Dummy prison	-0.023***	-0.025**	-0.010	0.005	0.010	0.017
	(0.003)	(0.010)	(0.013)	(0.007)	(0.012)	(0.014)
Due to technical violation						
Dummy prison	-0.018***	0.023*	0.067***	0.030***	0.071***	0.086***
	(0.005)	(0.011)	(0.013)	(0.009)	(0.013)	(0.015)
Panel B. 2SLS - Multiple cutoffs						
Overall						
Dummy prison	0.089	0.293***	0.355***	0.259***	0.355***	0.418***
	(0.047)	(0.086)	(0.097)	(0.065)	(0.091)	(0.103)
Due to new sentence						
Dummy prison	0.025	0.073	0.116	0.105*	0.109	0.162
	(0.030)	(0.066)	(0.081)	(0.045)	(0.072)	(0.087)
Due to technical violation						
Dummy prison	0.067	0.254***	0.295***	0.160**	0.283***	0.317***
	(0.036)	(0.064)	(0.073)	(0.049)	(0.068)	(0.076)
Observations	14708	14708	14708	14663	14647	14602

Notes: Robust standard errors. The outcomes are defined as the variables in the leftmost column on the table in the time frame specified in the headings of columns 2 to 7 (e.g. returned to prison within 1 year after sentence). OLS results are from regressions of each outcome on three treatment dummies which indicate the sentence type: prison, jail, and jail with probation. Regressions include a linear term on the PRV scores and the interaction between the treatment dummies and the PRV scores. In Panel B, 2SLS regressions run the same OLS regression instrumenting the dummy for prison with a dummy indicating whether the PRV score is to the right of the pooled cutoff. In Panel B, the dummy for prison is instrumented with 11 dummies indicating whether the PRV score is to the right of each of 11 cutoffs. All models include cutoff fixed effects and their interactions with the PRV scores, as well as a quadratic on age at sentence. The coefficients in the table are the point estimates an indicator variable equal to one if the offender was sentenced to prison.*** p<0.01, ** p<0.05, * p<0.1

CHAPTER II

Economic decision making along the college to labor market transition

2.1 Introduction

Tracking changes in economic decision-making along the life cycle is an important area of interest in economics. If individuals make decisions differently along the different phases of the life cycle, and in particular, if they engage in counterproductive decision-making in some of those phases, there may be room for policies to help them avoid costly mistakes. This motivates the literature studying how risk, time and social preferences measured through experimental games and survey measures change over time or in response to macroeconomic or idiosyncratic shocks (see Chuang & Schechter, 2015, for a survey). Even though economic models take preferences as given, there may be variation associated with age, cognitive ability and other individual characteristics (Benjamin et al., 2013), or with changes in the environment people face (Malmendier & Nagel, 2011; Nguyen, 2011; Beauchamp et al., 2012; Krupka & Stephens, 2013).

This paper studies the transition from being a student to working full-time and how decision-making, performance in cognitive tests, and self-reported feelings about financial situation and phychological measures change along this transition. Even though this is an important transition in the lives of many people, the changes it involves have not been sufficiently studied in the economics literature.[1] Economic theory predicts that students' behavior should not change when transitioning from

[1] One exception is Gustman and Stafford (1972) who study consumption changes among graduate students

being a student to a working person because they should incorporate all future income streams into their decisions before graduating and are, therefore, expected to dissave in order to smooth their life-time consumption. Empirical tests of the permanent income hypothesis (PIH), however, show that it fails when analyzing income shocks from tax refunds (Shapiro & Slemrod, 1995) and from receipt of normal income (Stephens, 2003). Even though we are not directly testing the PIH in this paper, these findings provide evidence that people do not completely smooth consumption and may also change behavior in settings like ours. In fact, the uncertainty around the specifics of a job, rather than whether they will get one or not are important enough to affect decision making, but this remains an understudied aspect of uncertainty.

A unique feature of our study is that, while in college, students cannot completely smooth their consumption in the presence of liquidity constraints, and they also face uncertainty about what kind of job they will get and their starting salary. In our sample, at baseline, only 6 percent of students have a credit card with a limit larger than US$1,000, and 10 percent have loans of US$5,000 or more. Hence, in this context, it may not be surprising that we observe changes in decision making along the transition from college to the labor market.

To measure changes in decision making, we compare experimentally-measured economic decision-making tasks across three main stages: job search (baseline), receiving and accepting a job offer, and receiving the first paycheck. We compare the second and third stages to the first to provide evidence on changes in decision making along these two important stages of the transition to the labor market. Our experimental measures include risk aversion (Eckel & Grossman, 2002; T. Tanaka, Camerer, & Nguyen, 2010), time preferences (Andreoni & Sprenger, 2012), ambiguity aversion (Y. Tanaka et al., 2014), cognitive measures (IQ test, cognitive reflection test, numerical Stroop task, Flanker task), and social preferences (dictator and ultimatum games). We also measure perceived financial situation and emotions through survey questions.

Our sample consists of students in a prestigious university in Colombia who are very likely to experience the college-to-labor-market transition, as their chances of finding a job soon after graduation are high. Specifically, we compare students in their last semester in college (i.e. those who are searching for jobs) to similar students in previous semesters (i.e. those who are involved in day-to-day college life). We pick comparison students to closely match the gender, major-choice and economic background of last-semester students to most accurately mimic what would have happened to last-semester students had they not finished college. With this comparison group, we perform a difference-in-differences (DID) analyses of risk, time and social preferences, performance in cognitive tests, perceived financial situation, and psychological factors. We also provide evidence of the stability of these outcomes across time from within-person analyses.

We find significant effects on decision making both in the after-offer and after-paycheck stages among last-semester students who transition to the labor market. What is striking is the strong series of effects observed when these students merely receive a job offer - they become less present-biased, perceive less hardship in raising funds for emergencies, become more altruistic and scale up both their spending in rent and groceries and their savings. This demonstrates how important resolving uncertainty around the details of an otherwise almost guaranteed event can be in affecting decision making.

This study also highlights how crucial self-reported emotion measurements can be. Students report feeling less tired, frustrated, depressed and worried when they receive their job offers. In fact, not controlling for these phychological factors can lead to erroneous conclusions about some of their economic decision making behavior. For instance, last-semester students appear to become less risk averse when they get a job offer, but this effect disappears after controlling for these psychological factors. The results on time preferences, on the other hand remain quite robust to controlling for these variables.

The fact that we observe these results demonstrates how the behavior of these students while in university, where they live quite frugally, is not only constrained by liquidity and credit constraints. This is evidence that uncertainty about the future is important in determining behavior. Once these students receive a paycheck, their behavior further changes. They perform worse on cognitive tasks relative to students in the comparison group. After being paid, these students might have to undertake greater responsibilities and may have more variables to consider, causing a greater cognitive load (Mullainathan & Shafir, 2013). Indeed, the share of their monthly income spent on groceries and savings goes up as early as the job offer stage lending credence to their added responsibilities. The feelings of less worry, tiredness, depression and frustration as well as the observed changes in preferences dissipate by the time they start getting paid.

We contribute to the literature studying changes in decision making along the life cycle and to the literature studying the stability of preferences by highlighting the implications of uncertainty and psychological factors in experimentally-elicited preferences. Experimental measures of risk, time and social preferences are used extensively in experimental economics and are increasingly being added to large-scale national surveys. Hence, it is important to understand how factors, internal and external to individuals, can affect their behavior in these tasks. Moreover, if measures of preferences elicited in the lab relate to behavior in real life, our findings suggest that when people resolve uncertainty about big life-cycle events, they are in a better position to make important decisions given that they would care more about the future, are more generous and are in a better psychological state. In the particular context that we study, after receiving a job offer is usually when people choose health and pension plans. A policy implication of this study is that major decisions such as the details of such plans may be best dealt with right after receiving a job offer. This is the period we identify as when people are most forward thinking, report less worry and tiredness and are most altruistic.

Our study compares favorably to other papers in the literature in terms of sample size, breadth of measures analyzed, and low attrition that is not systematically correlated with covariates or outcomes at baseline. Furthermore, the characteristics of our sample guarantee a large degree of homogeneity in terms of baseline cognition and education level. Because in this study we try to have individuals as similar as possible at baseline, this element constitutes an advantage over other studies because it is less likely that results are driven by correlation between cognitive ability and choices in the tasks (Benjamin, Brown, & Shapiro, 2013; Choi, Kariv, Müller, & Silverman, 2014).

The rest of this paper is organized as follows: Section 2.2 presents the background of our setting and the research design. Section 2.3 presents details about data collection and the experimental measures we use. Section 2.4 discusses the difference-in-differences results and section 2.5 presents the stability of preferences results. Section 2.6 concludes.

2.2 Background and research design

We study one of the most important transitions in life, i.e. from college to the labor market and how economic decision making changes along this period. Our setting involves students at a large public university in Colombia recruited primarily from the Engineering Deparment. In general, students from this university and engineering students specifically have very good prospects in the Colombian labor market given the academic quality and prestige of the university they attend.

A unique feature of this university is that admissions are solely determined by an admissions exam. About 40,000 to 60,000 applicants take the entrance exam every semester for admission to the Bogota Campus. The number of slots varies every semester but is usually around 2,000. Hence, the university admits the students who are at the very top of the admission score distribution. The admissions process guarantees that we will have students with similar cognitive ability so, our results will presumably not be affected by big differences in cognition or level of education. Indeed, this assumption bears out in the data where we observe parallel trends in the cognitive performance of last-semester students and their counterparts in lower years.

We use the fact that students in their last semester of college will experience a series of changes in their transition from college to the labor market. First, they will receive job offers that, once accepted, will resolve the uncertainty they may have regarding when and what kind of job they will secure and how much they will be paid. Second, once they start in the new job, they will receive a salary which will help them them ease their liquidity constraint. To construct an appropriate comparison group to have a benchmark to analyze the changes observed for students transitioning to the labor market, we use the fact that students in earlier semesters are very similar to last-semester students. Besides age and variables that are naturally different when one is farther along in college, we expect that students about to graduate and in other semesters are similar in most observable and unobservable characteristics. Our group of comparison students is selected to closely match the last-semester students on gender, major and economic background. In section 2.3 we provide evidence of similarities in characteristics we collected. Hence, we divide students in two groups: those who are about to graduate and will experience the transition within the next semester, and those who will remain in college for the duration of the study. We differentiate between the two groups by calling them “last-semester students” and “comparison”, respectively and find balance across the groups on all of the relevant characteristics.

A particular feature of this context is that, in general, students cannot perfectly smooth consumption by taking loans that will help them keep their standard of living constant before and after graduating from college. In our sample, only 6 percent of students have credit cards with a credit limit above $1,000 (the equivalent of about 1.5 times the expected salary in their first job after graduation) and about 10 percent have loans over $5,000 at baseline.

The different stages in the research design are summarized in Table 2.1. By observing behavior from the second and third stages relative to the first we provide evidence on how our outcomes of interest change across these important stages of the college to labor-market transition. Even though stage 2 is associated with the resolution of uncertainty, and stage 3 with easing of the liquidity constraint on economic decision-making, there may be other changes as students gradually become more independent.

The relevant timeline for our design is as follows. The two semesters in the academic year go from February to May and August to November. Graduation ceremonies take place in March and August. About half of our participants are in their last semester of college in the February to May semester and hence graduate in August. Students in their last semester of college typically work on a thesis, do an internship which may turn into a contract job after graduation or already have a job in an area related to their major. If they have already secured a job, the expectation

Table 2.1: Summary of research design

	Round 1 **Job search**	Round 2 **After job offer**	Round 3 **After pay**
Last-semester students	Send resumes, job interviews	Receive and accept offer	Cash on hand
Other students		Normal student life	
Timeline	April - May, 2016	October, 2016	December, 2016

is that their salary will increase when they graduate. Simultaneaously, they may look for other college-graduate jobs.

2.3 Data and experimental tasks

We collected data from five waves of surveys taking place at the recruitment stage and at three points in time according to Table 2.1: (i) Sign-up survey; (ii) Two surveys at baseline (during the last semester before graduation of last-semester students); (ii) One survey approximately after receiving and accepting a job offer[2]; (iii) One survey approximately after starting in the new job and receiving at least one paycheck. Participants responded to all surveys online on roughly the same dates between April, and December, 2016. All surveys except the sign-up questionnaire contained the same tasks, although in cognitive tests we varied the questions or worded them differently every time to reduce the role of memory in answering these questions. For other tasks, remembering would be harder because each task involved many choices.

2.3.1 Recruitment

The Engineering College at this university agreed to send an email inviting engineering students to participate in a research study about economic decision making.

[2] Notice that the timing of this survey does not coincide with graduation in August. Because all surveys were conducted either well before or well after graduation, we do not think that our results are driven by the excitement associated with graduation.

Students signed up in April, 2016 using an online form with questions about demographics, major, current semester in the major, GPA, tuition, socio-economic measures at the household level, whether they work, and perceived probability of finding a job between April and October, 2016 for those who plan to graduate in August. Students in undergraduate as well as master's programs were allowed to sign up.

Among students in their last semester and in other semesters, 767 signed up to participate in the study. Since we wanted similar numbers of last-semester students and students in other semesters to maximize power, we selected students in lower semesters to equal the number of students in their last semester who signed up. We did so by stratifying on gender, major, and tuition above or below the median.[3] Our number of observations at baseline is 363 of which 178 (or 49.1 percent of) students were in their last semester of college.

2.3.2 Tasks and incentives

Our online surveys contained of three types of questions: economic decision making tasks, cognitive tests, social preferences, and questionnaires about socioeconomic situation, consumption of durable goods, debt, stress, and salary expectations. In addition, we ask last-semester students about job offer and paycheck dates.

We measure economic decision making in terms of risk aversion, time preferences, ambiguity aversion, and inconsistencies in risk lotteries and time preference choices. We elicit risk aversion using the Eckel and Grossman (2002) measure (see example in Appendix 2.7.1). This method consists of presenting six different gambles varying the expected return, the standard deviation, and the implied CRRA range. Subjects are instructed to select one of the gambles to play. Each gamble has a 50 percent probability of receiving a low payoff and 50 percent probability of receiving a high payoff, except the first one in which both payoffs are the same. If this task is selected for payment at the end of the survey, the gamble they choose will actually be played. The expected payoff in gambles 1 to 5 increases linearly with risk. For gambles 5 and 6, the expected payoff is the same but the risk is bigger in gamble 6

[3] The median tuition per semester in our sample is COP 600,000 which is equivalent to around US$200.

as reflected by the higher standard deviation. Risk-averse subjects are expected to choose gambles with a lower standard deviation, while risk neutral subjects should choose the gamble with higher expected return (gamble 5) and risk-seeking subjects should choose gamble 6 (Charness, Gneezy, & Imas, 2013).

To analyze risk choices we use the risk lotteries from T. Tanaka et al. (2010). This method is intended to capture Prospect Theory parameters through a series of three lotteries which are much more complex than the Eckel and Grossman (2002) measure.[4] The lotteries consist of a given number of rows and two columns designated A and B. In each row, columns A and B contain two values each that represent payoffs and their probabilities appear at the top of the table. For each row subjects have to choose whether they prefer column A or B. They are explained that if this task is randomly chosen for payment at the end of the survey, one of the rows will be chosen at random, and the amount they will win will depend on the probability stated at the top of the column. The lotteries are designed such that, a risk neutral person will choose column A up to row 6 and column B starting in row 7 (see appendix 2.7.2). This is because the expected payoff of choosing column A is higher for rows 1 to 6 and higher for column B in rows 7 to 14. Ideally, subjects will switch columns only once but it has been found that if monotonic switching is not enforced, subjects often switch multiple times especially in populations with low education (T. Tanaka et al., 2010). Hence, most papers using this method only ask for the row in which the subject would switch. Because we want to study inconsistencies in choices, we do not enforce monotonic switching but rather ask for choices in every row. We are interested in seeing whether making mistakes (switching back and forth from Column A to Column B) changes across the three stages differentially for those who will find jobs while we control for learning or understanding the task better with the performance of the group of students in the comparison group.

Ambiguity aversion, the preference for known risks relative to unknown risks (Ellsberg, 1961; Camerer & Weber, 1992), is another measure of economic decision making that we analyze. To implement this measure we use a task based on

[4] The lotteries elicit the three Prospect Theory parameters: risk aversion, loss aversion, and non-linear probability weighting. Prospect Theory provides a different and more general characterization of risk preferences than Expected Utility Theory.

Y. Tanaka et al. (2014) in which subjects must choose between a gamble whose outcome objective probabilities are known relative to one in which they are unknown. In practice, participants are presented with a series of comparisons as in appendix 2.7.3. In each of 9 choices, they see two urns filled with 24 blue and red balls and are instructed that they will receive the monetary reward associated to the urn the choose to play if a red ball is drawn from that urn. In urn A (left-hand side), there are always 12 red and 12 blue balls completely visible to participants and the payment in case of drawing a red ball from that urn is 20,000 pesos (about US$7) in each of the 9 choices. Urn B (right-hand side) is partially covered so that it is impossible to know the mix of red and blue balls, hence urn B is the ambiguous urn. The occluder covers 1/4, 1/2 and 3/4 of the urn depending on the choice. In the first three choices, the value of urn A and B is the same (20,000 pesos) but in subsequent choices, the value of urn B increases to 30,000 (in choices 4-6) and to 40,000 pesos (in choices 7-9) if a red ball is randomly selected. To analyze ambiguity aversion we create a variable counting the number of times the students choose the ambiguous urn from a total of 9.

Time preferences is an important dimension of economic decision making that we measure in this study by adapting the elicitation task presented in (Andreoni & Sprenger, 2012). The idea of the task is that subjects are given a pre-specified monetary amount and are required to allocate it between two dates: earlier and later (see appendix 2.7.4). A difference between our implementation of the task relative to (Andreoni & Sprenger, 2012) is that the allocations are not continuous but discretized to increase in 1,000 pesos (about 33 dollar cents) intervals. Participants are instructed to allocate 50,000 pesos (about US$17) between two dates separated between each other by 4 or 9 weeks. The earliest payment they could receive is one week from the date they respond the survey because participants were responding the surveys online and that made it impossible to pay them the same day they finished it. The trade-offs they face are between weeks 1 and 5, 1 and 9, 5 and 9, and 5 and 13. For each of these trade-offs, they made 4 decisions with varying interest rates if money is allocated to the later date (1, 10, 50, and 100 percent interest rates). When they choose a value in the earlier date, the amount to be received in the later date was automatically calculated in the “later” column including interest. In total, in each round they made 16 choices allocating money to earlier and later dates. We analyze the monetary values assigned to early dates in each of the four time comparisons. The first decision we study is the number of non-monotonic decisions made by a student. Non-monotonocity in this case refers to allocating an increasing amount of money to an earlier period as the interest rate for the later period payout goes up. We also examine the "impatience" of students by examining how many times out of the 16 choices students allocate the entire 50,000 pesos to the sooner period. We also study a measure of present-biasedness by examining at each of the four interest rate levels, what the probability is of assigning a greater amount of money to week 1 vs. week 5 for a delay of 4 weeks and 8 weeks respectively. These probabilities are then weighted proportional to the interest rates in order to derive a percentage of "present-biasedness".

In terms of cognition, the bandwidth theory proposed by Mullainathan, Shafir and coauthors implies that scarcity (of time or resources) affects cognitive functioning which may compromise decision making (Shah et al., 2012; Mani et al., 2013; Mullainathan & Shafir, 2013). To measure different dimensions of cognition we use tasks such as a Raven's matrices-type IQ test, the Cognitive Reflection Test (CRT), the Flanker's test, and the numerical Stroop test.

The IQ test is a version of the Raven's test in which a pattern must be completed by the participant by choosing one of the choices given. This test provides a non-verbal measure of fluid intelligence which, as discussed in Mani et al. (2013), proxies the capacity to solve problems without prior knowledge. There were 9 questions in total and a time limit of 3 minutes to solve them. The test is difficult enough that very few people are able to correctly answer all questions in 3 minutes. Upon completion of the 3 minutes, participants were automatically directed to the next task. The same questions were given in Baseline 1 and after the job offer, and in Baseline 2 and after their first paycheck so the participants did not see the questions in at least 5 months.

The Numerical Stroop Test requires the subject to enter the number of digits displayed to them without getting distracted by the digit itself. For example if they see "3 3" they must respond "2" which is the number of objects displayed and not "3" which is the number that may come first to mind. This test has been used by Mani et al. (2013) and Carvalho et al. (2016) as a measure of cognitive control which is related to inhibitting innapropriate responses and selecting the appropriate information for processing. Because our surveys are taken online, our version of the Numerical Stroop Task involves using the keyboard to select the correct number of objects displayed out of 45 in total in 30 seconds. Participants receive 1,000 pesos for each correct answer if this task is selected for payment at the end of the survey.

In the Flanker Test subjects see a sequence of five arrows pointing to the left or to the right. They have to press the arrow in the keyboard that corresponds to the direction that the middle arrow in the sequence is poining to. This test measures the ability to ignore distracting information and supress inappropiriate responses. Again they have 30 seconds to correctly respond as many questions as possible.

The last cognitive measure we study is the Cognitive Reflection Test (CRT). This test measures the ability to supress incorrect intuitive and spontaneous answers to give the reflective correct answer (Frederick, 2005). The test usually consists of 3 questions (see appendix 2.7.5) but we add three more from Sinayev and Peters (2015) or change the wording of the original three questions so that it is harder for participants to recognize them from previous rounds.

Finally, for social preferences we use the dictator and ultimatum games. By introducing these games, we were interested in seeing whether altruistic behavior changes across the different stages. In the dictator game, participants were told that they will receive 20,000 pesos for sure. Then, they had to choose whether to give part of their allocation to another randomly chosen student participating in the study. If this task would be chosen for payment, the allocation chosen by the student would be implemented. A second question of this game changes the recipient of the gift from a randomly chosen student to a foundation that helps kids in need in Bogota. In the ultimatum game, the setup is the same except that now the subject proposes an allocation to the recipient student which can be rejected or accepted by the recipient.[5]

The order in which tasks appeared to participants was random although they always came before the questionnaire about psychological and stress measures, ex-

[5] The response from the recipient in this game was actually not implemented because participants were not responding the survey simultaneously. In practice, whatever amount the participant donated was assigned to a randomly chosen student.

penditures, salary expectations, and relevant dates of job offer and paycheck. No feedback about performance after each survey was given to participants. At the end of the survey, one of the tasks was selected for payment. The computer followed the instructions that participants read in the instructions in order to select the amount of the prize. In each survey excluding the recruitment survey, prizes ranged from the equivalent of US$7 to US$57. The mean prize across all three rounds of surveys was $30.

2.3.3 Summary statistics

As mentioned previously, at baseline, we expect that last-semester students do not differ substantially from students who will not experience the changes along the transition to the labor market except in variables such as age and degree of independence. Table 2.2 presents the means of variables collected at sign-up and the p-values of the differences between the two groups. We see that last-semester students are older, more likely to be employed at the time of the survey (during their last semester of college), to have accumulated more experience from part- and full-time jobs and internships, and less likely that their parents pay for most of their expenses. Importantly, they do not differ from comparison group students across other demographic or academic characteristics.

To construct our difference-in-differences, we collected information from students over a total three rounds. From the two surveys in the baseline period (round 1), we are able to establish parallel trends for most of our outcome variables. In the section showing the difference-in-differences results we include the two rounds of data collection to check for parellel trends. We only see one variable (inconsistencies in risk aversion) to have an interaction coefficient in the diff-in-diff regression in the baseline periods that is statistically significant. Hence, given that virtually all variables exhibit parallel trends in the period before the changes associated with the transition to the labor market take place, the difference-in-differences analysis is a valid method to analyze our data.

Round 2 and Round 3 were timed in such a way as to capture as many students as possible who met the criteria for these rounds, namely to have received a job offer

Table 2.2: Differences in baseline characteristics

Variable	**Comparison**	**Last-sem. students**	**Obs**	**p-value difference**
Poor (tuition<median)	0.54	0.59	363	0.29
Female	0.26	0.25	363	0.79
Age	22.98	25.07	352	0.00
Tuition	6.71	6.26	363	0.42
Undergraduate	0.87	0.88	363	0.73
Semester	6.20	10.42	355	0.00
GPA	3.80	3.81	356	0.70
Poor (SISBEN=1,2,3)	0.36	0.38	363	0.70
Residential stratum	2.85	2.93	362	0.34
Employed	0.43	0.65	363	0.00
Expected first salary (pesos)	2,046,757	1,957,584	363	0.33
Expected salary in 5 years (pesos)	4,508,649	4,819,663	363	0.19
No. semesters working full time	0.32	0.52	363	0.07
No. semesters working part time	1.97	2.84	363	0.00
No. semesters internship	0.09	0.53	363	0.00
How hard to find job after graduation	2.83	2.57	363	0.01
Parents in different hh	0.05	0.07	363	0.33
Parents pay most expenses	0.72	0.57	363	0.00

by Round 2, and to have started working by Round 3. However, there is no uniform way in which all students get jobs at the same time, and so while a majority of the last-semester students fulfilled the criteria, not all of them did. In order to interpret our results using a difference-in-differences strategy in which to compare their Round 2 or Round 3 behavior to baseline, we made various modifications to the observations in Round 2 and Round 3.

By Round 2, all of the last-semester students had job offers, but some of them had also started working. This increases the probability that they had already been paid, and therefore, if we wanted to quantify the effect of resolving their uncertainty with a job offer, before they had been paid, the students who had already started working may bias the results. For instance, we had 142 last-semester students in Round 2 originally and they all reported having received a job offer. Out of these, 46 students report having started working and thefore, get transferred to Round 3. Then in Round 3, we check if every student reports having received at least one paycheck, if they have not yet received one, they are moved from Round 3 to Round 2 - in this case, 12 students from Round 3 were moved to Round 2. Therefore, the students who remained in Round 2 were those who explicitly reported having a job offer and not working, or working and not having received a paycheck. The students who remained in Round 3 were those who reported at least one paycheck in the previous few months. By making this adjustment, we can now interpret the effect of Round 2 as being the effect of having a job, and therefore having one's uncertainty about the job being resolved. The effect of Round 3 would be capturing the easing of the liquidity constraint because these students would now be paid a salary, among other changes associated with starting a new job. Further details of the effect of these adjustments can be found in Table 2.3.

Table 2.3: Adjustments reflecting after offer and after paycheck stages

No. of Students	Original Observations				Observations after Adjustment			
	Round 1		Round 2	Round 3	Round 1		Round 2	Round 3
	Wave 1	Wave 2			Wave 1	Wave 2		
Total	365	363	304	285	365	363	258	273
Last semester	179	178	142	128	179	178	96	116
Job offer	-	-	142	-	-	-	96	-
Working	-	-	-	128	-	-	-	116

2.3.4 Attrition

Any longitudinal study involves some degree of attrition. In this section we assess the extent of attrition across different rounds and whether it can be predicted from baseline covariates, baseline outcomes, or the stability in the outcomes measured at two points in the baseline. If attrition happened differentially for students with certain characteristics, some of our results in the next section could be driven by selection into staying in the sample.

For the baseline, surveys 1 and 2, we collected data from 363 participants. The aggregate attrition rate after receiving a job offer is 16.7 percent and after receiving the first paycheck, it is 21.5 percent. Effective attrition, after making the adjustments in the definition of stages discussed in the previous section, is 28.9 percent for the after offer period (Round 2) and 24.8 percent for the after paycheck period (Round 3). We consider these rates to be exceptionally low among longitudinal studies. For the econometric analyses we use the sample that contains students who answered all surveys. In total, 64 percent of the students we observe at baseline responded to all surveys in the study.

As expected, attrition is higher among last-semester students who eventally graduate and find jobs (regression results not shown). Besides being correlated with this characteristic, students who stay in the sample are more likely to be undergraduates although the statistical significance of this variable dissapears when adjusting for multiple inference testing. The only other variable that is correlated with attrition is whether the student receives a salary at baseline. However, the coefficient of this variable is 0.000 so there seems to be no bias. Except for being a last-semester student, there is no indication that attrition is related to other baseline covariates.

Because comparison group students are more likely to respond to all surveys, we examine whether outcomes measured at baseline, and the stability of those outcomes, are related to staying in the sample for the comparison and last-semester students separately. Tables 2.4 and 2.5 show the attrition test results. The dependent variable in the two tables is an indicator equal to one if the student responds all surveys. We regress that indicator on each of the variables in the rows. Given the large number of regressors, we split these regressions in two tables. Table 2.4 shows the risk, time and social preferences outcomes at baseline. Similarly, table 2.5 shows cognitive tests, perceptions on personal finances and emotion measures. The six columns of results correspond to one of three samples (all, comparison group, and last-semester students) and two types of dependent variables. The first three columns look at the baseline outcomes in levels. Columns 4 to 6 look at the changes in the outcomes from baseline 1 to baseline 2, thus giving a measure of stability at baseline. For statistical significance, we report the usual tests without adjusting for multiple hypothesis testing (stars) and the tests adjusting for the Benjamini-Hochberg method within column (daggers).

Column 1 of Table 2.4 shows that students who remain in the sample are more likely to make non-monotonic switches in the risk lottery and less likely to be present-biased that students who leave the sample. Only the result for the significant correlation in inconsistencies in risk choices survives the multiple testing adjustment. Moreover, it is clear that this significant relationship is driven by last-semester students as can be seen in column 3. In Table 2.5 we do not see any significant differences between those who stay and those who attrit in the baseline outcomes measured in levels (Columns 1 to 3).

We also look at correlation between stability of the baseline outcomes and attrition in columns 4 to 6 of Tables 2.4 and 2.5. In this case, the variables in the rows are defined as the change in level from baseline 1 to baseline 2. Notice that this analysis could not have been be performed by other studies dealing with stability because they do not have two measurements for the full sample like we do.

We see in column 4 of Table 2.4 that even though stability of present-biasedness is negatively related with who stays in the sample, it does not pass the multiple hypothesis test.[6] The same is true for the change in the IQ test performance and the change in the variable enjoying myself. Column 5 shows the relationships between attrition and stability for comparison group students. We report that larger instability in the fraction of the endowment donated to a foundation in the dictator game is associated with being less likely to stay in the sample (significant after adjusting for multiple testing). Besides this variable, we do not see large correlations between stayers and stability of outcomes at baseline other than what we already saw in column 4.

There are more regressors related to who stays in the sample in the case of last-semester students (column 6). Even though they do not pass the strict multiple

[6] Each of the regressions shown in Tables 2.4 and 2.5 include 26 regressors whose point estimates are split in the two tables for ease of display. The multipe hypothesis tests are performed within regression in which case 26 hypotheses are tested simultaneously.

Table 2.4: Tests for sample attrition (Part 1)

	Levels in Baseline 2			**Change from Baseline 1 to 2**		
	All	Comparison	Last-sem. Students	All	Comparison	Last-sem. Students
Risk averse	0.095	0.049	-0.005	0.006	-0.042	-0.242**
	(0.086)	(0.094)	(0.131)	(0.075)	(0.088)	(0.110)
CRRA	-0.033	-0.021	-0.020	-0.015	-0.028	0.007
	(0.024)	(0.026)	(0.038)	(0.020)	(0.021)	(0.030)
Inconsistent risk	0.205***†††	-0.001	0.388***††	0.017	-0.069	0.092
lottery	(0.065)	(0.085)	(0.109)	(0.066)	(0.059)	(0.103)
Ambiguity averse	0.030	0.075	0.033	-0.031	0.023	-0.023
	(0.052)	(0.057)	(0.087)	(0.051)	(0.052)	(0.081)
Present biased	-0.002**	-0.002	-0.001	-0.002***	-0.002**	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Impatient	-0.003	-0.002	0.006	-0.005	-0.011	0.003
	(0.008)	(0.011)	(0.011)	(0.010)	(0.014)	(0.012)
Non-monotonic	0.022**	0.008	0.026	0.014	0.001	0.005
choices	(0.009)	(0.010)	(0.017)	(0.009)	(0.014)	(0.013)
Fraction to student	-0.027	0.049	0.131	-0.056	0.186	-0.077
- dictator	(0.134)	(0.158)	(0.181)	(0.142)	(0.201)	(0.204)
Fraction to foundation	-0.110	-0.128	-0.192	-0.097	-0.422***††	0.159
- dictator	(0.097)	(0.120)	(0.136)	(0.129)	(0.137)	(0.173)
Fraction to student	0.090	0.157	-0.046	0.144	0.359	-0.052
- ultimatum	(0.175)	(0.253)	(0.235)	(0.169)	(0.247)	(0.242)
Constant	0.651***††	0.812***	0.533	0.583***††	0.827***††	0.306***††
	(0.198)	(0.251)	(0.337)	(0.047)	(0.062)	(0.071)
N	352	178	171	317	157	157

Notes: Standard errors clustered at the individual level in parentheses
*** p<0.01, **p<0.05, * p<0.1
Multiple testing: ††† p<0.01, †† p<0.05, † p<0.1

hypothesis test, it seems that last-semester students with larger changes in risk aversion, and enjoying myself are more likely to stay in the sample. Similarly, larger changes in IQ tests performance and happiness are positively associated with staying in the sample. To have an idea of when these variables would be significant, they would pass the multiple hypothesis test if we were testing 6 hypothesis only.

Overall, we find that last-semester students are more likely to attrit. Students who remain in the sample are more likely to make inconsistent choices in the risk lottery and, to some extent, to be less present biased. Analyzing changes in the outcomes across the two baselines brings up more significant relationships between

Table 2.5: Tests for sample attrition (Part 2)

	Levels in Baseline 2			**Change from Baseline 1 to 2**		
	All	Comparison	Last-sem. Students	All	Comparison	Last-sem. Students
IQ test (Raven's)	0.012	0.001	0.026	0.035***	0.020	0.052**
	(0.018)	(0.019)	(0.027)	(0.013)	(0.015)	(0.021)
CRT test	0.008	-0.024	-0.017	-0.023	-0.005	0.008
	(0.041)	(0.045)	(0.059)	(0.031)	(0.034)	(0.053)
Stroop test	-0.001	0.002	0.004	0.004	-0.003	0.011**
	(0.005)	(0.004)	(0.009)	(0.004)	(0.004)	(0.006)
Flanker test	-0.003	0.000	-0.006	-0.004	-0.000	-0.008**
	(0.003)	(0.005)	(0.006)	(0.003)	(0.003)	(0.004)
Hard to come up with money	-0.008	-0.018	-0.117	0.022	-0.039	0.025
	(0.056)	(0.066)	(0.085)	(0.065)	(0.069)	(0.108)
Hard to cover expenses	-0.020	0.064	0.027	-0.040	0.089	-0.097
	(0.074)	(0.085)	(0.108)	(0.064)	(0.072)	(0.103)
Insatisfied HH finances	-0.026	0.002	0.001	-0.035	0.040	-0.050
	(0.066)	(0.068)	(0.097)	(0.058)	(0.059)	(0.089)
Stressed personal finances	-0.033	0.047	-0.071	0.025	0.002	-0.035
	(0.064)	(0.079)	(0.095)	(0.059)	(0.049)	(0.090)
Inconsistent in the value of money	-0.031	-0.031	-0.048	-0.007	-0.039	-0.089
	(0.055)	(0.065)	(0.096)	(0.060)	(0.066)	(0.102)
Happy	0.117*	0.087	0.131	0.094	0.042	0.229***
	(0.071)	(0.092)	(0.101)	(0.058)	(0.067)	(0.086)
Frustrated	0.028	-0.054	0.035	0.017	-0.023	0.089
	(0.070)	(0.073)	(0.133)	(0.054)	(0.057)	(0.085)
Depressed	-0.032	-0.063	0.015	-0.041	-0.110*	-0.044
	(0.074)	(0.082)	(0.144)	(0.057)	(0.058)	(0.097)
Worried	0.005	-0.039	-0.021	-0.011	0.023	-0.086
	(0.061)	(0.071)	(0.088)	(0.050)	(0.055)	(0.075)
Enjoying myself	-0.109*	-0.022	-0.215**	-0.126**	-0.013	-0.182**
	(0.066)	(0.084)	(0.089)	(0.052)	(0.068)	(0.076)
Tired	0.017	-0.001	-0.013	-0.029	0.024	-0.071
	(0.058)	(0.066)	(0.084)	(0.050)	(0.052)	(0.074)
Constant	0.651***††	0.812***††	0.533	0.583***††	0.827***††	0.306***††
	(0.198)	(0.251)	(0.337)	(0.047)	(0.062)	(0.071)
N	352	178	171	317	157	157

Notes: Standard errors clustered at the individual level in parentheses
*** p<0.01, **p<0.05, * p<0.1
Multiple testing: ††† p<0.01, †† p<0.05, † p<0.1

staying in the sample and baseline outcomes in particular for the group of last-semester students. A joint test of the hypothesis that the outcomes in changes predict who stays in the sample is significant at the 5 percent level in the case of last-semester students but not for comparison group students.

2.4 Difference-in-Differences Results

In order to examine whether the economic and social preferences, cognitive performance and survey responses of these students change along the transition from college to the labor market, we employ a difference-in-differences (DID) strategy. The two main periods of interest are Round 2, when last-semester students receive a job offer, and Round 3, when they finally start working and receive at least one paycheck. An important contribution of this paper is to separate these two periods in order to understand whether there are any changes in behavior along the transtition and whether the changes in decision making accompany a mere resolution of uncertainty after getting a job offer or whether there needs to be a real increase in their incomes. In order to tease out these various changes, we run the following regression specification:

$$y_{it} = \alpha_1 \text{Baseline 1} + \alpha_2 \text{Baseline 2} + \alpha_3 \text{After offer} + \alpha_4 \text{After paycheck} + \beta_1 \text{Baseline 1} \times \text{Last sem.} \\ + \beta_2 \text{Baseline 2} \times \text{Last sem.} + \beta_3 \text{After offer} \times \text{Last sem.} + \beta_4 \text{After paycheck} \times \text{Last sem.} \tag{2.1}$$

In the above specification, the dependent variables include risk-aversion, time preferences, social preferences, cognitive performance, personal finances and emotions. We collect four measurements of these variables so the index t goes from 1 to 4. On the right hand side, the first four independent variables represent the indicator variables for the 4 periods under study here: the two baseline rounds, after offer and after paycheck as described above. The last four terms represent the interaction terms between the last-semester dummy, in this case a dummy for whether a student is in his or her final semester, and these four periods. Note that this regression specification does not include a constant and therefore the first four coefficients (α_1, α_2, α_3, α_4) may be interpreted as the average values of the outcomes in each of the rounds for lower year students (the comparison group). The coefficients β_1, β_2, β_3 and β_4 similarly represent the differential effect of being a last-semester student. The average value of an outcome for a last semester student in the after offer stage, for instance, may then be interpreted as $(\alpha_3 + \beta_3)$. The standard errors are clustered at individual level.

Our main hypothesis was that even though these students attend one of the most pretigious universities in Colombia, they still have considerable uncertainty about when and where they will get their jobs. This uncertainty may be enough to affect their decision making behavior, in addition to when they receive their first paycheck and resolve a potential liquidity constraint.

We measure risk aversion using the Eckel and Grossman (2002) task. Extreme risk aversion is defined as an indicator variable for when the student picks the first three gambles. The ambiguity aversion variable counts the number of times the ambiguous urn is chosen out of nine possible choices between the visible and the ambiguous urn. For details on the tasks ot the definition of the variables see Section 2.3.

Table 2.6 demonstrates that when uncertainty regarding the labor market is resolved and students who previously were in their final semester receive a job offer, there is a decrease in extreme risk-aversion (first column) among all students, but significantly more so for students who receive a job offer. Therefore students who have their job uncertainty resolved appear to have a higher propensity to pick riskier gambles than lower year students by about 12.2 percentage points. It is worth pointing out that the majority of students were risk averse at baseline with about 70 percent of students choosing one of the three least risky gambles to play. There is a significant reduction in risk aversion among all students by job offer stage in which the proportion of risk-averse subjects is reduced to about 50 percent and 38 percent among comparison group and last-semester students, respectively.

However, in column 2, we run the same regression but with controls for psychological measures, specifically: self-reported measures of how tired, frustrated, worried, depressed and happy the students were, and how much enjoyment they took in life. With the controls in place, the after-offer result for last-semester students being differentially less risk averse than their counterparts in lower years vanishes. Additionally, extreme risk aversion falls by a far greater magnitude from baseline to

Table 2.6: DID results: Risk and ambiguity preferences

	(1)	(2)	(3)	(4)
	Extremely risk averse: first 3 gambles	**Extremely risk averse: first 3 gambles**	**Lower: more ambiguity averse**	**Lower: more ambiguity averse**
Baseline 1 * Comparison	0.696***	0.778***	3.793***	3.533***
	(0.0341)	(0.0674)	(0.142)	(0.310)
Baseline 2 * Comparison	0.707***	0.846***	3.740***	3.438***
	(0.0337)	(0.0674)	(0.150)	(0.307)
After offer * Comparison	0.500***	0.201***	4.226***	3.704***
	(0.0370)	(0.0528)	(0.152)	(0.323)
After paycheck * Comparison	0.424***	0.261***	4.377***	4.552***
	(0.0366)	(0.0579)	(0.154)	(0.414)
Baseline 1 * Last sem.	-0.0103	-0.0143	0.138	0.181
	(0.0488)	(0.0489)	(0.203)	(0.202)
Baseline 2 * Last sem.	-0.0548	-0.0690	0.0516	0.0236
	(0.0492)	(0.0492)	(0.226)	(0.225)
After offer * Last sem.	-0.122**†	-0.0287	-0.0634	-0.00240
	(0.0545)	(0.0552)	(0.264)	(0.269)
After paycheck * Last sem.	0.0457	0.0155	-0.147	-0.168
	(0.0593)	(0.0624)	(0.247)	(0.249)
Observations	1,355	1,355	1,232	1,232
R-squared	0.600	0.626	0.801	0.807
Emotion controls	NO	YES	NO	YES

Standard errors clustered at the individual level in parentheses.

*** p<0.01, **p<0.05, * p<0.1

Multiple Inference: † † † p_m<0.01, †† p_m<0.05, † p_m<0.1

the after offer stage and the after paycheck stage for all students, ranging from 78 percent in the first baseline to 26 percent after receiving a paycheck. This underlines the importance of sujective measures of wellbeing and emotions in economic decision making. Therefore controlling for such measures is crucial to understanding how economic decision making changes over time, and not including them gives rise to the risk of errors in measuring trends in such decisions.

The ambiguity aversion results show that at baseline students chose the ambiguous urn very few times (less than 4 times on average) independent of their last-semester or comparison group status. These preferences remain remarkably stable over time, with there being a slight trend towards less ambiguity aversion with every period. We observe no differential impact of being a last-semester student in any period. Controlling for emotional measurements does not change these results.

Because we are analyzing multiple outcomes simultaneously, we conduct a multiple inference test to study the effects of the resolution of job uncertainty, and of receiving a paycheck on the risk and ambiguity outcomes jointly (separately for each regression specification), following the Benjamini-Hochberg procedure to determine the false discovery rates. This procedure recalculates the p-values of coefficients of interest, and the new significance levels are denoted by the † symbol. We find that the risk aversion result in column 1 holds up to this multiple inference test at 10% significant level. However, in column 2, with the emotion controls, we do not find any differential effect of receiving a job offer on the last-semester students. We may conclude that these self-reported emotional measurements do affect experimentally-elicited risk decisions. We also examine inconsistencies in risk choice, calculated by counting the number of times subjects switch between option A and B (see description in Section 2.3). However, we do not consider these results because we do not observe parallel trends in the baseline measures and attritions seems to be correlated with this variable, and therefore cannot make conclusions about subsequent results (Table 2.16 in the Appendix).

In terms of time preferences, we look at three main measures calculated from a task following a similar format to Andreoni and Sprenger (2012). In this task, we ask subjects to allocate money (50,000 pesos or $17) between two periods. These include between week 1 and 5, week 1 and 9, week 5 and 9 and week 5 and 13. If they allocate money to the later date, they receive interest of 1%, 10%, 50% and 100% for each of the above four intertemporal decisions, totalling 16 intertemporal choices to be made in each round.

From Table 2.7, out of a possible 12 non-monotonic decisions, on average, at the first baseline students just make 2.26 of these inconsistent decisions and there is no differential effect of being a last-semester student. In the second baseline round,

Table 2.7: DID results: Time preferences

	(1)	(2)	(3)	(4)	(5)	(6)
	No. of non-monotonic choices	**No. of non-monotonic choices**	**No. of times subj. allocates full amt. for sooner period**	**No. of times subj. allocates full amt. for sooner period**	**% present biasedness (weighted by interest rate)**	**% present biasedness (weighted by interest rate)**
Baseline 1 * Comparison	2.258***	2.577***	2.577***	2.188***	30.07***	36.91***
	(0.239)	(0.528)	(0.247)	(0.522)	(2.889)	(6.150)
Baseline 2 * Comparison	2.088***	2.069***	3.434***	3.617***	28.81***	32.38***
	(0.241)	(0.457)	(0.281)	(0.596)	(2.688)	(5.558)
After offer * Comparison	1.515***	1.659***	4.571***	4.748***	34.18***	33.07***
	(0.225)	(0.482)	(0.315)	(0.698)	(3.074)	(6.611)
After paycheck * Comparison	1.344***	1.216***	4.487***	6.284***	29.10***	41.48***
	(0.224)	(0.405)	(0.330)	(0.732)	(3.061)	(7.253)
Baseline 1 * Last sem.	0.0468	0.0320	0.254	0.272	-1.274	-1.820
	(0.359)	(0.368)	(0.356)	(0.359)	(4.064)	(4.110)
Baseline 2 * Last sem.	-0.622*	-0.678**	0.623	0.516	0.502	-0.0710
	(0.317)	(0.319)	(0.410)	(0.413)	(3.805)	(3.779)
After offer * Last sem.	-0.452	-0.518	-0.549	-0.560	-10.07**†	-12.38**†
	(0.328)	(0.363)	(0.509)	(0.514)	(4.610)	(4.787)
After paycheck * Last sem.	-0.414	-0.478	0.320	0.249	5.229	3.994
	(0.302)	(0.325)	(0.495)	(0.493)	(4.806)	(4.774)
Observations	1,243	1,243	1,243	1,243	1,243	1,243
R-squared	0.264	0.272	0.508	0.520	0.480	0.492
Emotion controls	NO	YES	NO	YES	NO	YES

Standard errors clustered at the individual level in parentheses.

*** p<0.01, **p<0.05, * p<0.1

Multiple Inference: ††† p_m<0.01, †† p_m<0.05, † p_m<0.1

however, there appears to be a differential effect on making non-monotonic choices for last-semester students, violating the parallel trends assumption (at the 10 percent level) in this case.

Our measure of "impatience", which counts the number of times a student allocates the full endowment to the earlier period does not demonstrate any additional effect for last-semester students, On average, all students appear to become slightly more impatient over time, with them making around 2.6 "impatient" choices in the first baseline and going up to almost 5 out of 16 by Round 3 (when last-semester students receive their first paycheck). On controlling for emotional measurement, this range widens, but once again, there is no additional effect of being a last-semester student. Finally, we also examine a measure of present-biasedness, where we enumerate the instances where a subject allocated a greater amount to the sooner period when the sooner period was a week from now versus 5 weeks from now for the same delay length until the later period, i.e., week 1 vs. week 5 and week 5 vs. week 9. This is then weighted by the interest rate for the later period payoff in each row of the price list to end up with a percentage of present bias. Here, we find that last-semester students are differentially less present biased that their counterparts after receiving a job offer in Round 2, by about 10 percentage points. In fact, this result becomes even stronger when controlling for emotional measurements, with the gap widening to 12 percentage points. This demonstrates that time preferences may be less swayed by these emotions compared to risk decisions. A multiple inference test following the Benjamini-Hochberg procedure combining all three time preference measures reiterates that for present-biasedness, there is a definite effect from resolving job uncertainty for last-semester students. To contextualize this result, Carvalho et al. (2016) find that before payday, poor individuals in the US are more present biased when making choices about monetary rewards. We observe the opposite behavior when individuals have not received income yet but face less uncertainty regarding their future income and outcome of their college education investment.

We do find significant effects of receiving a job offer on the self-reported financial health of last-semester students. In Table 2.8 we show the result of regressions on outcomes such as whether it is hard to come up with money for an emergency, whether it is hard to cover next week's expenses with the money they have today and whether they are stressed about personal finances. In the first two cases, receiving a job offer has a significant and positive effect for last-semester students. To elaborate, they report finding it hard to come up with money or it being hard to cover expenses less frequently than the baseline, and when compared to students in lower years. Therefore, in terms of perception of own wealth, there appears to be a clearly positive effect of merely receiving a job offer, without having yet been paid. There is no significant differential effect for last-semester students after receivng a paycheck, which is telling of the immense effect that the resolution of uncertainty alone has on perception of one's coping ability.

Table 2.8: DID results: Financial status

	(1)	(2)	(3)	(4)	(5)	(6)
	Hard to come up with money	**Hard to come up with money**	**Hard to cover expenses**	**Hard to cover expenses**	**Stress level - personal finances**	**Stress level - personal finances**
Baseline 1 * Comparison	0.592***	0.602***	0.234***	0.184***	0.391***	0.371***
	(0.0364)	(0.0771)	(0.0313)	(0.0632)	(0.0361)	(0.0753)
Baseline 2 * Comparison	0.663***	0.568***	0.196***	0.257***	0.326***	0.236***
	(0.0350)	(0.0745)	(0.0294)	(0.0616)	(0.0347)	(0.0676)
After offer * Comparison	0.571***	0.253***	0.245***	0.0984**	0.337***	0.107**
	(0.0366)	(0.0464)	(0.0318)	(0.0385)	(0.0350)	(0.0444)
After paycheck * Comparison	0.543***	0.312***	0.212***	0.209***	0.337***	0.159***
	(0.0369)	(0.0566)	(0.0302)	(0.0484)	(0.0350)	(0.0480)
Baseline 1 * Last sem.	-0.0531	-0.0469	-0.0539	-0.0476	0.0357	0.0429
	(0.0522)	(0.0516)	(0.0426)	(0.0412)	(0.0519)	(0.0505)
Baseline 2 * Last sem.	-0.124**	-0.0993*	0.00660	0.0164	0.0672	0.0905*
	(0.0513)	(0.0516)	(0.0421)	(0.0414)	(0.0505)	(0.0495)
After offer * Last sem.	-0.253***†††	-0.139***††	-0.103**††	-0.0355	-0.0599	0.0399
	(0.0531)	(0.0511)	(0.0429)	(0.0444)	(0.0509)	(0.0492)
After paycheck * Last sem.	-0.0913	-0.155***††	-0.0120	-0.0263	0.0804	0.0167
	(0.0594)	(0.0589)	(0.0481)	(0.0503)	(0.0579)	(0.0564)
Observations	1,355	1,355	1,355	1,355	1,355	1,355
R-squared	0.553	0.595	0.207	0.275	0.368	0.438
Emotion controls	NO	YES	NO	YES	NO	YES

Standard errors clustered at the individual level in parentheses.

*** p<0.01, **p<0.05, * p<0.1

Multiple Inference: † † † $p_m<0.01$, †† $p_m<0.05$, † $p_m<0.1$

The results for the self-reported measure on it being hard to come up with money in an emergency holds up to the regression specification with the emotional controls - i.e. last-semester students less frequently report this in Round 2 after receiving a job offer. Indeed, after controlling for these emotions, there is an additional effect for last-semester students after receiving their paycheck as well. What is important to note with this specification however, is that all students, on average, report it being less hard to come up with money in an emergency over the 4 rounds, but we do observe differential effects for last-semester students. This, once again, underlines that controlling for such emotions when studying perceptions of financial status are important because they can affect not only economic decision making measures but subjective assessment of financial health.

Once again, on running a multiple inference test for these three financial health measures, the differential effect for last-semester students after their job offer persists at 1 and 5 percent significance levels in their perception of it being hard to come up with money in an emergency and it being hard to cover expenses (only for the regression specification without emotional measurement controls) respectively. The significant and different effect at after paycheck stage from column 2 also persists after the multiple inference test.

One reason that last-semester students and comparison students do not differ in their reporting of how hard it is to cover expenses, despite last-semester students perceiving it being less hard to come up with money in an emergency, may be due to their increased expenditures after receiving a job offer. Table 2.9 backs up the above pattern by demonstrating that the fraction of these job market students' monthly income spent on rent, groceries and savings goes up significantly in Round 2 when last semester students receive a job offer. Particularly their expenditures on groceries and savings remain differentially higher than lower year students in both regression specifications (with and without emotion controls.) In fact, after calculating the false discovery rates for the multiple inference test of how Round 2 affects these three expenditures, the results for groceries and savings remain highly significant. What is interesting here is that before having been paid in their jobs, these students already scaled up their expenditures in anticipation of receiving a paycheck.

There is also a differential but smaller effect of receiving a paycheck on last-semester students with their shares of expenditures on rent, groceries and savings being higher in Round 3 after paycheck - but this is more expected given their rise in income.

The controls that we include in our second regression specification quite often strongly affect our results regarding last-semester students, particularly in the case of risk aversion. There are some important patterns within these emotion measures as well. Last-semester students who receive a job offer report being differentially less tired, worried, depressed and frustrated (Figure 2.1). Further results are presented

Table 2.9: DID results: Spending behavior

	(1) **Share of monthly inc. spent on rent**	(2) **Share of monthly inc. spent on rent**	(3) **Share of monthly inc. spent on groceries**	(4) **Share of monthly inc. spent on groceries**	(5) **Share of monthly inc. spent on savings**	(6) **Share of monthly inc. spent on savings**
Baseline 2 * Comparison	9.388***	13.09***	6.546***	8.487***	12.84***	10.13***
	(1.334)	(2.968)	(0.733)	(1.713)	(1.228)	(2.403)
After offer * Comparison	11***	18.70***	7.856***	8.916***	11.34***	8.049***
	(1.568)	(3.848)	(0.879)	(1.712)	(1.143)	(2.826)
After paycheck * Comparison	10.46***	9.254***	8.168***	10.18***	12.14***	14.21***
	(1.545)	(3.292)	(0.835)	(1.981)	(1.147)	(2.958)
Baseline 2 * Last sem.	3.324	2.602	2.928**	2.687**	1.322	1.469
	(2.049)	(2.023)	(1.133)	(1.152)	(1.884)	(1.879)
After offer * Last sem.	4.234*†	3.083	3.059**††	2.715**†	6.014***††	5.728**††
	(2.495)	(2.518)	(1.378)	(1.377)	(2.195)	(2.256)
After paycheck * Last sem.	4.594*†	4.474*	2.650**†	2.606**	3.030	3.552*†
	(2.414)	(2.497)	(1.230)	(1.259)	(1.926)	(1.880)
Observations	884	884	884	884	884	884
R-squared	0.280	0.296	0.412	0.422	0.405	0.431
Emotion controls	NO	YES	NO	YES	NO	YES

Standard errors clustered at the individual level in parentheses.

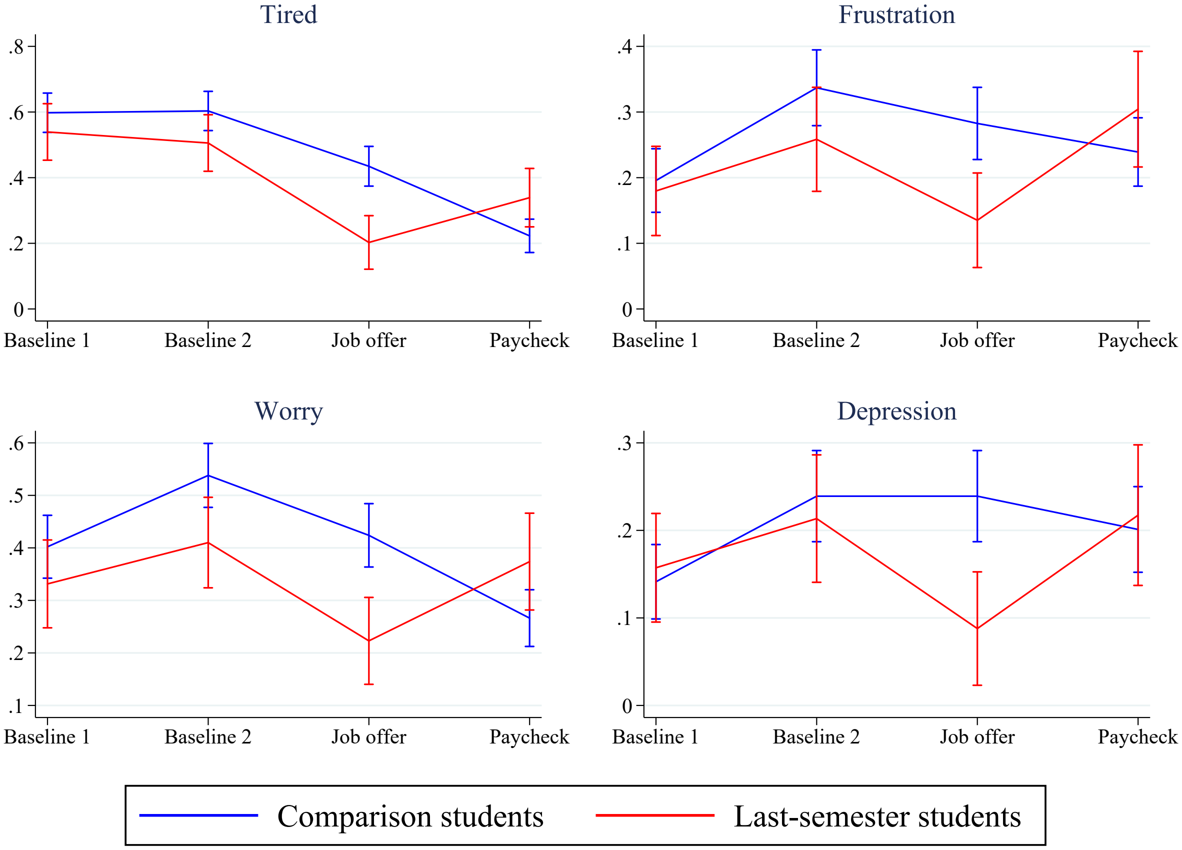
*** p<0.01, **p<0.05, * p<0.1

Multiple Inference: † † † p_m<0.01, †† p_m<0.05, † p_m<0.1

in Appendix table 2.18. What is clear is that while last-semester students report being less worried or tired in Round 2 after resultion of job uncertainty, by Round 3 after receiving a paycheck, these effects disappear, that is there is not difference in psychological measures between last-semester and comparison. Part of this may be related to the additional responsibilities they have to take care of, given their additional expenditures after resolving job uncertainty in Round 2 and after being paid in Round 3 (Table 2.9).

Corroborating the above pattern of reports of being less tired, depressed and worred dissipating by Round 3, we have further evidence that in Round 3, after receiving their paycheck there are increased responsibilities. We study changes in cognitive performance by looking at how students perform in tasks such as the Raven's

Figure 2.1: Patterns in self-reported feelings



Matrices and a Cognitive Reflection Test - CRT (Table 2.10); Flanker test and a Stroop test (Table 2.19 in the Appendix). These increased responsibilities may be contributing to an increasing cognitive load and we find that in after receiving a paycheck, last-semester students perform differentially worse than lower year students in the cognitive reflection tasks as well the Raven's Matrices. In fact, these results hold up even in the multiple inference test where we examine the hypothesis that receiving a paycheck significantly affects the four cognitive tasks jointly.

Specifically, the performance of the lower year students stays approximately stable across time in the cognitive reflections tasks. Their performance in the Raven's Matrices task does improve over time and this may be attributable to learning effects. They respond 3.5 questions (column 4 in Table 2.10) correct on average at baseline, and lower year students improve their score by almost 3, while last-semester students lag slightly behind at a score of 6 in after receiving a paycheck. This is consistent with the additional responsibilities and changes associated with starting a new job. It is possible that these changes impose a cognitive load on last-semester students and impairs their performance in these cognitive tasks.

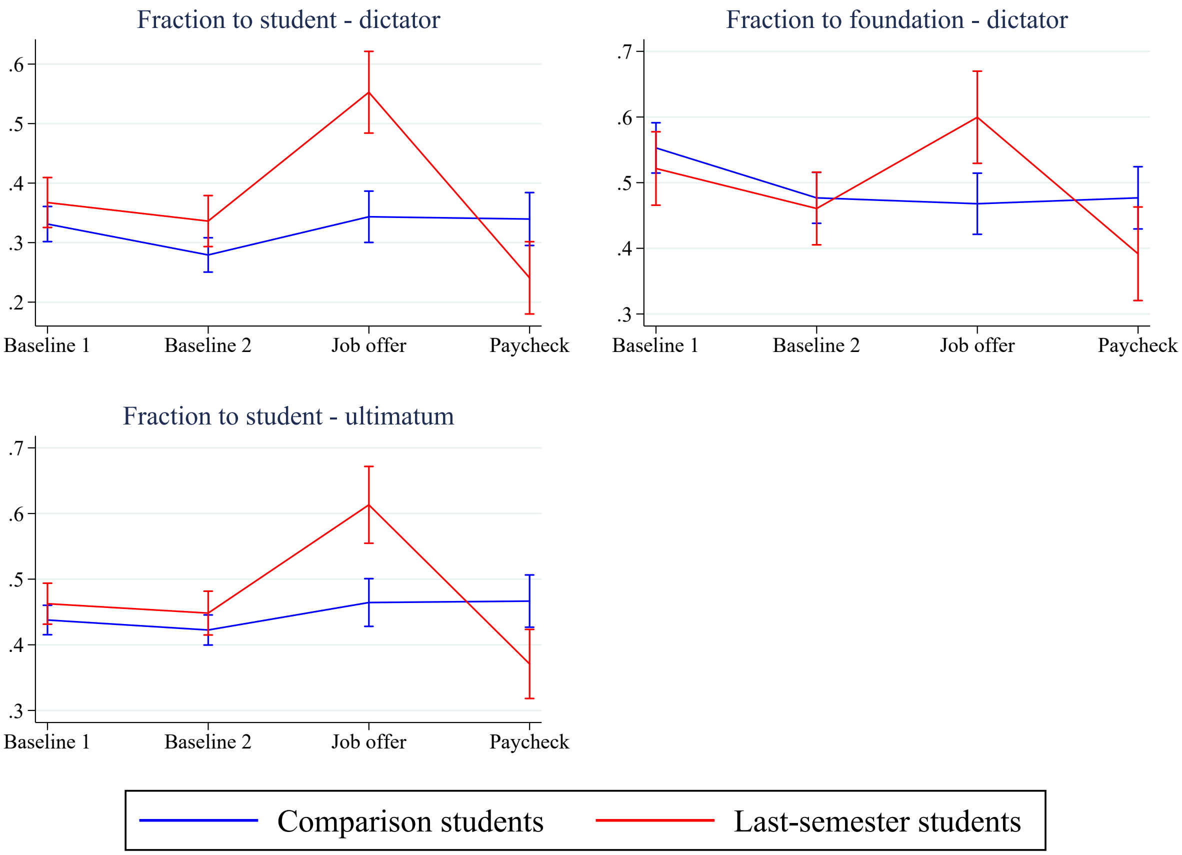
Table 2.10: DID results: Cognitive performance

	(1) **CRT: both questions**	(2) **CRT: both questions**	(3) **Raven's Matrices**	(4) **Raven's Matrices**
Baseline 1 * Comparison	1.142***	1.130***	4.038***	3.503***
	(0.0612)	(0.135)	(0.118)	(0.254)
Baseline 2 * Comparison	1.317***	1.256***	6.408***	6.050***
	(0.0548)	(0.115)	(0.110)	(0.242)
After offer * Comparison	1.188***	1.273***	5.154***	5.020***
	(0.0577)	(0.134)	(0.137)	(0.340)
After paycheck * Comparison	1.445***	1.324***	6.538***	6.447***
	(0.0595)	(0.134)	(0.116)	(0.265)
Baseline 1 * Last sem.	-0.119	-0.106	0.0919	0.117
	(0.0856)	(0.0863)	(0.171)	(0.174)
Baseline 2 * Last sem.	0.0107	0.0139	-0.00988	0.00265
	(0.0791)	(0.0805)	(0.154)	(0.156)
After offer * Last sem.	-0.0811	-0.0900	0.140	0.145
	(0.0991)	(0.101)	(0.229)	(0.237)
After paycheck * Last sem.	-0.210**††	-0.218**††	-0.512***††	-0.515***††
	(0.0937)	(0.0947)	(0.190)	(0.188)
Observations	1,243	1,243	1,249	1,249
R-squared	0.722	0.725	0.925	0.926
Emotion controls	NO	YES	NO	YES

Standard errors clustered at the individual level in parentheses.
*** p<0.01, **p<0.05, * p<0.1
Multiple Inference: † † † p_m<0.01, †† p_m<0.05, † p_m<0.1

Another hypothesis we started out with was that at Round 2 and Round 3, when job uncertainty is resolved and students start working at their first jobs, in dictator and ultimatum games, they may allocate more to others. From figure 2.2, we find that their generosity towards other individuals and charity foundation after job uncertainty is resolved is indeed higher compared to lower year students. Furthermore, we find that this differential pattern persists even after controlling for emotions in the case of allocating money to another student in a dictator game and allocating

Figure 2.2: Social Preferences (without controlling for emotion measurements)



money to another student in an ultimatum game (columns 2 and 6 in table 2.17 in the Appendix). On average, students allocate a far higher share to a foundation than a student in a dictator game, donating almost 61 percent of their endowment to a foundation versus 30 percent to a student in a dictator game (specification with emotion controls). What is puzzling under this specification is that both lower year and last-semester students scale up their donations to other parties in all three games. This is at odds with previous findings by Matthey and Regner (2013) that individuals who have participated in more experiments donate less money. One can hypothesize why last-semester students may donate more after receiving a job offer, but it is unclear why comparison students would also scale up at Round 2 (after last semester students receive a job offer) (columns 2, 4 and 6 in table 2.17 in the Appendix).

These results help us conclude that this life transition from college to working life is quite crucial in the way it affects the choices and decisions made by these subjects. An important takeaway from this analysis is the role played by emotions such as worry and tiredness in making economic decisions such as risk choices. What is interesting is that the resolution of job uncertainty alone is sufficient enough to affect time and social preferences. Their spending behavior also changes significantly in response to an expected wealth increase. Given that they are at a top university, and there is a reasonable guarantee of getting a good job, these large effects are important and provide an avenue for further research.

2.5 Stability of preferences and cognitive measures

Time and risk preferences are thought to be persistent over time for the same individual and incorporated as parameters in economic models. The literature studying stability spans several disciplines and the results tend to point towards stability of risk, time and social preferences as measured by the correlation coefficient of the preference measure in two different time periods. There is less evidence showing that preferences are affected by individual shocks such as changes in health or income. A recent survey of this literature can be found in Chuang and Schechter (2015).

Tests for stability of preferences are conducted by measuring the same underlying preference with the same task or question at different points in time or using different elicitation methods that attempt to capture the same preference at one point in time. Hence, preferences are not observed directly and may be a combination of the true underlying preference, the particular environment faced by the respondent, including macroeconomic or idiosyncratic shocks, and measurement error derived mainly from lack of understanding of the task. Given the longitudinal nature of our data, we focus on testing stability using the same experimental tasks across the relevant stages for our study. In particular, we are interested on whether stability changes along the transition from college to the labor market.

Because we collect a rich set of cognitive measures and survey questions regarding self-reported financial situation, we also analyze whether these measures change over time. We expect preferences, cognitive measures and survey questions to be relatively stable when comparing the two baseline periods. One factor that may contribute to stability is that the two baseline surveys were conducted only two weeks apart. On the other hand, better understanding of the tasks in the second baseline can contribute to lower stability if responses change relative to the first baseline due to learning effects. We expect that the later measurements, taken to coincide with the after offer and after paycheck periods for students who transition to the labor market, exhibit less stability when compared to the baseline if more learning takes place or if the responses students give are related to the changes they are experiencing.

In what follows, we report correlation coefficients that measure how persistent an outcome measure from a previous period is in a future period, as is standard in the stability of preferences literature. The standard test of stability consists of finding whether the correlation between the same outcome measured in two different periods is statistically different from zero. A test of whether the correlation is not statistically different from one is also possible but achieving perfect correlation is unlikely because there is usually measurement error in how individuals make choices in experimental tasks.[7] Given that the majority of the evidence points to stability of preferences and to low response of preference measures to individual shocks, we expect to find that correlations are significantly different from zero in line with the findings in the literature but that the farther apart the periods compared, the lower the correlation will be.

In the tables below, each column specifies which two periods are being compared in bold. For each comparison, two correlations are reported: The correlation coefficient is shown on the left-hand side, and the point estimate from a regression of the first variable in the column title on the second, on the right-hand side. All regressions control for whether the student is in the comparison group, their gen-

[7] Meier and Sprenger (2015) ask the question of what is a high enough correlation to conclude that time preferences are stable. They perform a simulation analysis including aggregate estimates of the key parameters of their model (present bias, discount rate and stochastic decision error) to obtain this persistence measure. They conclude that their simulated correlation of 0.452 is close to the observed correlation of 0.464 which suggests that stability of time preferences does not imply that the correlation has to be near 1. In fact, most studies in this literature find correlations in risk and time preferences that are below 0.5 (see literature review in Cuang and Schechter (2015).

der, poverty status as measured by the tuition they paid at baseline, and age. For the correlation coefficient, the standard errors are obtained by bootstrapping, and an adjustment for multiple testing is performed for both, correlation coefficient and regression estimates (see table notes).

Table 2.11 shows correlations between individual responses in risk and ambiguity aversion variables across pairs of periods. The variable indicating whether an individual is risk averse or not exhibits a correlation of about 0.4 in all periods in which the baseline is involved. The correlation increases to about 0.5 when comparing after paycheck and after offer periods and is lowest when comparing after paycheck with baseline. These figures are within the range of risk preference correlations reported by Chuang and Schechter (2015) of 0.13 to 0.55 in studies of more than 100 participants.

To our knowledge, previous studies have not analyzed persistence in inconsistent lottery choices. This is because they must enforce monotonic switching in order to obtain the CRRA or prospect theory parameters. In our case, however, we use a multiple price list to deliberately allow for mistakes in risk choices. In row 2 of Table 2.11 we analyze how persistent mistakes are. We see that making mistakes in the first baseline survey is highly correlated with making mistakes in the second baseline survey as the corretions between 0.4 and 0.46 show. This level of correlation persists when comparing the after offer and baseline periods but goes down to between 0.21 and 0.3 when comparing after paycheck and baseline, the two most distant periods compared.

Finally, the most drastic changes are observed in the ambiguity aversion measure. From a correlation of around 0.4 when comparing the two baseline periods, the correlation goes down to around 0.2 in the after offer vs. baseline and after paycheck vs. baseline. Interestingly, this measure did not show substantial changes for either of the groups (last-semester or comparison students) in the analysis of the previous section.

Table 2.12 shows correlation coefficients for the time preferences variables. Overall, the correlation for present biasedness is smaller than for the impatient and non-monotonic choices variables. This result suggests that allocating more to week 1 in the different trade-offs involving week 1 is less persistent over time than allocating

Table 2.11: Stability of risk preferences and ambiguity aversion

	(1)		**(2)**		**(3)**		**(4)**	
	Baseline 2		**After offer**		**After paycheck**		**After paycheck**	
	vs. Baseline 1		**vs. Baseline 2**		**vs. Baseline 2**		**vs. After offer**	
	Corr.	Reg.	Corr.	Reg.	Corr.	Reg.	Corr.	Reg.
Risk averse	0.387	0.394	0.391	0.412	0.226	0.254	0.499	0.505
	(0.065)	(0.069)	(0.061)	(0.066)	(0.063)	(0.070)	(0.056)	(0.058)
Inconsistent risk lottery	0.459	0.399	0.403	0.370	0.3	0.214	0.49	0.419
	(0.066)	(0.066)	(0.078)	(0.078)	(0.081)	(0.069)	(0.085)	(0.086)
Ambiguity averse	0.394	0.408	0.165	0.178	0.218	0.232	0.308	0.303
	(0.06)	(0.061)	(0.064)	(0.064)	(0.063)	(0.062)	(0.064)	(0.065)

Notes: Correlation coefficients are obtained from the Stata command `correlate`. Standard errors are in parenthesis are bootstrapped with 10,000 replications in the case of correlation and Eicker-Huber-White in the case of regression. The coefficients in the column Reg. are obtained from regressions of the variable on the left in the period mentioned first in the column title on the same variable in the period mentioned second in the column title. For example, in the Baseline 2 vs. Baseline 1 column, the coefficient displayed is from a regression of risk aversion at Baseline 2 on risk aversion at Baseline 1. All regressions control for whether the student is in the comparison group, gender, poverty status and age. All point estimates are significant at the 5% level after adjusting by the Benjamini-Hochberg multiple testing method. The sample size includes students observed in all rounds of data collection (234).

the full endowment to the early period as reflected in the correlations near 0.7 in the case of the impatience variable. Similar to the risk preferences analysis, the lowest correlations are for the comparison of after paycheck and baseline. In the case of present biasedness, the correlations go down to 0.13 - 0.15 and are no longer significant at the 5 percent level after correcting for multiple hypothesis testing.

Non-monotonic choices are essentially mistakes that violate the law of demand by allocating less to the later period when the interest rate is higher (see Giné, Goldberg, Silverman, & Yang, 2017). This behavior is relatively persistent when comparing the two baseline surveys and the two later periods with correlations of around 0.6. The correlations go down slightly when comparing the after offer and after paycheck periods with the baseline (columns 2 and 3 of Table 2.12).

We next look at stability social preferences in Table 2.13 in terms of the fraction of the endowment of 20,000 pesos given to a student or a foundation in the dictator and ultimatum games. The comparison of the two baselines gives the highest correlations for all three variables which are near 0.6, when allocating money to a randomly selected student. The correlations are higher in the case of donations to

Table 2.12: Stability of time preferences

	(1) **Baseline 2 vs. Baseline 1**		(2) **After offer vs. Baseline 2**		(3) **After paycheck vs. Baseline 2**		(4) **After paycheck vs. After offer**	
	Corr.	Reg.	Corr.	Reg.	Corr.	Reg.	Corr.	Reg.
Present biased	0.289	0.257	0.278	0.294	0.133	0.151	0.374	0.383
	(0.071)	(0.067)	(0.066)	(0.073)	(0.07)	(0.079)	(0.065)	(0.074)
Impatient	0.708	0.802	0.651	0.698	0.61	0.660	0.709	0.730
	(0.045)	(0.061)	(0.047)	(0.055)	(0.051)	(0.059)	(0.043)	(0.050)
Non-monotonic	0.616	0.594	0.555	0.454	0.489	0.388	0.633	0.615
choices	(0.067)	(0.063)	(0.063)	(0.073)	(0.074)	(0.072)	(0.065)	(0.078)

Notes: Correlation coefficients are obtained from the Stata command `correlate`. Standard errors are in parenthesis are bootstrapped with 10,000 replications in the case of correlation and Eicker-Huber-White in the case of regression. The coefficients in the column Reg. are obtained from regressions of the variable on the left in the period mentioned first in the column title on the same variable in the period mentioned second in the column title. For example, in the Baseline 2 vs. Baseline 1 column, the coefficient displayed is from a regression of risk aversion at Baseline 2 on risk aversion at Baseline 1. All regressions control for whether the student is in the comparison group, gender, poverty status and age. After adjusting by the Benjamini-Hochberg multiple testing method, the correlation on column 3 of the present-biased outcome is not significant at the 5% level. The sample size includes students observed in all rounds of data collection (234).

a foundation helping children in Bogota (around 0.7). The correlation coefficients go down substantially in the after offer and after paycheck comparisons relative to baseline by at least 0.2. This reduction is probably related to the fact that students, especially last-semester students, give more in subsequent rounds of the study. The correlations go back to near their baseline levels in the after paycheck vs. after offer comparison reflecting the fact that the difference between what they give in the last two rounds is not as big compared to the initial rounds.

The correlations between social preferences variables we find in this study seem stronger than what has been found in the previous literature. For example, Chuang and Schechter (2015) find that a few of the correlations in experimental games measuring social preferences are not different from zero even though there is a great deal of persistence in social preferences measured with survey questions. The literature on the stability of social preferences is certainly smaller than in other type of preferences so we are providing new evidence with sample sizes larger than the typical study (see Chuang & Schechter, 2015, for a review).

Table 2.13: Stability of social preferences

	(1) Baseline 2 vs. Baseline 1		(2) After offer vs. Baseline 2		(3) After paycheck vs. Baseline 2		(4) After paycheck vs. After offer	
	Corr.	Reg.	Corr.	Reg.	Corr.	Reg.	Corr.	Reg.
Fraction to student	0.627	0.602	0.348	0.365	0.313	0.321	0.514	0.484
- dictator	(0.049)	(0.055)	(0.072)	(0.088)	(0.074)	(0.081)	(0.067)	(0.071)
Fraction to foundation	0.729	0.702	0.539	0.580	0.511	0.577	0.717	0.746
- dictator	(0.037)	(0.049)	(0.059)	(0.073)	(0.064)	(0.076)	(0.048)	(0.053)
Fraction to student	0.585	0.616	0.314	0.391	0.3	0.418	0.543	0.582
- ultimatum	(0.068)	(0.080)	(0.092)	(0.120)	(0.084)	(0.114)	(0.081)	(0.092)

Notes: Correlation coefficients are obtained from the Stata command `correlate`. Standard errors are in parenthesis are bootstrapped with 10,000 replications in the case of correlation and Eicker-Huber-White in the case of regression. The coefficients in the column Reg. are obtained from regressions of the variable on the left in the period mentioned first in the column title on the same variable in the period mentioned second in the column title. For example, in the Baseline 2 vs. Baseline 1 column, the coefficient displayed is from a regression of risk aversion at Baseline 2 on risk aversion at Baseline 1. All regressions control for whether the student is in the comparison group, gender, poverty status and age. All point estimates are significant at the 5% level after adjusting by the Benjamini- Hochberg multiple testing method. The sample size includes students observed in all rounds of data collection (234).

Most of the studies in the psychology literature involving cognitive measures analyze how performance in these measures changes when inducing cognitive load. In economics, this often takes the form of comparing results before and after receiving a harvest payout or a paycheck (e.g., Mani et al., 2013; Carvalho et al., 2016). In this section we examine how persistent over time the measures of cognition are. Importantly, for a fraction of the participants, the periods analyzed coincide with important events in their transition from college to the labor market.

Table 2.14 shows correlation coefficients for the four cognitive measures. All are statistically different from zero at the 5 percent level after adjusting for multiple testing. What is striking in these results is that the correlations are not very high. In most cases, they are between 0.2 and 0.3 with the exception of the Cognitive Reflection Test in row 2 which reaches 0.5 in one of the pairwise comparisons. The lowest correlations are in the numerical Stroop test. This may partly be because participants faced some difficulties with this test because they were supposed to use the numerical keyboard but had issues with this in the baseline surveys.

Finally, Table 2.15 shows correlation coefficients for three of the survey measures. The individual responses to the question "How hard will it be for you to come up

Table 2.14: Stability of cognitive measures

	(1)		**(2)**		**(3)**		**(4)**	
	Baseline 2		**After offer**		**After paycheck**		**After paycheck**	
	vs. Baseline 1		**vs. Baseline 2**		**vs. Baseline 2**		**vs. After offer**	
	Corr.	Reg.	Corr.	Reg.	Corr.	Reg.	Corr.	Reg.
IQ test (Raven's)	0.269	0.220	0.346	0.388	0.388	0.339	0.32	0.222
	(0.052)	(0.049)	(0.053)	(0.073)	(0.06)	(0.054)	(0.063)	(0.050)
CRT test	0.408	0.360	0.386	0.397	0.538	0.512	0.37	0.337
	(0.056)	(0.055)	(0.056)	(0.063)	(0.048)	(0.055)	(0.056)	(0.060)
Stroop test	0.211	0.197	0.182	0.203	0.211	0.207	0.364	0.255
	(0.08)	(0.085)	(0.066)	(0.072)	(0.066)	(0.063)	(0.079)	(0.062)
Flanker test	0.363	0.304	0.284	0.336	0.292	0.308	0.246	0.209
	(0.064)	(0.058)	(0.072)	(0.093)	(0.072)	(0.081)	(0.071)	(0.072)

Notes: Correlation coefficients are obtained from the Stata command `correlate`. Standard errors are in parenthesis are bootstrapped with 10,000 replications in the case of correlation and Eicker-Huber-White in the case of regression. The coefficients in the column Reg. are obtained from regressions of the variable on the left in the period mentioned first in the column title on the same variable in the period mentioned second in the column title. For example, in the Baseline 2 vs. Baseline 1 column, the coefficient displayed is from a regression of risk aversion at Baseline 2 on risk aversion at Baseline 1. All regressions control for whether the student is in the comparison group, gender, poverty status and age. All point estimates are significant at the 5% level after adjusting by the Benjamini- Hochberg multiple testing method. The sample size includes students observed in all rounds of data collection (234).

with 3 million pesos in a week for an emergency?" are highly correlated across all periods compared. This is not the case for the question "How hard wil it be to cover next week's expenses with the money you have today?" which is highly correlated among the baseline periods but the strength of the correlation drops for all other periods. This finding is surprising because it could be suggesting that because part of the sample transitions to the labor market may have more access to resources and finds easier to cover expenses. However, we find no change in this variable in the difference-in-differences analysis of the previous section (adjusted for multiple inference testing).

In sum, we find correlations of the outcomes across time that are statistically different form zero in virtually all cases. The results do not differ substantially from the literature analyzing the stability of preferences in terms of the magnitude of the persistence. In addition to what other papers have done, we provide correlations of cognitive measures and personal finances variables. We find the highest correlations among social preferences variables and the lowest among one of the cognitive tests.

Table 2.15: Stability of personal finances and inconsistency in the value of money

	(1)		**(2)**		**(3)**		**(4)**	
	Baseline 2		**After offer**		**After paycheck**		**After paycheck**	
	vs. Baseline 1		**vs. Baseline 2**		**vs. Baseline 2**		**vs. After offer**	
	Corr.	Reg.	Corr.	Reg.	Corr.	Reg.	Corr.	Reg.
Hard to come up with money	0.586 (0.053)	0.557 (0.060)	0.531 (0.057)	0.489 (0.063)	0.579 (0.054)	0.552 (0.060)	0.604 (0.053)	0.590 (0.058)
Hard to cover expenses	0.404 (0.075)	0.383 (0.079)	0.207 (0.072)	0.215 (0.084)	0.311 (0.075)	0.340 (0.082)	0.218 (0.071)	0.196 (0.071)

Notes: Correlation coefficients are obtained from the Stata command `correlate`. Standard errors are in parenthesis are bootstrapped with 10,000 replications in the case of correlation and Eicker-Huber-White in the case of regression. The coefficients in the column Reg. are obtained from regressions of the variable on the left in the period mentioned first in the column title on the same variable in the period mentioned second in the column title. For example, in the Baseline 2 vs. Baseline 1 column, the coefficient displayed is from a regression of risk aversion at Baseline 2 on risk aversion at Baseline 1. All regressions control for whether the student is in the comparison group, gender, poverty status and age. All point estimates are significant at the 5% level after adjusting by the Benjamini- Hochberg multiple testing method. The sample size includes students observed in all rounds of data collection (234).

Surprisingly, the correlation among cognitive measures is not very high even though the same instruments were used across all rounds.

Because we find that some of our outcomes do in fact change along the different stages and that in some of them there is an evident increasing or decreasing trend, it is surprising that stability does not change more than we observe. For example, in Section 2.4 we see a significant change in present biasedness that does not translate into a lower correlation between this variable measured at baseline and at the after offer period. The reduction in the correlation coefficients emerges when comparing the baseline and the after paycheck period. Therefore, the standard way of measuring stability in the literature may be hiding important changes in the levels that we are able to describe thanks to our research design.

Also importantly, we provide evidence that mesures of psychological states may be behind some of the apparent changes in preferences as we saw in the case of risk aversion. Simple correlations of the responses in the experimental games are not capable of capturing the relationships between these variables.

2.6 Conclusion

This paper documents the changes in decision making that occur as a result of a major life transition - specifically transitioning from being a college student to a working member of society. When students join college, particularly if it is a prestigious school like the one from which we draw our participants in this study, it may be reasonable to assume that students have certain expectations of finding a job and having financial security. Therefore, similar to the Permanent Income Hypothesis' predictions for consumption, one may not expect to see changes in decision making behavior for risk preferences, time preferences, cognitive performance and other related tasks and decisions. However, our hypothesis was that even though students in such universities are somewhat assured of finding good jobs, there is considerable uncertainty of the specificities of the job, such as the when it will come and how much it will pay. These uncertainties may be large enough to cause changes in decision making merely in response to receiving a job offer, even before being paid for the first time.

In fact, our results bear out this hypothesis quite conclusively. We use a difference-in-differences strategy to study the effect on decision making of first transitioning from a being a (last-semester) college student to receiving a job offer, and then the effect receiving a paycheck. We employ the fact that students about to experience the transition are similar to students in lower years in many dimensions. Therefore, comparing last-semester students to students in lower years (pursuing similar major, having a similar gender distribution and having similar tuition levels) provides us with a reasonable research design. By having lower year students in the comparison group answer the same questions as the last-semester students at roughly the same times, we can effectively compare the their answers across rounds to determine differential trends among these final semester students. Of course, because we are unable to randomly assign the status of being a last semester student, we cannot make strong causal claims about the results. But the patterns we observe are strongly suggestive of the effects that transition to the job market can have on decision making behavior measured through experimental tasks.

We find that there is indeed a change in time preferences, perceptions of financial health and feelings about being tired and worried as a result of merely receiving a job offer. The finding that last-semester students become differencially less present-biased solely in response to a job offer demonstrates what a strong effect the resolution of this job uncertainty can have. These students report it now being less harder to come up with money for an emergency and it being less hard to cover expenses even though they have not been paid in their new jobs just yet. This contradicts the perception that students at a good college would have no uncertainty about getting a job since without any change in their earnings, their perception of their status changes significantly. Furthermore, these students report that they are less worried, tired, depressed and frustrated when they receive their job offer. While this is not surprising, what is striking is the large effect these emotions have in their decision making during the transition. Without accounting for these feelings, students appear to become less risk-averse on receiving their job offer. However, once we control for these emotions, these results vanish. In other cases, the effect of the transition is made stronger, like in the case of becoming less present-biased when the job uncertainty is resolved. Often when studying decision making behavior, such self-reported measures are not taken into account and this could be affecting the interpretation of results.

After receiving at least one paycheck from their new jobs, all the positive effects on perception of financial status we observed in the after job offer period dissipate and are no longer significant. There are no longer significant results on present-biasedness. Furthermore, these students also perform differentially worse on the cognitive reflection task and the Raven's Matrices-type cognitive tests. Finally, after receiving a paycheck, students report being more frustrated, worried and tired. These results are consistent with the hypothesis that after actually receiving some income, these students have to take on many more responsibilities relating to becoming more independent. They may also have to take care of other family members, adding to their stress levels and generating a decrease in the bandwidth available to solve problems (Mullainathan & Shafir, 2013).

It appears that the resolution of uncertainty regarding the details of their job is the crucial factor that induces changes in the decision making of students who transition to the labor market. Their perceptions of their financial health also change positively. However, after starting to work and being paid, there may be greater cognitive load that comes with having a lot more responsibilities that lead to changes in cognitive performance and feelings of worry and tiredness.

2.7 Appendices

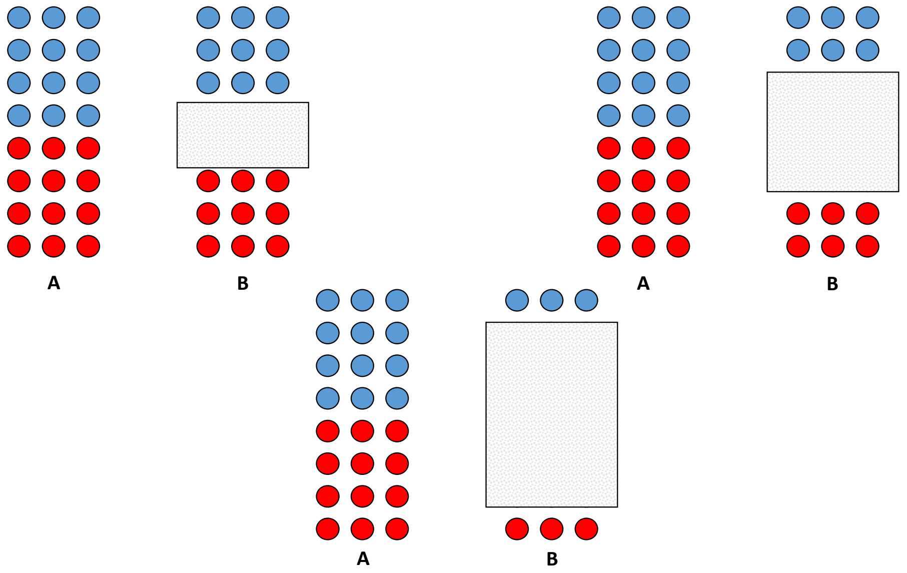
2.7.1 Risk lottery based on Eckel and Grossman (2002)

Row no.	**Column A**	**Column B**
	(if heads comes out)	(if tails comes out)
1	28,000 pesos	28,000 pesos
2	24,000 pesos	36,000 pesos
3	20,000 pesos	44,000 pesos
4	16,000 pesos	52,000 pesos
5	12,000 pesos	60,000 pesos
6	2,000 pesos	70,000 pesos

2.7.2 Risk lotteries based on Tanaka, Camerer and Nguyen (2010)

	Column A		Column B		Exp. payoff diff.
Row no.	**If 1 to 3 comes out**	**If 4 to 10 comes out**	**If 1 comes out**	**If 2 to 10 comes out**	(A - B)
1	4,000 pesos	1,000 pesos	6,800 pesos	500 pesos	770 pesos
2	4,000 pesos	1,000 pesos	7,500 pesos	500 pesos	700 pesos
3	4,000 pesos	1,000 pesos	8,300 pesos	500 pesos	620 pesos
4	4,000 pesos	1,000 pesos	9,300 pesos	500 pesos	520 pesos
5	4,000 pesos	1,000 pesos	10,600 pesos	500 pesos	390 pesos
6	4,000 pesos	1,000 pesos	12,500 pesos	500 pesos	200 pesos
7	4,000 pesos	1,000 pesos	15,000 pesos	500 pesos	-50 pesos
8	4,000 pesos	1,000 pesos	18500 pesos	500 pesos	-400 pesos
9	4,000 pesos	1,000 pesos	22,000 pesos	500 pesos	-750 pesos
10	4,000 pesos	1,000 pesos	30,000 pesos	500 pesos	-1,550 pesos
11	4,000 pesos	1,000 pesos	40,000 pesos	500 pesos	-2,550 pesos
12	4,000 pesos	1,000 pesos	60,000 pesos	500 pesos	-4,550 pesos
13	4,000 pesos	1,000 pesos	100,000 pesos	500 pesos	-8,550 pesos
14	4,000 pesos	1,000 pesos	170,000 pesos	500 pesos	-15,550 pesos

2.7.3 Ambiguity aversion based on Tanaka et al. (2014)



2.7.4 Time preferences based on Andreoni and Sprenger (2012)

EARLIER LATER

Allocate □ to be received next week AND □ to be received in five weeks with a 1% interest

Allocate □ to be received next week AND □ to be received in five weeks with a 10% interest

Allocate □ to be received next week AND □ to be received in five weeks with a 50% interest

Allocate □ to be received next week AND □ to be received in five weeks with a 100% interest

Allocate □ to be received next week AND □ to be received in nine weeks with a 1% interest

Allocate □ to be received next week AND □ to be received in nine weeks with a 10% interest

Allocate □ to be received next week AND □ to be received in nine weeks with a 50% interest

Allocate □ to be received next week AND □ to be received in nine weeks with a 100% interest

2.7.5 Cognitive Reflection Test (CRT)

The questions that were asked in Spanish are a translation or adaptation of the following questions:

- A bat and a ball cost $1.10 total. The bat costs $1.00 more than the ball. How much does the ball cost? (Intuitive error: 10; correct: 5)

- If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? (Intuitive error: 100; correct: 5).

- In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? (Intuitive error: 24; correct: 47)

- Jerry received both the 15th highest and the 15th lowest mark in the class. How many students are in the class? (Intuitive error: 15, 30; correct: 29)

- A man buys a pig for $60, sells it for $70, buys it back for $80, and sells it finally for $90. How much has he made? (Intuitive error: 10; correct: 20)

- Simon decided to invest $8,000 in the stock market one day early in 2008. Six months after he invested, on July 17, the stocks he had purchased were down 50%. Fortunately for Simon, from July 17 to October 17, the stocks he had purchased went up 75%. At this point Simon has (a) broken even in the stock market, (b) is ahead of where he began, (c) has lost money. (Intuitive error: b; correct: c).

2.7.6 Other difference-in-differences results

Table 2.16: DID results: More risk measures

	(3)	(4)	(5)	(6)
	Risk averse	**Risk averse**	**Fraction making inconsistent risk choices**	**Fraction making inconsistent risk choices**
Baseline 1 * Comparison	0.832***	0.897***	0.261***	0.223***
	(0.0277)	(0.0500)	(0.0325)	(0.0635)
Baseline 2 * Comparison	0.837***	0.925***	0.190***	0.235***
	(0.0273)	(0.0553)	(0.0290)	(0.0586)
After offer * Comparison	0.598***	0.216***	0.136***	0.0774**
	(0.0363)	(0.0518)	(0.0254)	(0.0311)
After paycheck * Comparison	0.554***	0.327***	0.0924***	0.0533*
	(0.0368)	(0.0592)	(0.0214)	(0.0277)
Baseline 1 * Last sem.	0.0112	0.00759	-0.0305	-0.0249
	(0.0390)	(0.0389)	(0.0454)	(0.0452)
Baseline 2 * Last sem.	-0.0841**	-0.0945**	-0.0779**	-0.0805**
	(0.0424)	(0.0432)	(0.0375)	(0.0381)
After offer * Last sem.	-0.132**	-0.0139	-0.0615*	-0.0531
	(0.0549)	(0.0526)	(0.0333)	(0.0365)
After paycheck * Last sem.	0.0891	0.0531	-0.0141	-0.0238
	(0.0580)	(0.0598)	(0.0330)	(0.0339)
Observations	1,355	1,355	1,355	1,355
R-squared	0.725	0.752	0.181	0.195
Emotion controls	NO	YES	NO	YES

Robust standard errors in parentheses

*** p<0.01, **p<0.05, * p<0.1

Multiple Inference: † † † p_m<0.01, †† p_m<0.05, † p_m<0.1

Table 2.17: DID results: Social preferences

	(1)	(2)	(3)	(4)	(5)	(6)
	Fraction to student - dictator	**Fraction to student - dictator**	**Fraction to foundation - dictator**	**Fraction to foundation - dictator**	**Fraction to student - ultimatum**	**Fraction to student - ultimatum**
Baseline 1 * Comparison	0.331***	0.299***	0.553***	0.608***	0.438***	0.417***
	(0.0179)	(0.0352)	(0.0232)	(0.0478)	(0.0136)	(0.0268)
Baseline 2 * Comparison	0.279***	0.323***	0.477***	0.491***	0.423***	0.474***
	(0.0175)	(0.0384)	(0.0235)	(0.0492)	(0.0139)	(0.0302)
After offer * Comparison	0.343***	0.716***	0.468***	0.778***	0.464***	0.792***
	(0.0261)	(0.0373)	(0.0282)	(0.0394)	(0.0220)	(0.0307)
After paycheck * Comparison	0.340***	0.604***	0.477***	0.612***	0.467***	0.680***
	(0.0270)	(0.0514)	(0.0287)	(0.0536)	(0.0241)	(0.0431)
Baseline 1 * Last sem.	0.0362	0.0358	-0.0314	-0.0374	0.0249	0.0262
	(0.0255)	(0.0257)	(0.0339)	(0.0340)	(0.0189)	(0.0188)
Baseline 2 * Last sem.	0.0569**	0.0552**	-0.0162	-0.0141	0.0258	0.0204
	(0.0260)	(0.0267)	(0.0336)	(0.0340)	(0.0202)	(0.0205)
After offer * Last sem.	0.209***†††	0.0924***††	0.132***†††	0.0359	0.149***†††	0.0505*
	(0.0416)	(0.0351)	(0.0426)	(0.0391)	(0.0355)	(0.0304)
After paycheck * Last sem.	-0.0988***	-0.0524	-0.0852**	-0.0543	-0.0957***	-0.0631**
	(0.0368)	(0.0361)	(0.0432)	(0.0445)	(0.0319)	(0.0314)
Observations	1,355	1,355	1,355	1,355	1,355	1,355
R-squared	0.588	0.669	0.673	0.702	0.779	0.826
Emotion controls	NO	YES	NO	YES	NO	YES

Robust standard errors in parentheses

*** p<0.01, **p<0.05, * p<0.1

Multiple Inference: † † † p_m<0.01, †† p_m<0.05, † p_m<0.1

Table 2.18: DID results: Psychological measures

	(1) **Frustra-tion**	(2) **Depression**	(3) **Worry**	(4) **Enjoyment**	(5) **Tired**
Baseline 1 * Comparison	0.196***	0.141***	0.402***	0.511***	0.598***
	(0.0294)	(0.0258)	(0.0363)	(0.0370)	(0.0363)
Baseline 2 * Comparison	0.337***	0.239***	0.538***	0.484***	0.603***
	(0.0350)	(0.0316)	(0.0369)	(0.0370)	(0.0362)
After offer * Comparison	0.283***	0.239***	0.424***	0.446***	0.435***
	(0.0333)	(0.0316)	(0.0366)	(0.0368)	(0.0367)
After paycheck * Comparison	0.239***	0.201***	0.266***	0.543***	0.223***
	(0.0316)	(0.0297)	(0.0327)	(0.0369)	(0.0308)
Baseline 1 * Last sem.	-0.0159	0.0160	-0.0707	-0.0727	-0.0585
	(0.0412)	(0.0376)	(0.0507)	(0.0526)	(0.0522)
Baseline 2 * Last sem.	-0.0785	-0.0256	-0.128**	0.0500	-0.0976*
	(0.0481)	(0.0441)	(0.0523)	(0.0527)	(0.0522)
After offer * Last sem.	-0.147***†††	-0.151***†††	-0.201***†††	-0.0808	-0.232***†††
	(0.0437)	(0.0393)	(0.0502)	(0.0541)	(0.0495)
After paycheck * Last sem.	0.0652	0.0163	0.108*	0.0217	0.116**
	(0.0534)	(0.0487)	(0.0559)	(0.0593)	(0.0540)
Observations	1,355	1,355	1,355	1,355	1,355
R-squared	0.257	0.201	0.398	0.492	0.491
Emotion controls	NO	NO	NO	NO	NO

Robust standard errors in parentheses

*** p<0.01, **p<0.05, * p<0.1

Multiple Inference: † † † p_m<0.01, †† p_m<0.05, † p_m<0.1

Table 2.19: DID results: Cognitive performance - Additional tasks

	(1) **Stroop Test**	(2) **Stroop Test**	(3) **Flanker Test**	(4) **Flanker Test**
Baseline 1 * Comparison	16.84***	14.75***	28.45***	27.95***
	(0.463)	(1.053)	(0.735)	(1.478)
Baseline 2 * Comparison	17.84***	16.77***	30.58***	28.69***
	(0.451)	(0.976)	(0.646)	(1.191)
After offer * Comparison	18.24***	16.94***	29.67***	29.96***
	(0.485)	(1.060)	(0.767)	(1.772)
After paycheck * Comparison	19.78***	18.26***	31.69***	29.90***
	(0.433)	(0.695)	(0.699)	(1.404)
Baseline 1 * Last sem.	-0.399	-0.259	-0.708	-0.638
	(0.714)	(0.726)	(1.046)	(1.056)
Baseline 2 * Last sem.	0.530	0.752	-0.856	-0.538
	(0.609)	(0.614)	(0.886)	(0.895)
After offer * Last sem.	0.314	0.460	0.847	1.011
	(0.842)	(0.824)	(1.327)	(1.418)
After paycheck * Last sem.	-0.183	-0.199	-0.397	-0.244
	(0.629)	(0.635)	(1.067)	(1.085)
Observations	1,228	1,228	1,229	1,229
R-squared	0.901	0.904	0.914	0.915
Emotion controls	NO	YES	NO	YES

Robust standard errors in parentheses

*** p<0.01, **p<0.05, * p<0.1

Multiple Inference: † † † p_m<0.01, †† p_m<0.05, † p_m<0.1

CHAPTER III

Biased beliefs, performance and effort: Experimental evidence from a pilot in Colombia

3.1 Introduction

Beliefs about ourselves and about others are an important input in decision-making under uncertainty. However, beliefs are not always accurate. In fact, laboratory experiments document large biases in how people perceive their ability to perform a task. In the lab, everyone is miscalibrated by overestimating own performance. Interestingly, women and men differ in their self-assessments. For example, the most influential paper in experimental economics of the last decade found that, when performing a simple math task, everybody overestimates own performance and women believe they are ranked lower than men of similar ability (Niederle & Vesterlund, 2007). In a related context, studies have shown that biased beliefs or incorrect updating can affect the decision to enter a competitive environment in the laboratory (Mobius, Niederle, Niehaus, & Rosenblat, 2011; Berlin & Dargnies, 2016).

The lab has also shown that feedback can potentially help correct biased priors and improve decision-making. Within the gender and competitiveness literature, Wozniak, Harbaugh, and Mayr (2014) show that providing exact feedback about relative rank closes the gender gap in willingness to compete in the lab. Noisy feedback as in Mobius et al. (2011) or feedback about being at the top or bottom of a performance distribution as in Berlin and Dargnies (2016) affects competitive-entry decisions but asymmetrically depending on the nature of the signal received (positive or negative).

Outside of the lab, however, there is little evidence of the existence and magnitude of biased beliefs and how they relate to real-life decisions in which performance in tests or at work can lead to fundamentally different paths among people of the same ability. In the human capital formation literature, for example, there is evidence that individuals' investments are affected by the perceived returns to education (Jensen, 2010), beliefs about future earnings (Wiswall & Zafar, 2015; Reuben, Wiswall, & Zafar, 2015), information about school quality (Hastings & Weinstein, 2008; Mizala & Urquiola, 2013), and information about application to schools and financial aid (Hoxby & Turner, 2015; Dinkelman & Martínez, 2014). There is less evidence on how beliefs about relative ability affect decisions like how to allocate study time and which college majors to choose based on comparative advantages in academic subjects. This research contributes to filling this gap.

I study gender differences in beliefs about ability / performance in academic subjects and in the updating process among students preparing for a high-stakes college entrance exam in Colombia. Specifically, my research questions are: Are beliefs about own ability biased? Do biases differ by gender? Does feedback correct biased priors and affect effort (hours of study), performance in tests, and perceived difficulty of tests?

I provide new evidence of the magnitude and direction of the biases in self-assessments of performance in a real-life scenario, and of the extent to which feedback can correct biased priors. To do this, I take advantage of the multiple practice tests that students take as part of their test-preparation course. Every week, students take a practice test after which I elicit beliefs about being in the four quartiles of the score distribution in each of the five areas covered by the test (math, science, social science, image analysis and text analysis). After eliciting beliefs from all participants for a few rounds, students are divided into treatment and control groups. Students in the treatment group receive feedback about how their scores relate to those of the rest of the students at the institute and to their stated priors. I then collect more data about beliefs, allocation of study time and perceived difficulty of the practice tests.

I find substantial biases in assessing own ability. Across all areas of the test, between 50 and 70 percent of the students fail to correctly predict the quartile in which their score will be. Women are more biased than men. For example, almost 50 percent of the males but only 32 percent of women accurately predict their score in the math section of the test. Women perform worse in math and at the same time are more likely than the men to underestimate their performance, especially when predicted performance is in the worst quartile but actual performance is in the best. In text analysis, women underperform men but are more likely to overestimate their performance. These findings point to biases in the direction predicted by stereotypes about relative academic advantages of men and women.

Feedback provision with the features implemented in this pilot does not have significant effects on performance, beliefs in subsequent practice tests or allocation of study time. However, feedback affects the perceived difficulty of subsequent practice tests, an effect entirely driven by males in the treatment group who rate the tests as harder independently of the type of signal they receive about their performance. Women tend to be less confident than men about their ability to be admitted at the university they are preparing for. When receiving feedback, women become more confident and men less confident about their expected admission outcome so the gender gap in confidence in this dimension seems to disappear.

Two papers that are directly relevant to this research are Bobba and Frisancho (2016) and Gonzalez (2017). Both provide evidence that students overestimate their performance in a mock test. The former paper studies how middle-school students in Mexico City update their priors after receiving information about a mock exam score which leads those who receive positive feedback to choose, on average, more academically-oriented high school options. The latter provides evidence on how students who receive (and are instructed about) the Advance Placement (AP) potential signal after taking the practice SAT test in the US are more likely to take and pass AP courses. Both papers point to substantial biases in beliefs about own ability and how informing students of their performance can affect decision making and better align skills with academic options.

Relative to those papers, this research finds that self-assessment varies depending on the academic subject. Hence, analyzing beliefs and feedback about a single score composed of many subjects may conflate different directions and magnitudes in the biases. Further, this study shows that women are more miscalibrated than men in opposite directions in different subjects which is something that those studies are not able to detect with their research design.

This paper is divided in sections. Section 3.2 presents relevant details of the background and the experimental design. Section 3.3 discusses the data and descriptive statistics emphasizing in the gender dimension. In section 3.4, I present the main results regarding performance and beliefs, the type of signals students in the treatment group received, as well as the treatment effects of feedback. The last section concludes.

3.2 Background and experimental design

I partnered with a test-preparation institute in Colombia that prepares students who plan to take a high-stakes college admission test. This test is taken by about 60,000 students every semester who compete for a fixed number of slots (about 5,000) in all campuses of the university. Upon admission, applicants have to declare two college-major options. Slots in the different majors are allocated based on the student's overall score and the number of available slots remaining at the time the student is enabled to access the system to choose his or her preferred majors.[1] All students take the same exam regardless of their intended majors. Because the only criterion for admission at this university is the entrance exam and given the prestige of its academic programs and subsidized tuition, there is high demand of test preparation services from students willing to attend this university.

The test preparation institute gave me permission to recruit its students in Bogota and Medellin who intended to take the entrance exam on April 23, 2017. Students enrolled in this course are prepared over two-and-a-half months in all five areas of the admissions exam: math, science (physics, chemistry and biology), social science (history, geography, philosophy), image analysis, and text analysis.[2] Students at-

[1] At this university and, generally in Colombia, students declare their major before starting college. Admissions are based on the entrance exam score only, and given the number of slots per major, they are very competitive.

[2] More information about each component of the test can be found in

tend three-hour classes from Tuesday to Friday and take a full-length practice test (3 hours) every Monday. Enrollment in these courses is of about 1,100 students in Bogota and 260 in Medellin. Furthermore, after every practice test the test preparation institute provided me with practice-test scores of all students taking the same preparation course as students in my sample.

My research consists of eliciting beliefs about relative performance in each area of the test and randomly assigning students to a treatment group which receives feedback about their performance in the last practice test relative to the rest of students, and to a control group which does not. After each practice test, I elicited beliefs about the probabilities of being at each of the four quartiles of the practice test score distribution for each of the five areas of the exam. The timeline of activities relevant for the design are as follows: 1) students take weekly full-length practice tests (given by the institute), 2) students take beliefs survey immediately after each practice test, 3) tests are graded, 4) institute sends the scores of all students to the researcher, 5) the researcher produces and sends feedback notifications depending on whether the student is in the treatment or the control group.

Feedback was given on paper to maximize the chances that students will see it. Due to lags in the release of practice test results, it was not possible to provide weekly feedback after each practice test as would have been ideal. Instead, a summary of students' predictions and performances was given for three rounds of the latest practice tests with scores available to the researcher. A series of five graphs (one for each section of the exam) containing actual and predicted quartiles allowed students in the treatment group to see a summary of their performance as well as how well calibrated they were in their assessment of performance relative to other students taking the same course as them.[3]

http://admisiones.unal.edu.co/pregrado/panel-2-informacion-sobre-las-pruebas/prueba-de-admision/. An example of image analysis questions are in appendix 3.6.1.

[3] Even though students enrolled in this type of courses are not representative of the usual pool of applicants, relative comparisons with this group is relevant given that other students taking the same course may have similar backgrounds and knowledge about the admissions exam. Further, the focus of this research is not on assessing the probability of passing an admissions test but rather on signaling relative comparative advantages / disadvantages across academic subjects among students with similar characteristics. A mapping of where the students are relative to all applicants would be ideal to have but is unfortunately out of the scope of this chapter. This is left as future work.

Besides the multiple rounds of the beliefs survey, students are required to fill out a baseline survey. At sign up they fill out a baseline survey with questions about demographics, previous education, favorite school subjects, intended majors, and a series of incentivized experimental questions to assess their IQ, confidence, risk aversion, and competitiveness as is standard in the gender and competitiveness literature.

The statistical comparison between beliefs and actual performance allows me to establish whether beliefs are biased and whether the biases differ by gender. By comparing the treatment to the control group, I test to what extent biased beliefs are updated, i.e. whether individuals are more likely to hold more accurate beliefs after receiving feedback. Moreover, the random assignment identifies the causal effect of receiving feedback on effort, performance in the next practice test, perceived difficulty of the practice tests, and confidence in being admitted to this university. Hence, if I observe that these outcomes differ between the treatment and control groups, I will be able to conclude that it is a result of belief updating through feedback provision given that ex-ante, students in both groups are equivalent in terms of observable characteristics.

To incentivize participation, a raffle of laptops and cash prizes of 100,000 pesos (about \$35) takes place at the end of the study. Moreover, smaller cash prizes of 15,000 pesos or about US\$5 were distributed every week to guarantee truthful reporting of beliefs using the crossover mechnism explained in Mobius et al. (2011). To win the prizes, students accumulate lottery tickets based on incentivized questions in the baseline and end surveys as well as in each of the beliefs surveys.

3.3 Data and descriptive statistics

I present results from 208 students taking the preparation course in the city of Bogota. Students were recruited by visiting classrooms in early February, 2017. Students intersted in participating filled out a sign-up sheet with their name, email and age. They were then contacted by email with instructions about how to sign the consent form and the parent's assent in case the were minors. After consenting, they were required to fill out the first survey online. Tables 3.1 through 3.3 present tabulations from the baseline survey.

Table 3.1 shows that women constitute about two-thirds of the sample. From conversations with the test preparation institution, women tend to have a higher demand of this type of courses. The average age is 18 and about 40 percent of the sudents live in a poor household according to their level in the Identification System for Social Program Potential Beneficiaries (SISBEN). About 56 percent studied in a private high school, and more women relative to men studied in a religious high school. About a third self-report to have received academic honors and to ever worked full time in the past. More women than men report having received academic honors but with the small sample size, the difference is not found to be statistically significant. Slighltly more than a third of students in the sample report that their parents' education level is college or higher.

Table 3.1: Demographic and schooling characteristics

	Male	**Female**	**p-value diff.**
Fraction	37.02	62.98	
Age	17.99	18.13	0.537
Poor (SISBEN)	0.35	0.44	0.194
HS private	0.57	0.56	0.843
HS religious	0.09	0.19	0.054
HS mixed-sex	0.25	0.31	0.368
Academic honors	0.29	0.39	0.132
Ever worked	0.32	0.33	0.958
Mother college or more	0.47	0.35	0.098
Father college or more	0.36	0.36	0.944

I collected a rich set of questions about their intended majors, reason for choosing them, and their experience and expectations related to the admissions exam. Table 3.2 shows that students who are enrolled in this course have taken the real admissions exam once in the past, on average. The exam score is stardardized with a mean of 500 and a standard deviation of 100. Scores of admitted students for most majors in the Bogota Campus of the university are of at least 625 points although it varies on a semester basis.[4] On average, men and women predict that the score the will obtain

[4] This minimum score is based on statistics published by the university but it varies depending on the majors. The majors with the highest demand have much higher minimum scores. Admission scores can be consulted at: http://www.admisiones.unal.edu.co/servicios-en-linea/estadisticas-del-proceso-de-admision/

Table 3.2: Intended majors and predicted scores in admissions test

	Male	**Female**	**p-value diff.**
No. times taken exam	1.01	1.08	0.528
Predicted score before course	629.34	630.62	0.937
Predicted score after course	753.07	741.23	0.378
Expected increase in score	122.42	110.61	0.489
Min score to pass to 1st choice	715.11	725.25	0.275
Min score to pass to 2nd choice	685.31	703.68	0.039
First choice major in science	0.04	0.09	0.158
First choice major in econ sci	0.06	0.02	0.056
First choice major in hum	0.04	0.15	0.011
First choice major in health sci	0.32	0.41	0.211
First choice major in arts	0.14	0.12	0.670
First choice major in law	0.08	0.09	0.736
First choice major in engineering	0.31	0.11	0.000
Reason for 1st choice:interesting	0.78	0.85	0.165
Reason for 1st choice:unsure	0.08	0.11	0.496
Second choice major in science	0.09	0.09	0.987
Second choice major in econ sci	0.03	0.03	0.850
Second choice major in hum	0.17	0.22	0.364
Second choice major in health sci	0.13	0.19	0.259
Second choice major in arts	0.12	0.15	0.474
Second choice major in law	0.06	0.05	0.733
Second choice major in eng	0.23	0.12	0.036
Doesn't plan to choose 2nd major	0.06	0.08	0.621
Second choice major undecided	0.1	0.05	0.176
Reason for 2nd choice:interesting	0.43	0.44	0.843
Reason for 2nd choice:unsure	0.26	0.35	0.173
Reason for 2nd choice:by default	0.17	0.05	0.006
Reason for 2nd choice:Able to pass	0.04	0.09	0.158

with the preparation they have up to the point in which they take the survey is of about 630 points. Moreover, including the preparation from that point to the date of the exam, they expect to obtain a score of around 750 points (which is enough to pass to any major in most semesters). That is, they expect an increase in their score from baseline to the exam date of 122 points in the case of men and 111 for women.

Students in my sample taking this preparation course seem to be relatively accurate about the minimum score needed to pass to their intended first choice. Men report that they need about 715 points to pass to their desired first choice major and women report that they need 725 points. For second-choice major, which is supposed to be a back-up option for students who are not admitted to their first major, men report that they would need a score of 685 points to pass to their intended second option. Women, on the other hand, report that they would need a score of 704 points on average. The difference between these two scores is statistically significant and, coupled with the fact that women report higher scores needed to pass in both the first and second choices, it may suggest that women's perceptions about scores may be different than men's perceptions even though both must have the same information as it is provided by the test preparation institute. These figures are even more intriguing given that women tend to choose majors in which the threshold score is normally lower relative to the majors that men typically choose.

In terms of intended first-choice majors, there are statistical differences in the choices of men and women in fields of economic sciences at the 10 percent level, and in majors related to humanities and engineering at the 5 percent and 1 percent levels, respectively. There are not statistical differences in the second choices that men and women expect to declare except in the case of engineering.

Finally, I asked a series of experimental tasks that are standard in the gender and competitiveness literature starting with Niederle and Vesterlund (2007). As previously documented by Eckel and Grossman (2002) and Eckel and Grossman (2008), women are more risk averse (p-value <0.05). The confidence and competitive questions are based on their performance on an IQ test similar to the Raven's progressive matrices test which does not depend on numerical or verbal ability. Out of 9 questions to be solved in 3 minutes, men respond 3.81 questions correctly on average while women correctly respond 3.26 on average. The difference is not large but is statistically significant. On average they are able to respond 7 questions in total.

When asked about which quartile of the distribution of scores in the IQ test they thought their score would be in, about 43 percent of participants guessed their quartile correctly regardless of gender (Table 3.3). In this context, women seem to be more likely to overestimate their performance relative to men although the gender differences are not statistically significant. In the Niederle and Vesterlund (2007) questions, we see that men are more likely than women to think that they are ranked first in a group of 4 randomly chosen participants. Similarly, men are more likely to choose the tournament in which only the winner in the group of four wins the compensation. These two findings replicate the Niederle and Vesterlund (2007)

Table 3.3: Experimental tasks

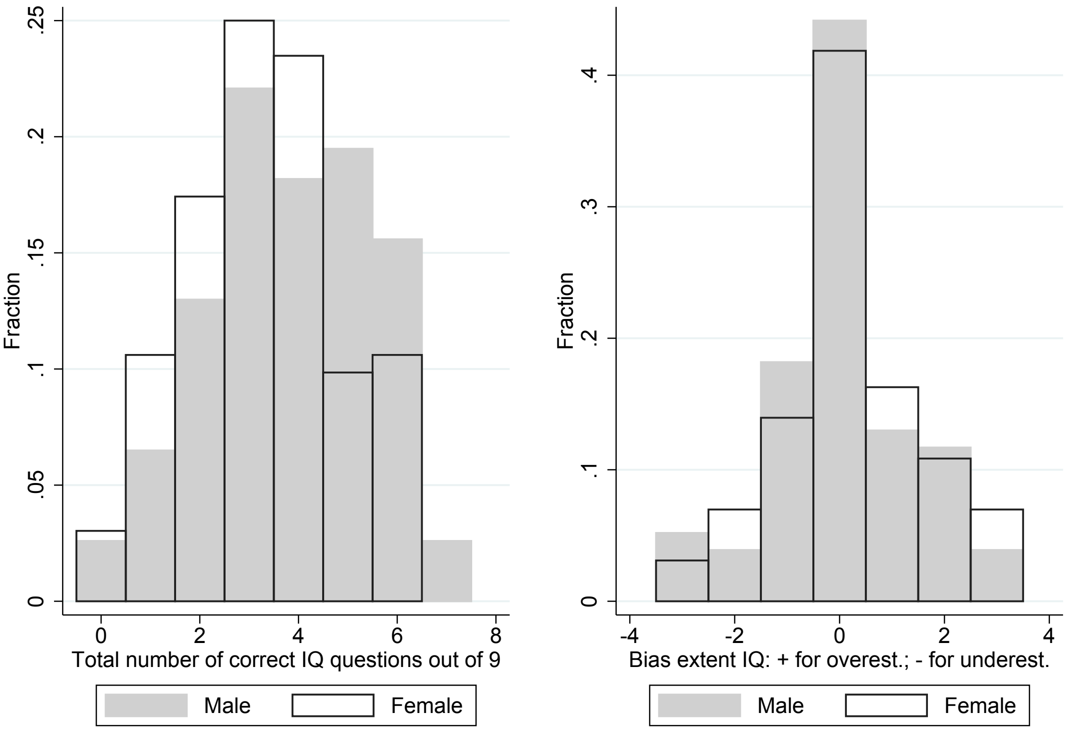
	Male	**Female**	**p-value diff.**
Risk row:1 most risk averse	3.08	2.82	0.253
Risk averse	0.75	0.86	0.047
IQ total correct	3.81	3.26	0.018
IQ total responded	6.78	7.08	0.233
Correct quartile guess	0.44	0.43	0.844
Overestimated	0.27	0.32	0.449
Undrestimated	0.29	0.23	0.381
Rank 1st in group of 4	0.2	0.1	0.175
Rank 2nd in group of 4	0.74	0.78	0.654
Rank 3rd in group of 4	0	0	.
Rank 4th in group of 4	0.07	0.12	0.344
Chose tournament	0.21	0.15	0.255

results although they are not as extreme in this setting as in their original study.[5]

Figure 3.1 shows the distribution of performance and bias in the quartile prediction in the IQ test along with the p-value of the difference between the two distributions. From now on, the first quartile represents the top scores and the fouth quartile the bottom scores. The hollow bars denote females while the gray bars denote males. Even though the two performance distributions do not differ from each other according to the chi-square test, in the left panel it is evident that the distribution for women is shifted to the left of the men's distribution. In the right panel, we see that an important group of men and women are right about their prediction. However, a large proportion hold biased beliefs about their performance. In the graph, negative numbers mean that students underestimate their performance, i.e. they think they performed worse than they actually did. The positive numbers mean that they thought they performed better than they did. Although, as mentioned previously, there are no statistical differences in over- or under-predicting, the graph shows that women are more likely to be in the positive side of the distribution, that is, they tend to overestimate. As expected, most over- or under-estimation occurs by one quartile of difference between the actual and guessed quartile. Less than 20 percent of the students over- or under-estimate their performance by the maximum amount, i.e., thinking they are in quartile 1 while in reality they are in quartile 4 and the reverse.

[5] Keep in mind that the task in Niederle and Vesterlund (2007) is different from the one used here. The addition of two-digit numbers is not possible in the online survey context because students can easily use calculators.

Figure 3.1: Performance and beliefs in IQ test



P-value diff.=0.182

P-value diff.=0.795

In terms of randomization, I stratified randomization at the classroom level because it was not feasible to provide feedback on paper to some students and not to others in the same classroom. The stratification consisted in classifying classrooms according to gender composition and average IQ level of the participants in the same classroom. Four strata are then generated: high number of women and average IQ above the median, high number of women and average IQ below the median, low number of women and average IQ above the median, and low number of women and average IQ below the median. Once the classrooms are assigned to these strata, randomization of the feedback treatment was done at the classroom level. In total, about 55 percent of the students in the sample received feedback. Only 4 out of 52 covariates were unbalanced. I control for these covariates in the effects of feedback regression.[6]

[6] The imbalanced covariates were: Having ever worked full time, number of times that the exam has been taken (0.2 difference), whether the first choice intended major is in the humanities, and whether the second choice is selected because they think they are able to be admitted to that major.

3.4 Results

In this section I characterize the extent and direction of biases in self-assessment of relative performance by presenting evidence from four rounds of beliefs elicitation before providing feedback. Furthermore, I analyze how the outcomes of interest change relative to the control group once students in the treatment group receive feedback.

Recall that rather than giving a realistic idea of where the students are in the distribution of all applicants to help them predict their chances of admission, the purpose of this research is to point out students' miscalibration on their perceptions about how good they are in certain subjects. I argue that having a more accurate idea of thier strengths and weaknesses even if not in relation to the actual pool of applicants, can help students make better decisions in terms of allocating study time and even inform them of what college majors are more aligned with their relative strengths.

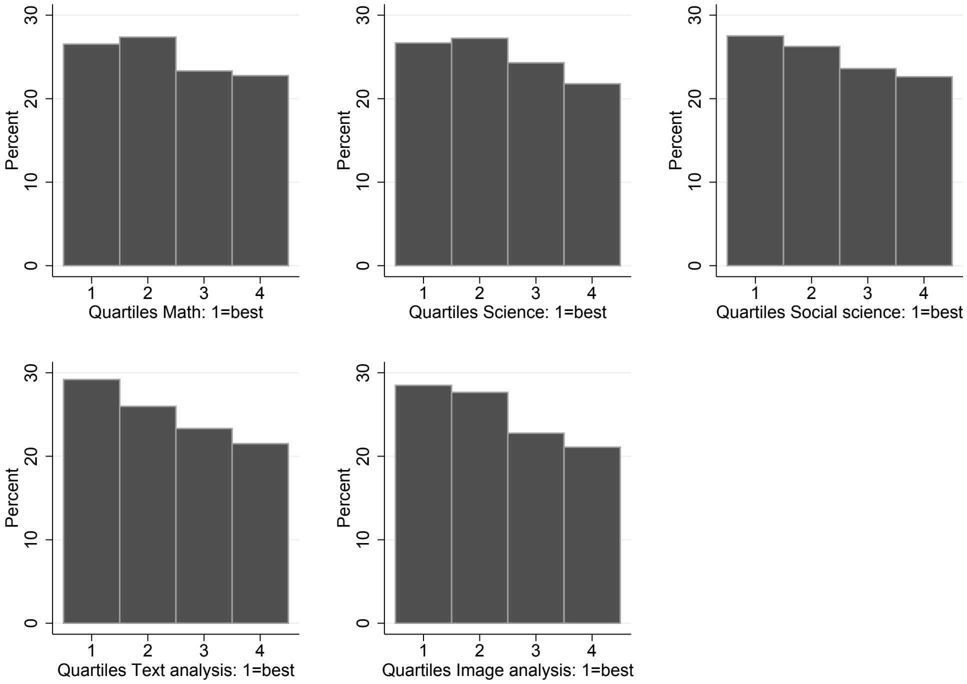
3.4.1 Sample selection

Because not all students at the test preparation institute signed up to participate in the study, one important question is how the sample who selected to be part of the study differs from the rest of the students who are taking the test preparation course. For each section of the practice test, Figure 3.2 shows that my sample is more or less representative of the population of students taking the same course than them. The only cases in which selection may be occuring from the top of the distribution is in text and image analysis in which the percentage of students in the best quartile is closer to 30 percent that to 25 percent.

3.4.2 Performance and beliefs

Figure 3.3 shows the performance of study participants in the five sections of the test: Math, science, social science and image analysis. As before, the statistical difference in the two histograms is given by the p-value at the bottom of each graph. The scale of the scores is from 0 to 10, where 10 means that the students had all questions correct in that area of the exam. According to conversations with the test preparation institute, the practice tests are designed to be much harder than the

Figure 3.2: Quartiles represented in sample by area of the test

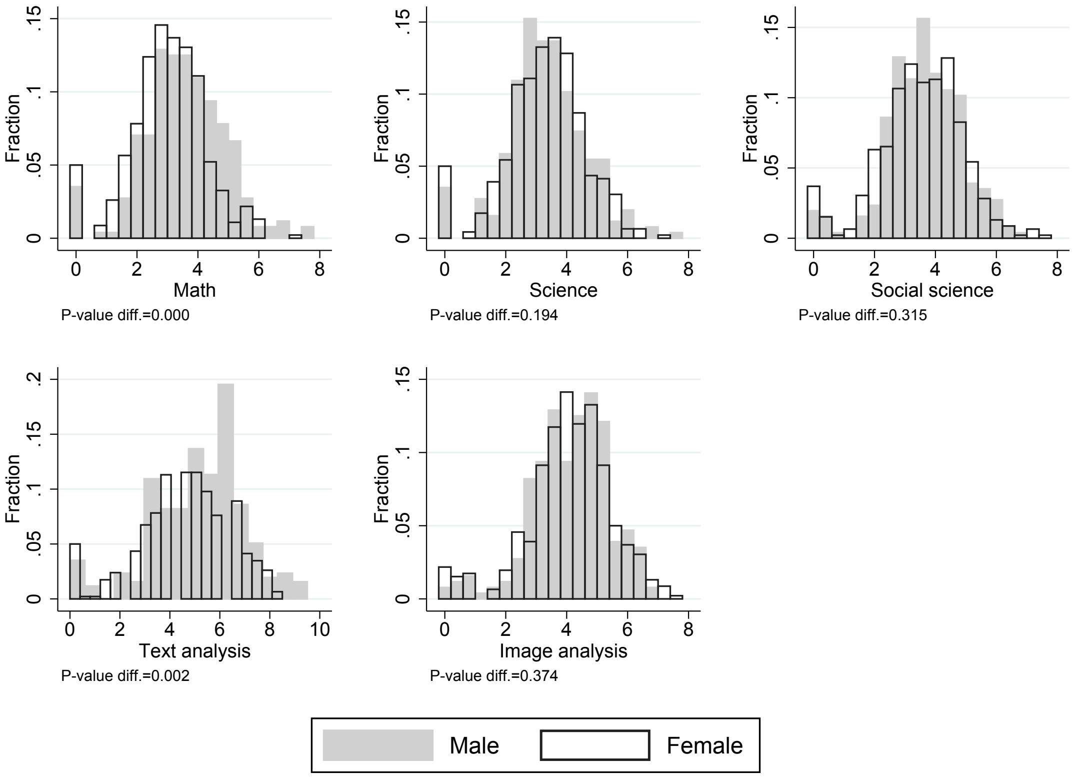


actual entrance exam. Hence, it is very unlikely that anyone correctly answers all questions in any given area. In the figure, all the distributions seem to have a bell shape although there is a mass point at zero.

In math and text analysis, the performance distribution of females is shifted to the left of the distribution of males. In fact, the p-value of the difference between the two histograms is below 0.001 in math and 0.002 in text analysis. The performance in other areas of the test is not statistically different across genders.

I now analyze beliefs regarding performance in each area of the practice test. As before, a value of zero in the histogram means that they were right in their assessment of the quartile of the distribution in which their score will be located. Negative values mean that the students underestimate their quartile, that is, are in a higher quartile than they thought. Positive values mean that they overestimate or think they they belong to a quartile with higher scores than they actually do.

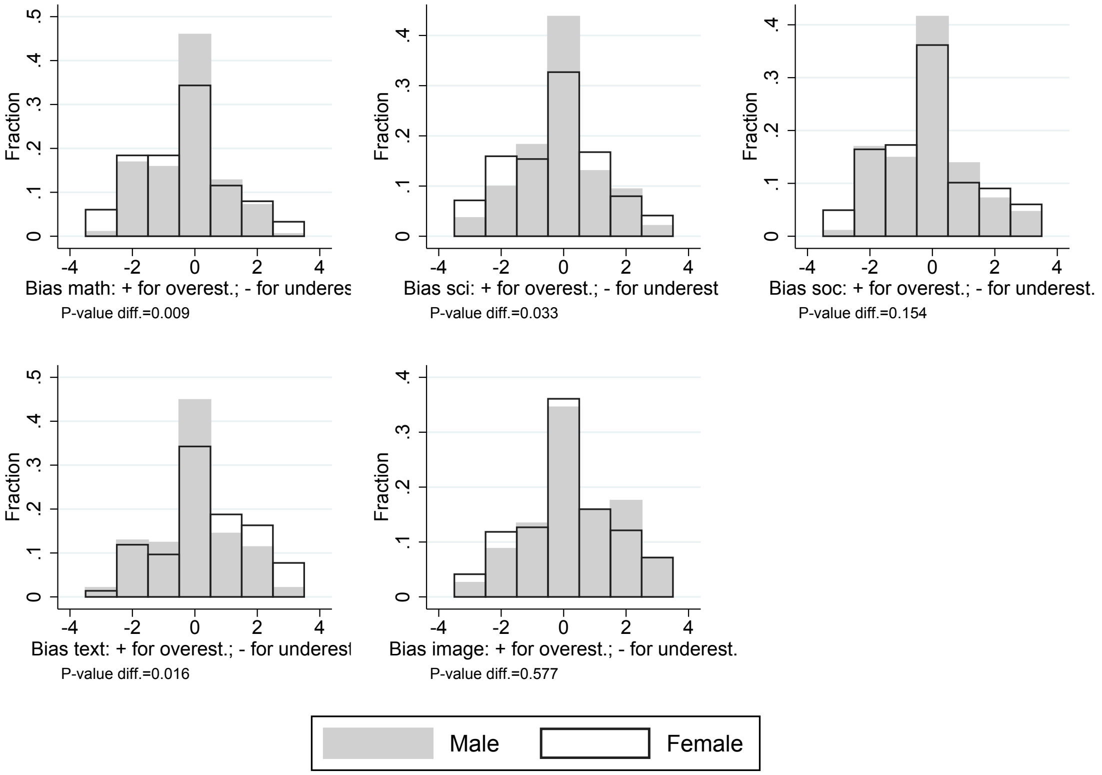
Figure 3.3: Performance in practice test



The top left panel of Figure 3.3 shows the extent of the biases for math. While almost 50 percent of the men accurately predict their math quartile, less than 35 percent of women have a correct assessment. The p-value of the difference in the distribution of biases is 0.009 suggesting that men and women are differentially biased. Overall, the graph shows that students tend to underestimate more than they overestimate and that women underestimate more relative to men. In fact, in the highest degree of underestimation (thinking that they were in the worst quartile but they are actually in the best) there are virtually zero men but about 6 percent of all women. Women are also more likely than men to overestimate but this happens to a smaller extent than underestimation.[7]

[7] Bobba and Frisancho (2016) did not find differences in beliefs between boys and girls. Some reasons why their finding differs from the finding in this paper are that the students in their sample are in middle school, the nature and stakes of the test they analyze are substantially different than in this context, and they only elicit aggregate beliefs across all subjects asked in the test.

Figure 3.4: Beliefs about performance in practice test



In science, 42 percent of the men and 32 percent of the women are right in their quartile prediction. Women are more likely to underestimate and overestimate but there is no clear pattern of which type of bias is more prevalent. In social science we see again that females are less accurate in their predicition than men altough the difference is smaller than in other subjects. Women are also more likely to underestimate to the largest extent (-3 in the graph) more than men.

In text analysis, men are more accurate than women by about 10 percentage points. In this case the direction of women's bias is to think that they performed better than that actually did. Recall that the two areas in which women were performing worse than men were math and text analysis but clearly their perceptions about their performance are opposite across these two subjects.

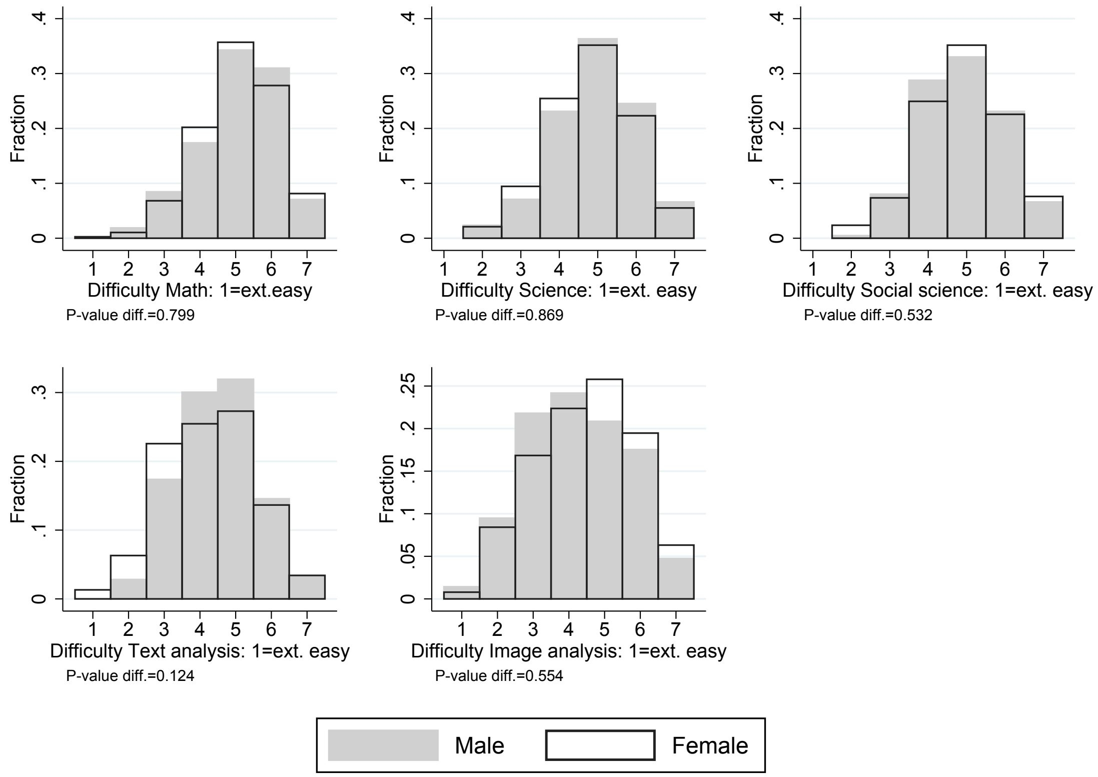
Finally, the only area in which there are no differences between the predictions of men and women is in image analysis. In this case the women are slighlty more likely to make a correct prediction although this is only in about 35 percent of the cases. In this sense, across all areas of the exam, on average 30 percent of the women make correct predictions but men are better at predicting across all areas except image analysis. It is worth noting that image analysis is not a subject taught in high school so it could be the case that women have more biases in subjects that they are familiar with but not in what is not so familiar to them.

One possible explanation of the higher biases and degree of the bias in the case of women is that they perceive that the exam is harder than what men perceive. In a scale from 1 to 7, where 1 is extremely easy and 7 is extremely hard, they are asked to rate the difficulty of each section of the practice test they just took. Figure 3.5 shows that there are no substantial gender differences in their perception. Actually, in text analysis, proportionally more women think that the test was less hard than men. In image analysis more women give a rating of 5, 6 and 7. Despite these differences, the statistical test conclude that they are not significant.

I also collect information about how many hours they studied for each section of the exam during the last week. My prior was that if you feel weak in certain area as reflected by underestimating your performance in the test, it may be the case that you study more for that area than for others. From that perspective, we would see women dedicating more study hours to math and text analysis than to other areas. Figure 3.6 shows that, overall, women are studying more than men in all subjects except science. However they do not necessarily study more forthe areas in which they feel weak. They study more math in which they feel weak but also study more text analysis in which they feel strong.

Regarding confidence about passing the exam on April 23 (obtaining a high enough score to be admitted to their intended majors), I construct a measure in which 5 means that the are very confident in passing. The two histograms in Figure 3.9 show observations from February on the left panel and from March in the right panel. In February, women seem substantially less confident than men but their ratings become closer to those of men in March. Lower confidence in being admitted may be capturing the fact that women perform worse than men in some areas of the

Figure 3.5: Perceived difficuty of each section of the test

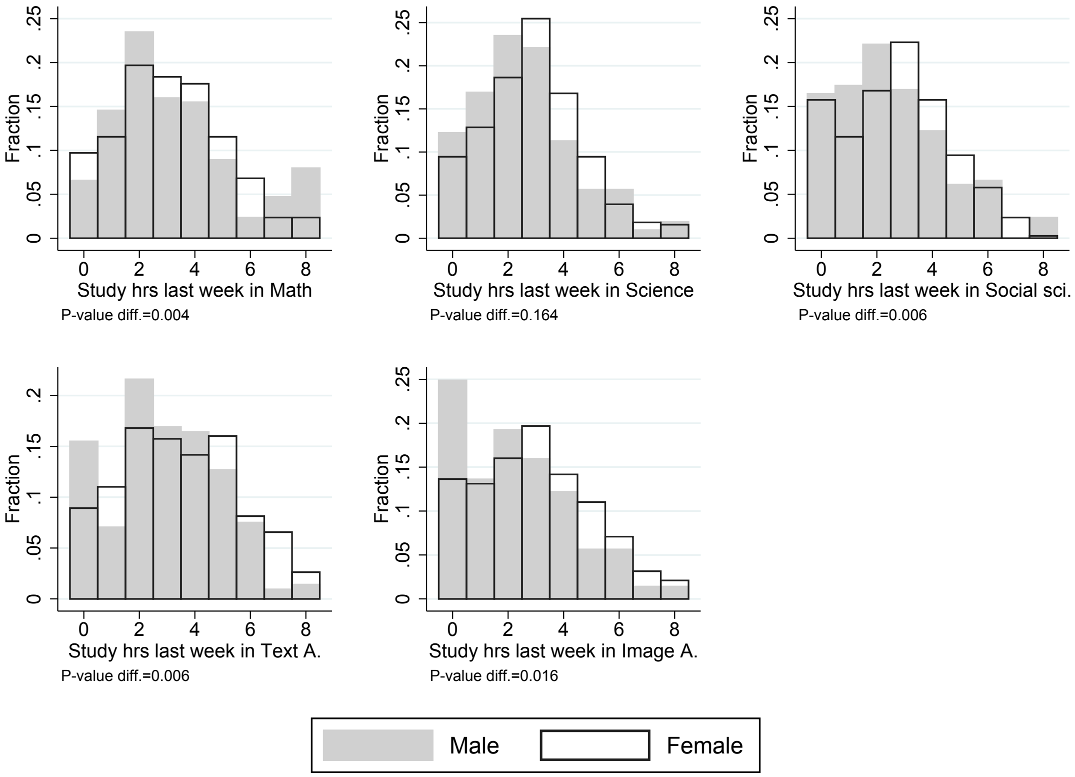


test but also a higher extent of biases regarding where they believe they are in the distribution of test-takers at this institute.

Finally, I report regression results from some of the variables described in this section in Table 3.4. Each entry in the table is the coefficient of female in a regression of the outcome (e.g. performance in math) on the indicator for female and controls (where indicated). As previosuly stated, female students perform significantly worse than males in math and text analysis. On average, women respond 0.5 fewer questions correctly in these two areas. This holds after controlling for a measure of intelligence, that is, among equally smart men and women, the women score worse on average in math and text analysis. The performance across other subjects of the test is not significantly different by gender.

The performance results in text analysis is at odds with widespread findings that,

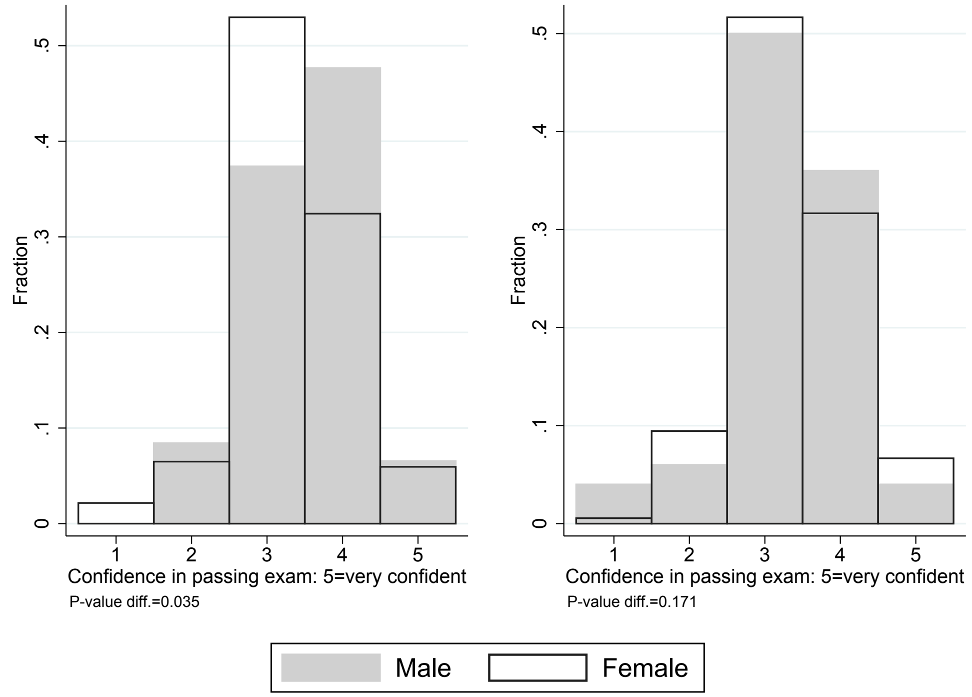
Figure 3.6: Study hours last week for each section of the test



on average, girls perform better in reading than boys. Evidence from the Programme for International Student Assessment (PISA) test administered to 15-year olds reveals that girls have consistently outperformed boys in since 2000 by the equivalent of one year of school (OECD, 2015). In Colombia, however, the average reading scores in the national standardized text (ICFES) from 2000 to 2013 are slightly higher for men. In fact, men outperform women in all subjects of this test except in philosophy. In the PISA 2012 results for Colombia there is a gender gap in reading favoring women but, despite being statistically significant, it is the second smallest gender gap among all countries tested that year. Therefore, the usual advantage of women in reading does not seem as clear in Colombia as in other countries so the underperformance I document is not as surprising as it may seem at first.

Table 3.4 women are less correct than men in their assessment of performance in math, text analysis, and to a less extent, in science. The difference in average

Figure 3.7: Confidence in passing the admissions test



accuracy in the beliefs of men and women in math, science and text analysis is of around 10 percentage points even after controlling for IQ. Women overestimate their performance in text analysis by nearly 15 percentage points relative to men, and underestimate in math by about 9 percentage points. The biases in science are, on average, almost equally split into overestimation and underestimation.

To reiterate, I find no evidence of differential biases by gender in social science and image analysis.

3.4.3 Effects of feedback provision

The results in this section are derived from two beliefs surveys collected from all students after students in the treatment group received feedback. Examples of what feedback looked like are in Appendix 3.6.3.

First, I present evidence of what type of signal men and women in the treatment

Table 3.4: Average performance and biases in self-assessment of performance

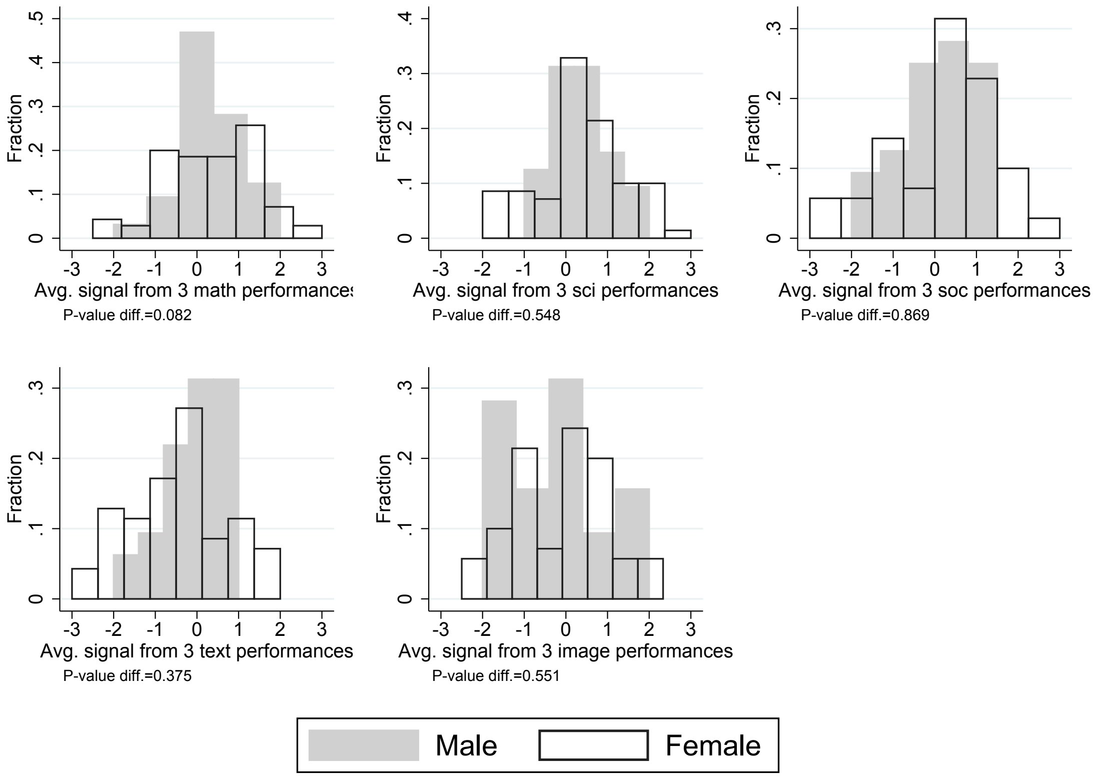
		Math	**Science**	**Social Sci.**	**Text A.**	**Image A.**
Performance	No controls	-0.536***	-0.059	-0.125	-0.555***	-0.045
		(0.137)	(0.128)	(0.123)	(0.195)	(0.121)
	Controls: IQ	-0.493***	-0.048	-0.094	-0.387**	-0.015
		(0.137)	(0.130)	(0.124)	(0.188)	(0.123)
Correct	No controls	-0.117**	-0.103**	-0.056	-0.106**	0.013
		(0.051)	(0.050)	(0.046)	(0.050)	(0.047)
	Controls: IQ	-0.104*	-0.098**	-0.048	-0.091*	0.028
		(0.053)	(0.049)	(0.047)	(0.050)	(0.049)
Overestimated	No controls	0.025	0.048	-0.002	0.151***	-0.052
		(0.038)	(0.043)	(0.044)	(0.051)	(0.055)
	Controls: IQ	0.009	0.052	-0.005	0.124**	-0.045
		(0.038)	(0.043)	(0.045)	(0.052)	(0.056)
Underestimated	No controls	0.092*	0.073	0.061	-0.042	0.041
		(0.049)	(0.048)	(0.048)	(0.048)	(0.050)
	Controls: IQ	0.094*	0.068	0.058	-0.027	0.023
		(0.049)	(0.048)	(0.047)	(0.050)	(0.051)

Notes: Standard errors clustered at the individual level in parentheses. Each entry is the female coefficient in a regression of the outcome (column 1) in the specified section of the test.
*** p<0.01, ** p<0.05, * p<0.1

group saw. Because feedback was given in the form of a graph plotting the performance quartile and the student's prediction of quartile over three practice tests, the signal I compute is the average over three performances. For every practice test shown in the feedback report, I calculate the difference between the actual quartile and the predicted quartile so that a negative difference means that the student overestimated and a positive difference means that the student overestimated. A zero difference means that the student was right in the assessment. After calculating these test-specific differences I average across all three tests shown in the feedback report. Hence, if a student is consistently overestimating, he or she will receive a negative signal or a positive signal if he or she is underestimating. A null signal is slightly more complicated to interpret because it could be result of the students accurately predicting performance or that the individual negative and positive signals cancel each other out.

Figure 3.8 shows that men were more likely to receive a null signal than women presumably because they were more likely to be correct in their assessment of performance. The figure also shows that the distribution of signals for women is more

Figure 3.8: Type of signal received by treatment group



disperse than for men.

Table 3.5 shows the average signal that men and women in the treatment group received (column 1) and the fractions receiving each type of signal in subsequent columns. The stars mean that the p-value of the difference by gender is below the specified significance levels. On average, men and women received positive signals in math, science and social science. As expected, on average, women received a negative signal in text analysis and this is statistically diferent from the average signal men received in the same subject. Both genders received a negative signal in image analysis on average but there is no statistical difference in the strength of this signal by gender.

In terms of proportions, relative to men, women are less likely to receive a null signal in math and more likely to receive a negative signal in math and in text analysis. In the remaining of the analysis the null signal will be split between receiving a null signal because the student correctly predicted the quartile and receiving a null signal that emerges from feedback that is too noisy (overestimating in some tests and underestimating in others).

Table 3.5: Average signal and fractions of men and women seeing each type of signal

			Fraction receiving:		
		Average	**Positive signal**	**Null signal**	**Negative signal**
Math	Male	0.271	0.469	0.375	0.156
	Female	0.284	0.543	0.114***	0.343**
Science	Male	0.359	0.563	0.188	0.25
	Female	0.304	0.486	0.272	0.243
Social science	Male	0.125	0.531	0.156	0.312
	Female	0.140	0.528	0.143	0.328
Text analysis	Male	-0.042	0.344	0.281	0.375
	Female	-0.460**	0.272	0.171	0.557*
Image analysis	Male	-0.292	0.313	0.125	0.563
	Female	-0.084	0.429	0.129	0.443

P-value of the difference in means: *** p<0.01, ** p<0.05, * p<0.1

Tables 3.6 to 3.11 present results for all variables collected from administrative data from the test preparation institute and the beliefs surveys. In each table, the outcome in the table title for each subject of the test is regressed on indicators for type of feedback (signal received), average of the dependent variable before feedback, average performance in that section of the test before feedback, randomization strata fixed effects and unbalanced covariates. The rows show the difference between the treatment and control group where the treatment is split into four mutually-exclusive indicators depending on the type of signal the student received. For example, the indicator for positive signal measures the difference in peformance in each of the subjects indicated by the column titles between students in the treatment group who received a positive signal and students in the control group. The mean of the control group and the number of observations are in the last two rows.

As mentioned earlier, the null signal is split in two to differentiate students who get an average zero signal from correctly predicting their performance from those whose performance and predictions are noisy enough to cancel each other out across the three rounds of practice tests. Standard errors are clustered at the group level because randomization was performed at this level. Stars and daggers show statisical significance from standard hypothesis tests and after correcting from multiple testing within the table, respectively.

The first set of results relates to how feedback affects performance in subsequent practice tests. If students feel more motivated after receiving positive feedback, they may be more confident or put more effort at the moment of taking the test. Similarly, if they receive negative feedback, the opposite effect may take place. Another hypothesis is that positive or negative feedback generates effects in the same direction. For example, research by Gill et al. (2016) reports that feedback increases performance in the lab and that performance improves in a U-shape form with students receiving positive and negative feedback improving the most. The evidence in Table 3.6 shows that in math, most students who receive feedback perform worse relative to students in the control group except those who receive a null signal by correclty predicting their quartile. The changes are not large in general and those that are relatively large, such as the decrease of about one-fifth of a standard deviation among students who receive a positive signal, are not statistically significant due to large standard errors.

The only subjects in which the changes are large enough to be identified as significant are social science and image analysis. In social science students receiving a noisy signal or a negative signal perform worse by 0.78 and 0.37 standard deviations, respectively, than students in the control group. Despite being substantial, after adjusting for multiple testing these changes are not statitically different from zero. The only significant effect in image analysis is among students receiving noisy feedback. They perform 0.77 standad deviations worse than students in the control group and this is significant at the 1 percent level with and without multiple testing adjustment.

Overall, peformance, at least in the short run does not seem to be affected by receiving feedback or by the type of signal received. The only result substantially large in this sample is that students receiving noisy feedback perform worse than students

Table 3.6: Effect of feedback on performance (standardized scores)

	Math	**Science**	**Social science**	**Text analysis**	**Image analysis**
Positive signal	-0.208	-0.047	-0.137	0.026	-0.164
	(0.168)	(0.125)	(0.106)	(0.145)	(0.132)
Null signal (correct)	0.203	0.009	-0.001	0.123	-0.280
	(0.202)	(0.249)	(0.235)	(0.100)	(0.225)
Null signal (noisy)	-0.098	-0.119	-0.794**	0.183	-0.762***†††
	(0.316)	(0.230)	(0.338)	(0.189)	(0.162)
Negative signal	-0.055	0.121	-0.364**	-0.097	-0.147
	(0.194)	(0.174)	(0.162)	(0.125)	(0.192)
Mean of control	0.036	-0.007	0.112	0.016	0.042
Observations	325	325	325	325	325

Notes: Standard errors clustered at the group level in parenthesis. Each regression controls for strata fixed effects, unbalanced covariates, average belief before feedback, and average performance in the corresponding section of the test before feedback.
*** p<0.01, ** p<0.05, * p<0.1
Multiple testing: ††† p<0.01, †† p<0.05, † p<0.01

in the control group particularly in social science and image analysis. Definitely a larger sample is needed to have a definitive conclusion but this finding suggests that providing feedack that may give a mixed message may harm students. Further, this highlights the importance of feedback content and presentation decisions in educational settings.

Feedback can also change future beliefs about relative performance. For example, a student who receives a positive signal and has thus been underestimating his or her performance, may update beliefs in subsequent tests. Table 3.7 presents changes in beliefs relative to the control group after receiving feedback. In this case, a negative value means that the average quartile predictions are in the direction of being in a better quartile (recall quartile 1 is the quartile with the best scores so a negative coefficient means going to smaller numbers). In general we see that people who receive a positive signal update in the correct direction, that is, after controlling for previous performance, students with a positive signal think they belong in a better quartile than students who do not receive feedback. The changes are small and not significant after adjusting for multiple testing.

Table 3.7: Effect of feedback on belief about quartile

	Math	**Science**	**Social science**	**Text analysis**	**Image analysis**
Positive signal	-0.074	-0.231	0.075	-0.092	-0.224
	(0.153)	(0.143)	(0.141)	(0.122)	(0.218)
Null signal (correct)	-0.361*	-0.272	-0.162	-0.002	-0.073
	(0.203)	(0.269)	(0.313)	(0.259)	(0.385)
Null signal (noisy)	0.251	-0.144	0.438	0.446*	-0.340
	(0.299)	(0.203)	(0.368)	(0.225)	(0.280)
Negative signal	-0.038	0.003	-0.059	0.045	0.180
	(0.156)	(0.264)	(0.181)	(0.148)	(0.197)
Mean of control	2.528	2.512	2.389	1.976	1.944
Observations	288	288	287	287	287

Notes: Standard errors clustered at the group level in parenthesis. Each regression controls for strata fixed effects, unbalanced covariates, average performance in the corresponding section of the test and average of dependent variable before feedback.
*** p<0.01, ** p<0.05, * p<0.1
Multiple testing: ††† p<0.01, †† p<0.05, † p<0.01

The sign of the change for students receiving a negative signal is not as clear. In math, science and social science there is some evidence that they update in the wrong direction (think they are in a better quartile) although the magnitudes of these coefficients are small. In text analysis and image analysis, the size of the coefficients is larger and they point in the right direction. The evidence for null signals is mixed and subject dependent. The only coefficient that stands out, although not after correcting for multiple testing, is that students with correct preditions in the past think they are in a better quartile in the following two tests after receiving feedback. This does not necessarily mean that they are wrong about their prediction. For this it is better to look at the variable measureing whether they are correct in their prediction.

A more precise analysis of beliefs is given in Tables 3.8 and 3.9. The first table shows whether treated students are correct in their quartile prediction while the second shows whether they are more likely to underestimate their place in the distribution relative to the control group.[8] The hypothesis is that students in the treatment group are better at assessing their performance because they have seen how their past performances relate to their predictions. Hence, they may be more likely to make a correct estimate relative to the control group. Across all subjects, about 45 percent of students in the control group are right about their quartile predictions. Table 3.8 shows that students receiving feedback are not better at correctly placing their performance. If anything, in some cases like social science and image analysis, receiving feedback makes students less correct in their assessment although this is observed in the case of receiving a null signal.

Table 3.9 shows how much more or less students in the treatment group underestimate their performance. If behavior would be as expected, treated students will be more correct in the previous table and less likely to underestimate their quartile in this table. There are negative signs, especially in math, but the results are far from displaying lower rates of underestimation among treated students. The two highly significant coefficients after correcting for multiple testing have opposite signs. In social science, it seems that students who receive a negative signal are less likely to underestimate. In text analysis, the coefficient for those receiving a noisy signal is actually positive, suggesting that not being able to see a clear pattern may make students more likely to understate their performance.

In the beliefs survey, students also report how difficult they thought each section of the practice test was in a scale from 1 to 7 where 1 is extremely easy and 7 is extremely hard. Depending on the type of feedback, we may see different effects in students' ratings relative to the control group. One possible case is that a positive signal makes the students think that the tests in that area are less hard because they are performing above what they expected. Another possibility is that because they are doing better than they expected they want to convince themselves that the test

[8] The table showing overestimation results is omitted because the results can more or less inferred from the these two tables analyzed together.

Table 3.8: Effect of feedback on holding a correct belief

	Math	**Science**	**Social science**	**Text analysis**	**Image analysis**
Positive signal	0.022	0.004	-0.058	0.036	-0.111
	(0.057)	(0.074)	(0.054)	(0.062)	(0.071)
Null signal (correct)	0.108	-0.097	-0.167	-0.079	-0.357***†
	(0.138)	(0.125)	(0.156)	(0.118)	(0.125)
Null signal (noisy)	0.007	-0.006	-0.298***	-0.196*	0.015
	(0.170)	(0.077)	(0.100)	(0.101)	(0.168)
Negative signal	-0.101	-0.068	0.020	0.005	0.040
	(0.075)	(0.092)	(0.104)	(0.060)	(0.076)
Mean of control	0.457	0.465	0.441	0.425	0.441
Observations	289	289	289	289	289

Notes: Standard errors clustered at the group level in parenthesis. Each regression controls for strata fixed effects, unbalanced covariates, average performance in the corresponding section of the test before feedback, beliefs before feedback, and average of dependent variable before feedback.

*** p<0.01, ** p<0.05, * p<0.1

Multiple testing: ††† p<0.01, †† p<0.05, † p<0.01

is really hard and may rate it as more difficult than students who do not receive feedback.

Table 3.10 shows that, on average, students in the control group rate the different sections between 4.2 and 4.8 or slighly above the middle point in the scale. Overall, students receiving feedback do not rate the exam as harder or easier than students in the control group with one salient exception. Students who were correct in their past assessments in image analysis rate that section of the exam as easier than students in the control group. However, there are very few students receiving this type of feedback so caution must be taking in interpreting this result beyond this small sample. At the end of this section I present evidence that the null results in this outcome are mixing together differential results by gender.

Table 3.9: Effect of feedback on underestimating performance

	Math	**Science**	**Social science**	**Text analysis**	**Image analysis**
Positive signal	-0.030	-0.055	0.027	0.053	-0.003
	(0.071)	(0.061)	(0.070)	(0.096)	(0.076)
Null signal (correct)	-0.192**	0.032	-0.027	0.200*	0.059
	(0.090)	(0.083)	(0.106)	(0.099)	(0.120)
Null signal (noisy)	0.097	-0.102	0.210	0.333***†††	-0.168
	(0.155)	(0.078)	(0.177)	(0.069)	(0.123)
Negative signal	-0.058	0.026	-0.163***††	0.042	0.008
	(0.069)	(0.110)	(0.046)	(0.074)	(0.056)
Mean of control	0.299	0.307	0.317	0.151	0.222
Observations	286	286	285	285	285

Notes: Standard errors clustered at the group level in parenthesis. Each regression controls for strata fixed effects, unbalanced covariates, average performance in the corresponding section of the test before feedback, beliefs before feedback, and average of dependent variable before feedback.
*** p<0.01, ** p<0.05, * p<0.1
Multiple testing: ††† p<0.01, †† p<0.05, † p<0.01

Regarding inputs into the preparation process, students may adjust study time depending on the type of feedback they receive. If they see a negative signal they may study more so that they can perform better next time. Alternatively, if they become discouraged by their performance below their expectations, they may study less. Similar hypotheses could be formulated in the case of positive signals. In the case of signals telling them that they are correct in their assessment, they may adjust study time up or down depending on whether they were right about being at the top or at the bottom of the distribution. For noisy signals, students will most likely not change study time.

Table 3.11 shows that, across subjects, studnets in the control group study more math and text analysis and, on average, students dedicate about 3 hours to study each of the subjects.[9] In most subjects, students receiving a positive signal are study-

[9] As instructed in the survey, study time excludes class time and homework from high school in the subjects evaluated by this test.

Table 3.10: Effect of feedback on perceived difficulty of last practice test

	Math	**Science**	**Social science**	**Text analysis**	**Image analysis**
Positive signal	-0.000	-0.083	-0.215	-0.217	0.224
	(0.241)	(0.209)	(0.262)	(0.243)	(0.323)
Null signal (correct)	-0.301	0.080	0.055	0.812**	-1.194***††
	(0.400)	(0.287)	(0.481)	(0.390)	(0.332)
Null signal (noisy)	0.429	0.067	-0.117	0.531	-0.882**
	(0.280)	(0.289)	(0.441)	(0.391)	(0.337)
Negative signal	-0.427	0.010	-0.022	0.154	0.238
	(0.281)	(0.204)	(0.273)	(0.220)	(0.314)
Mean of control	4.732	4.780	4.693	4.173	4.276
Observations	288	288	288	288	288

Notes: Standard errors clustered at the group level in parenthesis. Each regression controls for strata fixed effects, unbalanced covariates, average performance in the corresponding section of the test before feedback, beliefs before feedback, and average of dependent variable before feedback.

*** p<0.01, ** p<0.05, * p<0.1

Multiple testing: ††† p<0.01, †† p<0.05, † p<0.01

ing less hours for those subjects than the control group. The most clear patterns, although not statistically significant, are for math and text analysis in which students receiving positive feedback study almost 20 minutes less than students in the control group.

It also seems that students receiving a negative signal are studying less than students in the control group at least in math, science and text analysis. In fact, most of the coefficients across the table are negative although they are always below 1 hour of difference relative to the control group and not statistically significant.

One of the focuses of this paper is on gender differences in beliefs and reactions to feedback. I perform analyses in the same spirit of Tables 3.6 to 3.11 with gender indicators but given the small sample sizes it would not be advisable to draw definitive

Table 3.11: Effect of feedback on study hours per subject

	Math	**Science**	**Social science**	**Text analysis**	**Image analysis**
Positive signal	-0.279	0.032	0.038	-0.264	0.010
	(0.273)	(0.290)	(0.349)	(0.395)	(0.338)
Null signal (correct)	-0.574	-0.116	0.740	-0.045	-1.084*
	(0.376)	(0.595)	(0.482)	(0.687)	(0.564)
Null signal (noisy)	-0.843**	-0.224	-0.035	-0.144	0.563
	(0.320)	(0.453)	(0.503)	(0.889)	(0.724)
Negative signal	-0.512	-0.376	0.028	-0.124	0.031
	(0.421)	(0.233)	(0.305)	(0.349)	(0.345)
Mean of control	3.370	3.000	2.898	3.488	2.890
Observations	288	288	288	288	288

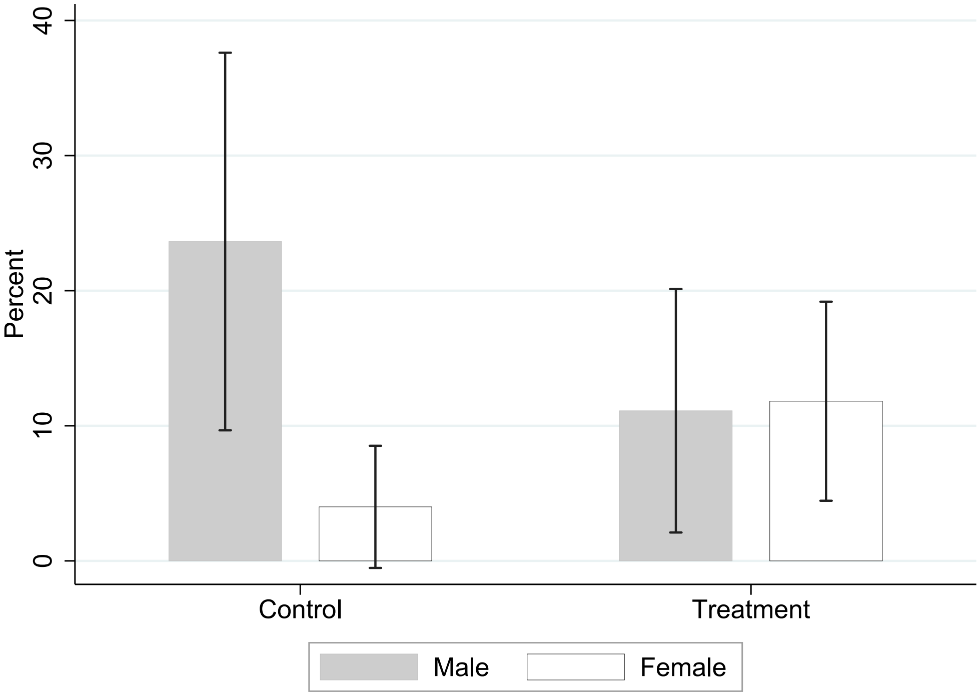
Notes: Standard errors clustered at the group level in parenthesis. Each regression controls for strata fixed effects, unbalanced covariates, average performance in the corresponding section of the test before feedback, beliefs before feedback, and average of dependent variable before feedback.
*** p<0.01, ** p<0.05, * p<0.1
Multiple testing: ††† p<0.01, †† p<0.05, † p<0.01

conclusions. Thus, I restrict my analysis to confience variable that does not need to be split into signal type and to the variable that shows the most systematic pattern across gender.

Figure 3.9 shows the proportions of men and women in the treatment and control groups who report being very confident about gaining admission along with 95 percent confident intervals. This variable is interesting to analyze because it aggregates all signals received across subjects of the test so it is the only variable that measure the effect of feedback at a global level.

About 24 percent of men in the control group report being very confident of passing the exam in their intended major. This is in stark contrast with what women report. Almost no women in the control group feel very confident in passing the

Figure 3.9: Effect of feedback on feeling very confident about gaining admission



exam and the difference of about 20 percentage points is statistically significant. In fact, the confidence interval for the proportion of women in the control group reporting feeling very confidence about passing the admissions exam includes zero. Fewer men who receive feedback are very confident in passing and the difference between these men and those in the control group is significant when controlling for strata fixed effects, unbalanced covariates, and average of dependent variable before feedback. Interestingly, more women in the treatment than in the control group report feeling very confident about gaining admission and this difference is also statisically significant. Overall, in the treatment group, about 12 percent of men and women are very confident that they will pass the admissions exam which suggests that this treatment may contribute to close the gender confidence gap in this particular aspect.

Studies in economics and psychology show that individuals rarely place themselves at the bottom 40 percent of a relative skill distribution (see Burks et al., 2013) and men tend to overestimate more than women (e.g., Burks et al., 2013). The effects shown in Figure 3.9 seem to suggest that this kind of treatment may put men and women closer together in terms of their confidence, in this case about being admitted to this university. Because feeling confident that one is capable of achieving an important goal may provide encouragement and an extra-push during the real exam, I believe this is an important result that may help reduce the gap in test performance and admissions rates by gender.

The only variable that when dissagregating by signal type and gender shows a clear pattern is the ratings of difficulty for each section of the practice test. Table 3.12 presents estimates from regressions on treatment status and gender by signal type. If men receive feedback they are more likely to report that the next practice test is harder than students in the control group and this is independent of signal type. The only exception to this result is in the case of image analysis when the signal says that they were correct in their past assessments. Because women do not respond in terms of perceived difficulty to feedback, their point estimates have the opposite sign and almost equal magnitude than those of men.

Potential interpretations of this result are that when men receive a negative signal, they try to justify a possible repetition of this outcome in the following test by reporting that the test was hard. When they perform better than they expected, they may want to feel good about themselves because they are performing good in a test they rate as hard. When they are correct about their prediction the two explanations could apply depending on whether they were correct about performing well or poorly. This may be related to the literature in psychology that finds that men tend to attribute success to skill and failure to bad luck.

Table 3.12: Effects of treatment on perceived difficulty by gender

	Math	**Science**	**Social science**	**Text analysis**	**Image analysis**
Panel A. Positive signal					
Treated	1.279***†††	0.346	0.325	0.220	0.299
	(0.267)	(0.237)	(0.274)	(0.352)	(0.540)
Female	0.551**†	0.038	0.227	0.292	0.329
	(0.199)	(0.193)	(0.151)	(0.220)	(0.363)
Treated x female	-1.725***†††	-0.697**	-0.827**	-0.760*	-0.233
	(0.322)	(0.287)	(0.360)	(0.433)	(0.633)
Mean of men in control	4.732	4.780	4.693	4.173	4.276
Panel B. Null signal (correct)					
Treated	-0.021	0.630**†	0.145	1.210**	-4.201***††
	(0.437)	(0.244)	(0.447)	(0.576)	(0.982)
Female	0.488**	0.077	0.145	0.249	0.257
	(0.218)	(0.184)	(0.157)	(0.201)	(0.403)
Treated x female	-0.796	-0.783*	-0.467	-0.910	3.868***†
	(0.693)	(0.426)	(0.769)	(0.595)	(1.201)
Mean of men in control	4.732	4.780	4.693	4.173	4.276
Panel C. Null signal (noisy)					
Treated	0.415	0.639	1.230***††	1.059**	0.236
	(0.323)	(0.707)	(0.318)	(0.480)	(0.217)
Female	0.593**†	0.123	0.209	0.272	0.275
	(0.205)	(0.209)	(0.157)	(0.205)	(0.395)
Treated x female	0.025	-0.820	-2.011***††	-1.153	-1.637***††
	(0.411)	(0.753)	(0.512)	(0.809)	(0.400)
Mean of men in control	4.732	4.780	4.693	4.173	4.276
Panel D. Negative signal					
Treated	-0.296	0.772	0.614	0.793**	0.708
	(0.634)	(0.514)	(0.379)	(0.343)	(0.430)
Female	0.524**†	0.037	0.186	0.306	0.282
	(0.192)	(0.187)	(0.171)	(0.212)	(0.346)
Treated x female	-0.476	-0.996*	-0.916**	-0.852**†	-0.762
	(0.520)	(0.519)	(0.384)	(0.336)	(0.459)
Mean of men in control	4.732	4.780	4.693	4.173	4.276

Notes: Standard errors clustered at the group level in parenthesis. Each regression controls for strata fixed effects, unbalanced covariates, and average performance in the corresponding section of the test and average of dependent variable before feedback.
*** p<0.01, ** p<0.05, * p<0.1
Multiple testing: ††† p<0.01, †† p<0.05, † p<0.01

3.5 Conclusion

Laboratory experiments show that, in that contrived environment, individuals overestimate their abilities to perform tasks and that this bias differs by gender. There is however, little evidence of this outside of the lab. Having a wrong perception of one's abilities may induce people to make choices that may not be in their own best interest so correcting biases in real-life situations is of the utmost importance. This paper seeks to fill this gap by analyzing beliefs, the updating process, and choices in a high-stakes context in a developing country.

I partner with a test preparation institution in Colombia to conduct this research with students taking a preparation course to increase their chances of admission at a highly-selective public univerisity. I elicit beliefs about their practice-test performance (in quartiles) in each section of the test after every practice test given by the institute. I randomize students to a treatment group in which they receive feeback about the quartiles to which their performance actually belongs, and to a control group who does not receive feedback.

I analyze responses to beliefs surveys and performance across four rounds of practice tests. I further analyze the type of signals that men and women in the treatment group receive. I find that women hold performance beliefs that are significantly more biased than those of men. In fact, only about 30 percent of women accurately predict their quartile in the distribution across all areas of the test. Men are significantly less biased given that at least 40 percent of them accurately predict their quartile across most areas (almost 50 percent in math). The only area in which men are equally biased than women is in image analysis and in an IQ test.

Overall, the most important and novel finding of this study is that women perfom worse than men in math and text analysis and, at the same time, exhibit larger biases regarding their perfomance in these two subjects. In math, women underestimate their performance meaning that they think they are worse than they actually are. Conversely, in text analysis, even though women also perform worse than men, they think they performed better than they actually did.

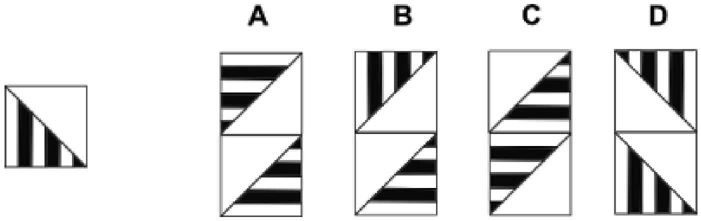
The results from the feedback intervention are less clear. Students who receive feedback do not seem to perform better or have more accurate beliefs in post-feedback practice tests than students in the control group. There is some evidence though that receiving a noisy signal, in the sense that positive and negative signals are mixed together, may hurt students' performance. The two most salient results are that women in the treatment group gain confidence about being admitted to the university they are preparing for although this comes at the expense of a reduction in men's confidence. It seems thus that this type of intervention can reduce the confidence gap in this dimension with important potential implications in terms of encouragement and assetiveness at the moment of taking the real admissions exam. Further, men who receive feedback rate their next practice test as harder than students in the control group irrespective of the type of signal received. This could be interpreted within theories of self-serving bias that suggest individuals attribute success to skill and failure to bad luck to keep self-esteem high. In the context of this study and under this theory, men who receive a positive signal want to think they are succeding in hard tests whereas men receiving a negative signal may blame their performance on having bad luck in getting a hard test.

Future work will focus on obtaining more precise estimations by increasing the sample size and analyzing outcomes related to performance in the actual admissions exam, and college major choices.

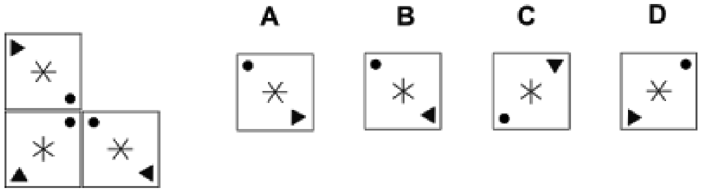
3.6 Appendices

3.6.1 Example of image analysis questions

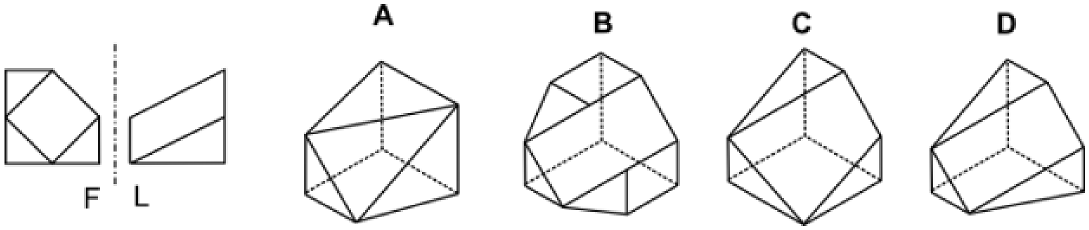
Identify the design that cannot be built with the figure on the left:



Identify the missing piece in the figure on the left:



Identify the figure that contains the frontal (F) and lateral (L) views of the figure on the left:



3.6.2 Beliefs questionnaire

Suppose we organize the scores in the practice exam of the participants in this study from highest to lowest. Suppose we divide the scores in 4 equal groups called quartiles where:

- Quartile 1 contains the 25% of participants with the highest scores
- Quartile 2 contains the 25% of participants with scores below quartile 1
- Quartile 3 contains the 25% of participants with scores below quartile 2

- Quartile 4 which contains the 25% of participants with the lowest scores.

Question 1
What is, according to you, the probability in % that your score in the practice test you just took belongs to each of the 4 quartiles? Make sure your answers are between 0 and 100. For each of your four answers, the computer will randomly pick a number between 0 and 100. If the probability in your answer is larger than that number, you will receive 1 lottery ticket if your score is in that quartile. If the probability in your answer is smaller than that number, you will win 1 lottery ticket with a probability equal to the number chosen by the computer. To maximize your chances of winning lottery tickets you should try to approximate to the best of your ability the probability that your score will be in each quartile.

Please write the probabilities in the following table and verify that the sum of all probabilities equals 100%. If you think that it is unlikely that your score will be in a given quartile, please with 0%.

Probability that my score is in quartile 1	%
Probability that my score is in quartile 2	%
Probability that my score is in quartile 3	%
Probability that my score is in quartile 4	%
Sum of 4 rows	%

Question 2
Please rate the level of difficulty of each section of the practice test you just took from 1 (very easy) to 7 (extremely hard):

Mathematics	1	2	3	4	5	6	7
Natural sciences	1	2	3	4	5	6	7
Social sciences	1	2	3	4	5	6	7
Text analysis	1	2	3	4	5	6	7
Image analysis	1	2	3	4	5	6	7

Question 3
Approximately, how many hours did you spend studying for each section of the exam during last week? Include time reviewing contents and working on practice questions.

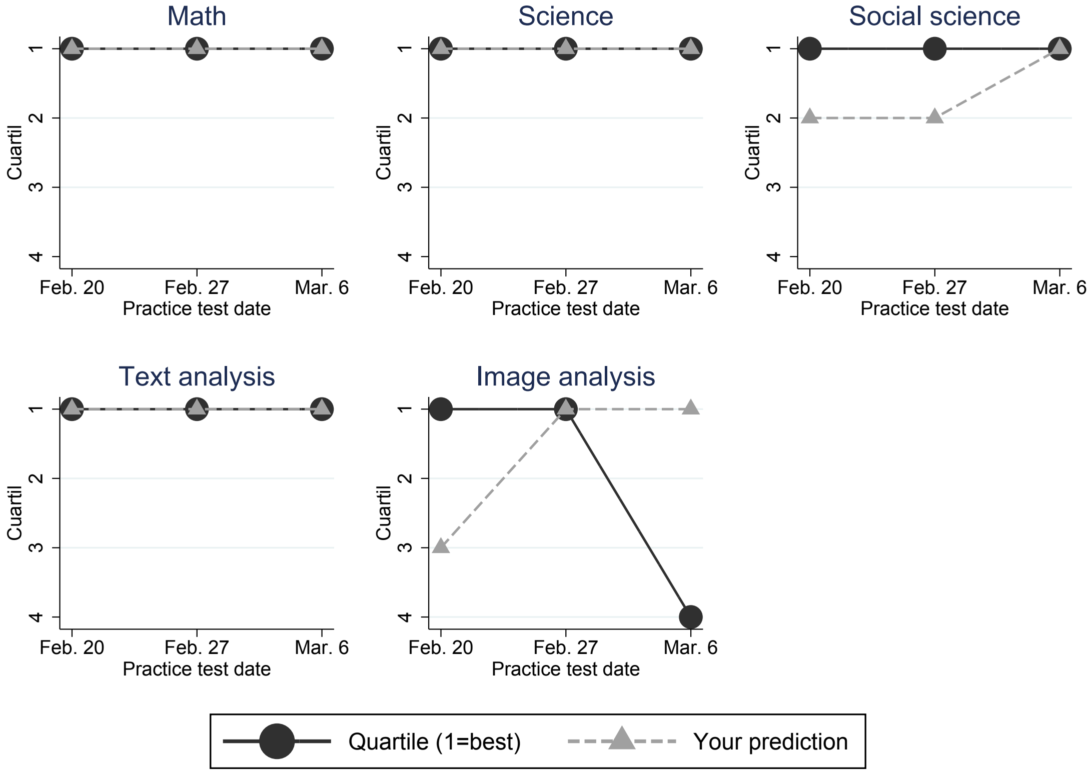
Exclude class time and time spent doing homework. Incluye el tiempo que dedicaste a resolver preguntas de práctica y revisar contenidos. Circle cero if you didn't study for a specific section of the test last week.

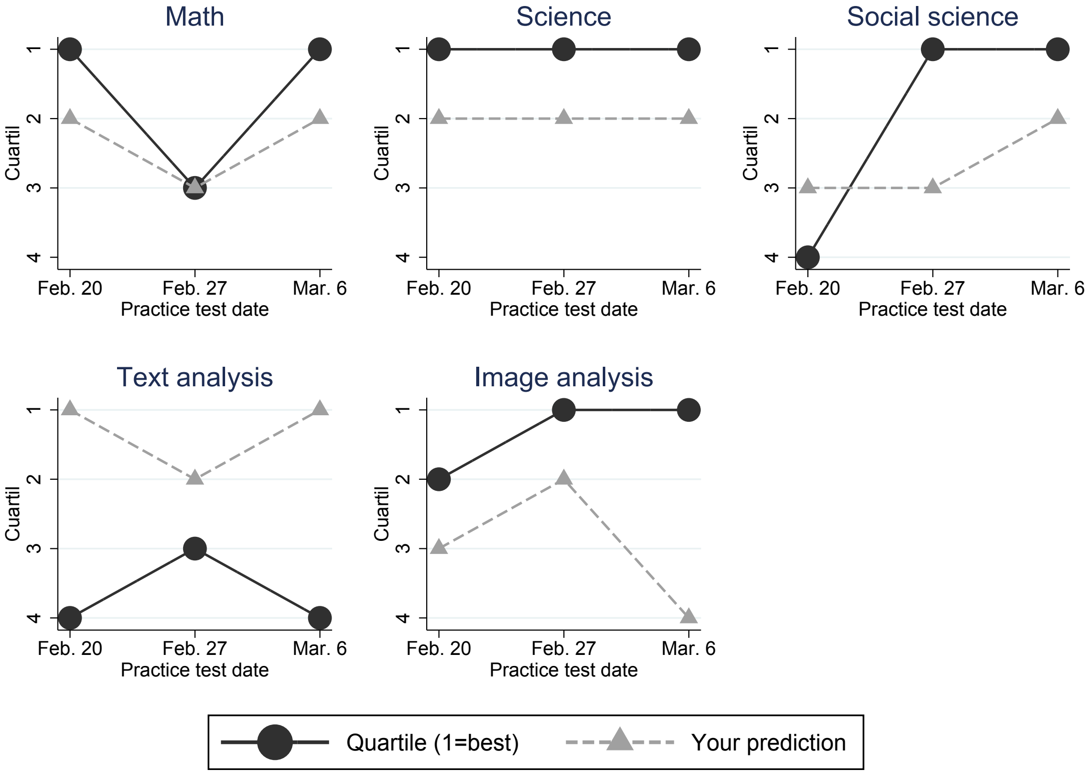
Mathematics	0	1	2	3	4	5	6	7	8+
Natural sciences	0	1	2	3	4	5	6	7	8+
Social sciences	0	1	2	3	4	5	6	7	8+
Text analysis	0	1	2	3	4	5	6	7	8+
Image analysis	0	1	2	3	4	5	6	7	8+

Question 4

How confident are you that you will pass the admissions exam on April 23, 2017? (5 smiley faces to choose from)

3.6.3 Feedback examples




Math
Science
Social science
Text analysis
Image analysis
Cuartil
1
2
3
4
Feb. 20
Feb. 27
Mar. 6
Practice test date
Quartile (1=best)
Your prediction

BIBLIOGRAPHY

BIBLIOGRAPHY

Abrams, D. (2009). Building criminal capital vs. specific deterrence: The effect of incarceration length on recidivism.

Aizer, A., & Doyle, J. J. (2015). Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges. *The Quarterly Journal of Economics*, qjv003.

Alexander, M. (2012). *The new Jim Crow: Mass incarceration in the age of colorblindness.* The New Press.

Andreoni, J., & Sprenger, C. (2012). Estimating time preferences from convex budgets. *The American Economic Review*, *102*(7), 3333–3356.

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.

Austin, J., & Krisberg, B. (1981). Nccd research review: Wider, stronger, and different nets: The dialectics of criminal justice reform. *Journal of research in crime and delinquency*, *18*(1), 165–196.

Barreca, A. I., Lindo, J. M., & Waddell, G. R. (2016). Heaping-induced bias in regression-discontinuity designs. *Economic Inquiry*, *54*(1), 268–293.

Beauchamp, J. P., Benjamin, D. J., Chabris, C. F., & Laibson, D. I. (2012). *How malleable are risk preferences and loss aversion* (Tech. Rep.). Harvard University Mimeo.

Benjamin, D. J., Brown, S. A., & Shapiro, J. M. (2013). Who is 'behavioral'? cognitive ability and anomalous preferences. *Journal of the European Economic Association*, *11*(6), 1231–1255.

Berlin, N., & Dargnies, M.-P. (2016). Gender differences in reactions to feedback and willingness to compete. *Journal of Economic Behavior & Organization*, *130*, 320–336.

Berube, D. A., & Green, D. P. (2007). The effects of sentencing on recidivism: Results from a natural experiment. *Second Annual Conference on Empirical Legal Studies, New York.*

Bobba, M., & Frisancho, V. (2016). Learning about oneself: The effects of signaling ability on school choice. *Inter-Am. Dev. Bank, Discuss. Pap*, *450*.

Burks, S. V., Carpenter, J. P., Goette, L., & Rustichini, A. (2013). Overconfidence and social signalling. *The Review of economic studies*, *80*(3), 949–983.

Bushway, S. D., & Paternoster, R. (2009). Do prisons make us safer? the benefits and costs of the prison boom. In S. Raphael & M. A. Stoll (Eds.), *Do prisons make us safer? the benefits and costs of the prison boom.* Russell Sage Foundation.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, *82*(6), 2295–2326.

Camerer, C., & Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of risk and uncertainty*, *5*(4), 325–370.

Carson, E. A., & Golinelli, D. (2013). Prisoners in 2012: Trends in admissions and releases, 1991–2012. *Washington DC: Bureau of Justice Statistics*.

Carvalho, L. S., Meier, S., & Wang, S. W. (2016). Poverty and economic decision-making: Evidence from changes in financial resources at payday. *The American economic review*, *106*(2), 260–284.

Cattaneo, M. D., Frandsen, B. R., & Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the us senate. *Journal of Causal Inference*, *3*(1), 1–24.

Charness, G., Gneezy, U., & Imas, A. (2013). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior and Organization*, *87*, 43–51.

Choi, S., Kariv, S., Müller, W., & Silverman, D. (2014). Who is (more) rational? *The American Economic Review*, *104*(6), 1518–1550.

Chuang, Y., & Schechter, L. (2015). Stability of experimental and survey measures of risk, time, and social preferences: A review and some new results. *Journal of Development Economics*, *117*, 151–170.

Cullen, F. T. (2005). The twelve people who saved rehabilitation: How the science of criminology made a difference. *Criminology*, *43*(1), 1–42.

Cullen, F. T., & Jonson, C. L. (2011). Rehabilitation and treatment programs. In J. Q. Wilson & J. Petersilia (Eds.), *Crime and public policy* (pp. 293–344). Oxford University Press New York, NY.

Deming, S. R. (2000). Michigan's sentencing guidelines. *Michigan Bar Journal*, *79*(6), 652–655.

Dinkelman, T., & Martínez, C. (2014). Investing in schooling in chile: The role of information about financial aid for higher education. *Review of Economics and Statistics*, *96*(2), 244–257.

Eckel, C. C., & Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and human behavior*, *23*(4), 281–295.

Eckel, C. C., & Grossman, P. J. (2008). Men, women and risk aversion: Experimental evidence. *Handbook of experimental economics results*, *1*, 1061–1073.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The quarterly journal of economics*, 643–669.

Frandsen, B. R. (2014). Party bias in union representation elections: Testing for manipulation in the regression discontinuity design when the running variable is discrete. *Manuscript, Brigham Young University, Department of Economics.*

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, *19*(4), 25–42.

Gelman, A., & Imbens, G. (2014). *Why high-order polynomials should not be used in regression discontinuity designs* (Tech. Rep.). National Bureau of Economic Research.

Gill, D., Kissová, Z., Lee, J., & Prowse, V. L. (2016). First-place loving and last-place loathing: How rank in the distribution of performance affects effort provision.

Giné, X., Goldberg, J., Silverman, D., & Yang, D. (2017). Revising commitments: Field evidence on the adjustment of prior choices. *The Economic Journal.*

Gonzalez, N. (2017). *How learning about one's ability affects educational investments: Evidence from the advanced placement program* (Tech. Rep.). Mathematica Policy Research.

Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, *78*(6), 1360–1380.

Green, D. P., & Winik, D. (2010). Using random judge assignments to estimate the effects of incarceration and probation on recidivism among drug offenders. *Criminology*, *48*(2), 357–387.

Gustman, A. L., & Stafford, F. P. (1972). Income expectations and the consumption of graduate students. *Journal of Political Economy*, *80*(6), 1246–1258.

Hastings, J. S., & Weinstein, J. M. (2008). Information, school choice, and academic achievement: Evidence from two experiments. *The Quarterly journal of economics*, *123*(4), 1373–1414.

Hjalmarsson, R. (2009). Juvenile jails: A path to the straight and narrow or to hardened criminality? *The Journal of Law and Economics*, *52*(4), 779–809.

Hoxby, C. M., & Turner, S. (2015). What high-achieving low-income students know about college. *The American Economic Review*, *105*(5), 514–517.

Hughes, T. A., & Wilson, D. J. (2003). *Reentry trends in the united states.* US Department of Justice, Bureau of Justice Statistics Washington, DC.

Imbens, G., & Kalyanaraman, K. (2011). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of economic studies*, rdr043.

Jaman, D. R., Dickover, R. M., & Bennett, L. A. (1972). Parole outcome as a function of time served. *The British journal of criminology*, *12*(1), 5–34.

Jensen, R. (2010). The (perceived) returns to education and the demand for schooling. *The Quarterly Journal of Economics*, *125*(2), 515–548.

Kling, J. R. (2006). Incarceration length, employment, and earnings. *The American economic review*, *96*(3), 863–876.

Kolesár, M., & Rothe, C. (2016). Inference in regression discontinuity designs with a discrete running variable. *Unpublished Working Paper*.

Krupka, E. L., & Stephens, M. (2013). The stability of measured time preferences.

Journal of Economic Behavior & Organization, *85*, 11–19.

Kuziemko, I. (2013). How should inmates be released from prison? an assessment of parole versus fixed-sentence regimes. *The Quarterly Journal of Economics*, *128*(1), 371–424.

Lee, D. S., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, *142*(2), 655–674.

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

Loeffler, C. E. (2013). Does imprisonment alter the life course? evidence on crime and employment from a natural experiment. *Criminology*, *51*(1), 137–166.

Malmendier, U., & Nagel, S. (2011). Depression babies: Do macroeconomic experiences affect risk-taking? *The Quarterly Journal of Economics*, *126*(1), 373-416.

Mani, A., Mullainathan, S., Shafir, E., & Zhao, J. (2013). Poverty impedes cognitive function. *science*, *341*(6149), 976–980.

Matthey, A., & Regner, T. (2013). On the independence of history: experience spill-overs between experiments. *Theory and decision*, *75*(3), 403–419.

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

Michigan Judicial Institute. (2016). *General information and instructions for using the statutory sentencing guidelines.* `https://mjieducation.mi.gov/documents/felony-sentencing-resources/67-2005-sgm/file`.

Miles, T. J., & Ludwig, J. (2007). The silence of the lambdas: deterring incapacitation research. *Journal of Quantitative criminology*, *23*(4), 287–301.

Mizala, A., & Urquiola, M. (2013). School markets: The impact of information approximating schools' effectiveness. *Journal of Development Economics*, *103*, 313–335.

Mobius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2011). *Managing self-confidence: Theory and experimental evidence* (Tech. Rep.). National Bureau of Economic Research.

Mueller-Smith, M. (2016). The criminal and labor market impacts of incarceration. *Unpublished Working Paper*.

Mueller-Smith, M., & Schnepel, K. T. (2016). Punishment and (non-) deterrence: Evidence on first-time drug o enders from regression discontinuities. *Unpublished Working Paper*.

Mullainathan, S., & Shafir, E. (2013). *Scarcity: Why having too little means so much.* Macmillan.

Nagin, D. S., Cullen, F. T., & Jonson, C. L. (2009). Imprisonment and reoffending. *Crime and justice*, *38*(1), 115–200.

Nagin, D. S., & Snodgrass, G. M. (2013). The effect of incarceration on re-offending: Evidence from a natural experiment in pennsylvania. *Journal of Quantitative Criminology*, *29*(4), 601–642.

National Research Council. (2008). *Parole, desistance from crime, and community integration.* National Academies Press.

Nguyen, Q. (2011). Does nurture matter: theory and experimental investigation on the effect of working environment on risk and time preferences. *Journal of Risk and Uncertainty*, *43*(3), 245–270.

Niederle, M., Segal, C., & Vesterlund, L. (2013). How costly is diversity? affirmative action in light of gender differences in competitiveness. *Management Science*, *59*(1), 1–16.

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, *122*(3), 1067–1101.

OECD. (2015). *The abc of gender equality in education: Aptitude, behaviour, confidence.* `http://dx.doi.org/10.1787/9789264229945-en`.

Pager, D. (2008). *Marked: Race, crime, and finding work in an era of mass incarceration.* University of Chicago Press.

Palumbo, D. J., Clifford, M., & Snyder-Joy, Z. K. (1992). From net widening to intermediate sanctions: The transformation of alternatives to incarceration from benevolence to malevolence. *Smart sentencing: The emergence of intermediate sanctions*, 229–244.

Petersilia, J. (2011). Community corrections: Probation, parole, and prisoner reentry. *Crime and public policy*, 499–531.

Raphael, S., & Stoll, M. A. (2009). Why are so many americans in prison? *Do prisons make us safer*, 27–72.

Rees, A. (1966). Information networks in labor markets. *The American Economic Review*, *56*(1/2), 559–566.

Reuben, E., Wiswall, M., & Zafar, B. (2015). Preferences and biases in educational choices and labour market expectations: Shrinking the black box of gender. *The Economic Journal.*

Sampson, R. J., & Laub, J. H. (1995). *Crime in the making: Pathways and turning points through life.* Harvard University Press.

Schmitt, J., Warner, K., Gupta, S., et al. (2010). The high budgetary cost of incarceration. *Washington, DC: Center for Economic and Policy Research.*

Shah, A. K., Mullainathan, S., & Shafir, E. (2012). Some consequences of having too little. *Science*, *338*(6107), 682–685.

Shapiro, M. D., & Slemrod, J. (1995). Consumer response to the timing of income: Evidence from a change in tax withholding. *The American Economic Review*, *85*(1), 274–283.

Sinayev, A., & Peters, E. (2015). Cognitive reflection vs. calculation in decision making. *Frontiers in psychology*, *6*, 532.

Stephens, M. (2003). " 3rd of tha month": Do social security recipients smooth consumption between checks? *The American Economic Review*, *93*(1), 406–422.

Tanaka, T., Camerer, C. F., & Nguyen, Q. (2010). Risk and time preferences: linking experimental and household survey data from vietnam. *The American Economic Review*, *100*(1), 557–571.

Tanaka, Y., Fujino, J., Ideno, T., Okubo, S., Takemura, K., Miyata, J., … others (2014). Are ambiguity aversion and ambiguity intolerance identical? a neuroeconomics investigation. *Frontiers in psychology*, *5*.

Travis, J. (2005). *But they all come back: Facing the challenges of prisoner reentry.* The Urban Insitute.

Tyler, J. H., Kling, J. R., et al. (2007). Prison-based education and reentry into the mainstream labor market." in barriers to reentry?: The labor market for released prisoners.

Van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression–discontinuity approach. *International Economic Review*, *43*(4), 1249–1287.

West, H. C., Sabol, W. J., & Greenman, S. J. (2010). Prisoners in 2009 (ncj 231675). *Washington, DC: Bureau of Justice Statistics, US Department of Justice.*

Western, B. (2006). *Punishment and inequality in America.* Russell Sage Foundation.

Wiswall, M., & Zafar, B. (2015). Determinants of college major choice: Identification using an information experiment. *The Review of Economic Studies*, *82*(2), 791–824.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.* MIT press.

Wozniak, D., Harbaugh, W. T., & Mayr, U. (2014). The menstrual cycle and performance feedback alter gender differences in competitive choices. *Journal of Labor Economics*, *32*(1), 161–198.