LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

ESSAYS ON CONDITIONAL CASH TRANSFERS, TARGETING AND EDUCATIONAL OUTCOMES: EVIDENCE FROM CHILE

CRISTIAN EDUARDO CRESPO ROJAS

A thesis submitted to the Department of Social Policy of the London School of Economics and Political Science for the Degree of Doctor of Philosophy,

London, April 2019

Declaration of Authorship

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent.

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 69,825 words.

Declaration of editorial help

I can confirm that my thesis was copy edited for conventions of language, spelling and grammar by Clare Sandford.

Cristian Eduardo Crespo Rojas
London, April 2019

Acknowledgements

When people ask me how it feels to pursue a PhD, I answer that it is like running a marathon. It is not a good idea to sprint or to stop running. To be honest, I have never run a marathon. I guess you need proper training and stimulation to succeed. Completing this PhD Thesis would not have been possible without the support and encouragement of many people along the way.

In the first place I would like to thank my supervisors. I am especially grateful to Stephen for his availability and dedication, from helping with basic editing to facilitating finding the path forward when crucial decisions needed to be made. It is tough to figure out how his supervision could have been any better. I am also indebted to Grace for her advice, especially for highlighting *ex-ante* every path that could have led to uncertain outcomes. In retrospect, I realise that without their guidance my PhD experience would have been very challenging.

I would like to thank the Chilean Ministry of Social Development and the Ministry of Education. My research would not have been possible without their work assembling datasets. Additionally, I want to express my gratitude to each colleague that has provided insightful comments. During the last three years the Social Policy Department has become my second home. It has been an extraordinary environment in which to carry out my PhD Thesis. I am grateful to the professors and administrative staff for their support. I will miss the invaluable learning that such a multidisciplinary platform provides.

Many friends made this journey a delightful one. I was blessed to belong to a bright and very social cohort of PhD students within my Department. I was fortunate to develop meaningful relationships in each of the places I lived. I am grateful that I reconnected in London with old friends from Chile. Each of these special people helped me to achieve the right balance between work and joy, career and personal development. My gratitude to you all.

Finally, and more importantly, I want to thank my wife Carolina. From the very beginning we embraced this project together. We chose to come to London because it is a city where we could both thrive. Time will tell but I think we made the right decision. Surely nothing of this would have been possible without our mutual support, understanding and love for each other.

Abstract

This thesis studies key questions located at the intersection of conditional cash transfers (CCTs), targeting mechanisms of CCTs, especially proxy means tests (PMTs) and outcomes in primary and secondary education. The research relies entirely on large and rich administrative datasets from Chile. The thesis is built around three empirical chapters or papers.

The first paper contributes to the social policy targeting field. The chapter analyses whether a PMT can identify the poor and future school dropouts effectively. Despite both being key target groups for CCTs, students at risk of dropping out are rarely considered for CCT allocation and in targeting assessments. Using simulations, I compare the PMT with other mechanisms based on a predictive model of school dropout. I build this model using machine learning algorithms, one of their first applications in regard to school dropout outside a developed nation. Using the outputs of the predictive model in conjunction with the PMT increases targeting effectiveness except when the social valuation of the poor and future school dropouts differs to a large extent.

The second paper analyses whether it is convenient to reward children for their academic performance. The chapter estimates the impact of a cash for grades programme on future attendance and average grade using a regression discontinuity (RD) design. The main causal estimates for the outcomes are not statistically significantly different from zero.

The third paper contributes to the causal inference literature, particularly about RD designs. Despite the rapid development of the RD methodological literature, some threats to internal validity have been overlooked. The chapter elaborates on two threats, administrative sorting and intermediate contamination, in the context of three impact evaluations of CCTs.

This thesis contributes to advancing knowledge both methodologically and for policy. Although the study focuses its analysis on one country, its results have implications for multiple contexts.

Table of Contents

Chapter 1 Introduction 1

1.1 Conditional Cash Transfers: Components, Goals and Impact 1

1.2 Targeting of Conditional Cash Transfers and Proxy Means Tests 6

1.3 Why Study Conditional Cash Transfers? 8

1.4 The Primary and Secondary Education System in Chile 10

1.5 Thesis Outline 14

Chapter 2 Two Become One: Improving the Targeting of Conditional Cash Transfers With a Predictive Model of School Dropout 16

2.1 Introduction 16

2.2 Data and Methods 21

2.3 Results: Predictive Model of School Dropout 36

2.4 Results: Targeting Assessment 48

2.5 Conclusion 56

Appendix A. Predictors Included in Machine Learning Algorithms 61

Appendix B. Indicators Derived from a Classifier of School Dropout 62

Appendix C. Out of Time Validation for the Predictive Model 63

Appendix D. Summary Statistics for Other Measures of School Dropout 65

Appendix E. Sensitivity Analysis for Total Leakage Targeting Assessment 66

Appendix F. Targeting Assessment Including Administrative Costs 69

Chapter 3 Cash for Grades or Money for Nothing? Evidence from Regression Discontinuity Designs 71

3.1 Introduction 71

3.2 Programme Description 74

3.3 Data and Methods 76

3.4 Results 89

3.5 Conclusion 110

Appendix G. Summary of Cash for Grades Evaluations in Schools 113

Appendix H. Using Relative Ranking in a Regression Discontinuity Design 115

Appendix I. Local Randomisation Regression Discontinuity Framework 120

Appendix J. Effects of Receiving the BLE in 2014 on Educational Outcomes of 2015 ... 125
Appendix K. Effects of Receiving the BLE in 2013 for Population Subgroups ... 126

Chapter 4 (Don't) Call me by Your Name: Reassessing Threats to Internal Validity in Regression Discontinuity Designs ... 130
4.1 Introduction ... 130
4.2 Case 1: *Beca de Apoyo a la Retención Escolar* ... 136
4.3 Case 2: *Subsidio Familiar* & *Asignación Social* ... 159
4.4 Conclusion ... 174

Chapter 5 Conclusion ... 178
5.1 Main Findings and Contributions ... 178
5.2 Policy and Methodological Implications ... 181
5.3 Further Research ... 185
5.4 Concluding Remarks ... 187

References ... 189

List of Tables

Table 1.1: Grade Length and Age Requirements in Primary Education 12

Table 1.2: Grade Length and Age Requirements in Secondary Education 13

Table 2.1: Confusion Matrix or Contingency Table of a Classifier of Student Dropout.........28

Table 2.2: Indicators Used in the Predictive Model of School Dropout.................................29

Table 2.3: Indicators Used in the Targeting Assessment...31

Table 2.4: Successful Targeting and Targeting Errors in a Context of Two Target Groups...32

Table 2.5: Social Valuation (Welfare Weights γ) in Different Scenarios...............................34

Table 2.6: Dataset Structure...36

Table 2.7: Summary Statistics 39

Table 2.8: Area Under the ROC Curve for Models Predicting School Dropout 42

Table 2.9: True Positive Rate (Sensitivity), False Positive Rate (1–Specificity) and Accuracy for Models Predicting School Dropout 46

Table 2.10: Mean and Relative Frequency of Poor and School Dropout by Quintile Groups 49

Table 2.11: Targeting Indicators by Independent Approach and Budget Available 51

Table 2.12: Targeting Indicators by Combined Approach and Budget Available 52

Table 2.13: Total Leakage by Targeting Mechanism and Budget Available 53

Table 2.14: Leaked Welfare by Social Valuation of Target Groups 55

Table 2.15: Predictors of School Dropout Included in Machine Learning Algorithms...........61

Table 2.16: Examples of Indicators Derived from a Classifier of School Dropout.................62

Table 2.17: Area Under the ROC Curve for Models Predicting School Dropout 63

Table 2.18: True Positive Rate (Sensitivity), False Positive Rate (1–Specificity) and Accuracy for Models Predicting School Dropout 64

Table 2.19: Mean Values and Relative Frequency for School Dropout in Years t+1 and t+2 65

Table 2.20: Total Leakage: Sensitivity Analysis by Poverty Line and Income Definition 66

Table 2.21: Total Leakage: Sensitivity Analysis by Measure of School Dropout..................66

Table 2.22: Total Leakage: Sensitivity Analysis Using a Composite Score 68

Table 2.23: Total Leakage: Sensitivity Analysis by Machine Learning Algorithm................68

Table 3.1: Dataset Structure by BLE Year-Cohorts 78

Table 3.2: Descriptive Statistics for the BLE in 2013 81

Table 3.3: RD Estimates for Outcomes (Using Mean Squared Error Optimal Bandwidth)....93

Table 3.4: RD Estimates for Future Average Grade (Sensitivity Analysis to Bandwidth) 94

Table 3.5: RD Estimates for Future Attendance (Sensitivity Analysis to Bandwidth) 94

Table 3.6: RD Estimates for Pre-Treatment Variables (Using Mean Squared Error Optimal Bandwidth)...95

Table 3.7: RD Estimates for Outcomes 102

Table 3.8: RD Estimates for Future Average Grade (Sensitivity Analysis to Bandwidth) ...103

Table 3.9: RD Estimates for Future Attendance (Sensitivity Analysis to Bandwidth) 103

Table 3.10: RD Estimates for Pre-Treatment Variables 104

Table 3.11: Summary of Cash for Grades Evaluations (Randomised Control Trials) in Primary and Secondary Education..113

Table 3.12: RD Estimates for Pre-Treatment Variables in Local Randomisation Framework 117

Table 3.13: RD Estimates for Pre-Treatment Variables in Continuity-Based Framework (Using Mean Squared Error Optimal Bandwidth)..118

Table 3.14: Theoretical Academic Cohort Size by Intervals of Relative Ranking...............119

Table 3.15: Summary Statistics (Mean Values) by Intervals of Relative Ranking 119

Table 3.16: RD Estimates for Outcomes in Local Randomisation Framework 122

Table 3.17: RD Estimates for Pre-Treatment Variables in Local Randomisation Framework 123

Table 3.18: RD Estimates for Outcomes in Local Randomisation Framework (*w*=0.1).......124

Table 3.19: RD Estimates for Pre-Treatment Variables in Local Randomisation Framework (*w*=0.1)...124

Table 3.20: RD Estimates for Outcomes in 2015 125

Table 3.21: RD Estimates for Outcomes in 2014 by Halves of Lowest PFSE Scores in Continuity-Based Framework (Using Average Grade in 2012 as a Running Variable)........126

Table 3.22: RD Estimates for Average Grade in 2014 (by Subgroups) 129

Table 3.23: RD Estimates for Attendance in 2014 (by Subgroups) 129

Table 4.1: IVSE Variables 138

Table 4.2: Descriptive Statistics (Mean Values) by BARE Cohorts and IVSE scores..........140

Table 4.3: Mean Values by IVSE Scores and RD Estimates for Pre-Treatment Variables in Local Randomisation Framework...144

Table 4.4: RD Estimates for Pre-Treatment Variables in Continuity-Based Framework (Observed IVSE as a Running Variable)..145

Table 4.5: Three Types of Adolescents With IVSE Scores Close to the BARE Threshold..146

Table 4.6: Functions $f_k(\cdot)$ Transforming IVSE Variables into Scores.................................147

Table 4.7: Functions $f_k'(\cdot)$ Transforming IVSE Variables into Scores....................................148

Table 4.8: RD Estimates for Pre-Treatment Variables in Continuity-Based Framework (Hypothetical IVSE as a Running Variable)..150

Table 4.9: Definition of $\Delta days$ by Categorical Boundaries of Δage and Day of Birth 154

Table 4.10: RD Local Randomisation Estimates for Pre-Treatment Variables.....................158

Table 4.11: RD Local Randomisation Estimates for Age...158

Table 4.12: Descriptive Statistics (Mean Values) by Conditional Cash Transfer With Its Respective SPF Threshold 164

Table 4.13: Continuity-Based RD Estimates for *Chile Solidario* Variables 173

Table 4.14: Summary of CCTs, Running Variables, Thresholds and Threats to Validity 174

List of Figures

Figure 2.1: ROC Curve for Models Predicting School Dropout in Year *t*+1 or *t*+2................41

Figure 2.2: ROC Curve for Models Predicting School Dropout in Year *t*+1 45

Figure 2.3: ROC Curve for Models Predicting School Dropout in Year *t*+2 45

Figure 2.4: Variable Importance for glmnet Algorithm...47

Figure 2.5: Variable Importance for gbm Algorithm...47

Figure 3.1: PFSE Score Density and Frandsen Manipulation Test 91

Figure 3.2: Future Outcomes by PFSE Score in 2012 (Students in the Highest 30% of Academic Performance in 2012)...96

Figure 3.3: Pre-Treatment Variables by PFSE Score in 2012 (Students in the Highest 30% of Academic Performance in 2012) 97

Figure 3.4: Distance of Average Grade to the Threshold Density and Frandsen Manipulation Test (Students With PFSE ≤ 98 in 2012)...101

Figure 3.5: Future Outcomes by Distance of Average Grade to the Threshold in 2012 (Students With PFSE ≤ 98 in 2012)...105

Figure 3.6: Pre-Treatment Variables by Distance of Average Grade to the Threshold in 2012 (Students With PFSE ≤ 98 in 2012)..106

Figure 3.7: Summary of RD Point Estimates and Confidence Intervals for Average Grade 108

Figure 3.8: Summary of RD Point Estimates and Confidence Intervals for Attendance 109

Figure 4.1: BARE Recipient by IVSE Scores (by Different Years).....................................141

Figure 4.2: Observed IVSE Density and Hypothetical IVSE Density (Year 2014).............148

Figure 4.3: Hypothetical IVSE Density & McCrary Test (Year 2014) 149

Figure 4.4: Pre-Treatment Variables by Hypothetical IVSE Scores 151

Figure 4.5: Distribution of $\Delta days$ (by Reference Day of Birth) 155

Figure 4.6: *Asignación Social* and *Subsidio Familiar* Recipient by SPF Scores 165

Figure 4.7: SPF Scores Distribution for *Asignación Social* and *Subsidio Familiar*..............166

Figure 4.8: McCrary Test for *Asignación Social* and *Subsidio Familiar* 167

Figure 4.9: Social Protection File Scores Distribution 168

Figure 4.10: One Hundred Bins of Income per Capita in CASEN and the SPF....................170

Figure 4.11: Income per Capita Distribution (SPF Prediction and CASEN) 171

Figure 4.12: SPF Predicted Income per Capita Distribution and McCrary Test...................171

Figure 4.13: *Chile Solidario* Variables by SPF Predicted Income per Capita ($CLP)..........172

List of Acronyms

AS	*Asignación Social*
AUC	Area Under the Curve
BARE	*Beca de Apoyo a la Retención Escolar*
BLE	*Bono por Logro Escolar*
CASEN	*Encuesta de Caracterización Socioeconómica de Hogares*
CCTs	Conditional Cash Transfers
CLP	Chilean Pesos
CS	*Chile Solidario*
DC	Distributional Characteristic
EITC	Earned Income Tax Credit
GAMs	Generalised Additive Models
GDP	Gross Domestic Product
HH	Head of Household
IEF	*Ingreso Ético Familiar*
JUNAEB	*Junta Nacional de Auxilio Escolar y Becas*
ME	Chilean Ministry of Education
MLA	Machine Learning Algorithms
MSD	Chilean Ministry of Social Development
OECD	Organisation for Economic Co-Operation and Development
PMTs	Proxy Means Tests
PPP	Purchasing Power Parity
RD	Regression Discontinuity
ROC	Receiver Operating Characteristics
RV	Running Variable
SH	Science and Humanities or Scientific-Humanistic
SIMCE	*Sistema de Medición de la Calidad de la Educación*
SPF	Social Protection File
SUF	*Subsidio Único Familiar* or *Subsidio Familiar*
TP	Technical-Professional
USD	United States Dollars

Chapter 1 Introduction

Income support programmes have changed substantially in recent decades across the globe. Increasingly, developing and developed countries have implemented or expanded programmes where cash transfers are exchanged in return for socially desired behaviour from their recipients.

This thesis analyses key questions located at the intersection of conditional cash transfers (CCTs), targeting mechanisms of CCTs and outcomes in primary and secondary education. The research relies entirely on large and rich administrative datasets from Chile. The thesis contributes to advancing knowledge both methodologically and for social policy. Although the study focuses its analysis on one country, its results have implications for multiple contexts.

This introductory chapter describes the policy and institutional context of my research. The first part of the chapter summarises essential features of CCTs. The second section focuses on the targeting mechanisms used by CCTs, especially proxy means tests. The third part explains the motivation behind the thesis. The fourth section provides details of the rules governing primary and secondary education in Chile. The last section of the chapter gives an outline of the entire document and summarises its most essential findings, contributions and implications.

1.1 Conditional Cash Transfers: Components, Goals and Impact

CCTs are social policies defined by distinctive characteristics. Most generally, these schemes consist of a regular (monthly, bi-monthly or annual) cash transfer. This cash transfer is given if certain conditions related to behaviour are met. Additionally, CCTs tend to be targeted at low-income households or individuals. These schemes can be found in both developing and developed countries. However, different formats of CCTs can be observed.

Developing nations have primarily adopted cash transfers that focus their conditions on infants, children and adolescents. The popularity of these cash transfers, especially in Latin America, can be ascribed to the success of the *Bolsa Família* and *Progresa* programmes in Brazil and Mexico respectively in the late 1990s (Handa & Davis, 2006). CCTs rapidly expanded across the continent and then the world. By the end of the following decade, all the countries in Latin

America, except Cuba and Venezuela (Stampini & Tornarolli, 2012), and a few countries in Asia, the Middle East and Africa had implemented a CCT (Fiszbein & Schady, 2009). By 2015, CCTs were present in at least 64 countries (Honorati, Gentilini, & Yemtsov, 2015).

CCTs in developing countries are commonly paid to the mother (Handa & Davis, 2006). The amount each household receives varies according to the number of children and how many of them meet the conditions. In a typical CCT design, the mother receives a health-related payment if all her children (no older than nearly 6 years old) regularly attend health check-ups. The education-related payment is generally given per child (aged between 7 and 17 years old proximately) attending primary or secondary school (Ibarrarán, Medellín, Regalia, & Stampini, 2017). Deviations from this typical design do occur. For example, some CCTs are provided directly to secondary students while others require children to have completed their immunisations.

One of the goals of CCTs is to increase the income of poor households or individuals. This goal has been framed in diverse but complementary ways in the literature, for example as providing poor households with a minimum consumption floor (Fiszbein & Schady, 2009) or as current poverty alleviation, reducing the incidence and depth of poverty (Handa & Davis, 2006).

The impact of CCTs on income-related indicators has been well documented. Overall, these schemes have increased food expenditure and consumption, reduced the poverty headcount index and the poverty gap (at the programme and the country levels) and in some cases contributed to drops in inequality (Fiszbein & Schady, 2009). For example, Stampini and Tornarolli (2012) estimate, for 13 Latin American countries, that the short-run poverty headcount indicator would have been on average 13% higher if CCTs had not been in place. Additionally, Kabeer, Piza, and Taylor (2012) synthesise six studies estimating the effect of CCTs on consumption and find a statistically significant average effect of 7%.

Despite these findings a common criticism of CCTs is that these schemes could reduce adults' labour supply. This hypothesis has been analysed empirically. The research on this topic in Latin America has found results of very small magnitude or that are not statistically significant (Alzúa, Cruces, & Ripani, 2013; Fiszbein & Schady, 2009). In a similar vein, the average effect of all the studies available regarding this matter is not statistically different from zero (Kabeer et al., 2012).

Additionally, CCTs are designed to promote human capital accumulation. Because of their conditions these cash transfers are expected to increase health care utilisation, school enrolment and attendance.[1] Economic theory suggests that these schemes are useful for the last two purposes. Poor households face budget constraints that may deter human capital accumulation activities (especially secondary schooling) in favour of labour. However, households are expected to maximise their utility considering not only the marginal cost of children's time at school but also the future benefits of human capital accumulation. A CCT then is expected to have an impact by lowering the opportunity cost of children's time spent in school. This is presumed in particular among households for which the budget constraints are binding (Skoufias, 2005).

Fiszbein and Schady (2009) and Skoufias (2005) argue that the use of conditions is justified. The reason is the potential existence of factors that cause households to underinvest, from society's standpoint, in the human capital accumulation of their children. Fiszbein and Schady (2009) elaborate more on these factors, citing as examples imperfect information about the expected returns on education or inconsistent day-to-day behaviour relative to long-term goals.

Although it has been shown that multiple factors are correlated with the decision to drop out of school (Hunt, 2008; Rumberger & Lim, 2008), with economic need being just one of them, the effectiveness of CCTs has been corroborated by an extensive body of research. In Latin America, positive effects have been found on school enrolment in Brazil (Soares, Ribas, & Osório, 2010), Colombia (Attanasio et al., 2010; Barrera-Osorio, Bertrand, Linden, & Perez-Calle, 2011), Ecuador (Schady & Araujo, 2008), Honduras (Galiani & McEwan, 2013; Glewwe & Olinto, 2004), Mexico (Schultz, 2004) and Nicaragua (Maluccio & Flores, 2005). More generally, García and Saavedra (2017) conclude from a meta-analysis that the average effect of CCTs on primary and secondary school enrolment has been 3.0 and 7.1 percentage points, respectively. Likewise, a systematic review of 26 CCTs shows that these policies have increased the odds of a child being enrolled in school by 41% (Baird, Ferreira, Özler, &

[1] At this point, it is important to acknowledge that human capital accumulation could also occur as a consequence of raising income, disregarding the changes in behaviour that might be introduced by the conditions. For example, households could have more food access (or of better quality) because of reduced income restrictions. Then, improved food access could have a positive effect on nutrition and health status for all or some household members.

Woolcock, 2014).

There is also extensive research on the effect of CCTs on health-related outcomes. CCTs have had a substantial impact on health service utilisation in Colombia (Attanasio, Battistio, Fitzsimons, Mesnard, & Vera-Hernández, 2005) and Honduras (Morris, Flores, Olinto, & Medina, 2004); however the evidence is less conclusive for immunisation coverage (Fiszbein & Schady, 2009; Lagarde, Haines, & Palmer, 2007; Owusu-Addo & Cross, 2014). The empirical literature has not exclusively analysed the impact of CCTs on human capital in the short-run. For example, recent studies have evaluated whether CCTs are responsible for effects, on former children or adolescents, at later stages of the life-cycle. The evidence regarding the long-term impact of CCTs though is inconclusive (Molina-Millan, Barham, Macours, Maluccio, & Stampini, 2016).

Developed nations, such as the United States, have primarily implemented cash transfers whose conditions are linked to labour or school performance. The former type of cash transfer is provided on a regular basis mostly to parents who are employed. These policies are recognised by different names such as tax credits, employee subsidies or worker subsidies.[2] The latter type of transfer is given to students, or a household member, as a reward for their achievement in primary or secondary education. These programmes are commonly known as cash for grades.

The impact of worker subsidies on labour force participation has been analysed from a theoretical perspective. Specifically, the Earned Income Tax Credit (EITC) in the United States has been studied in more detail (Dickert-Conlin & Holtz-Eakin, 2000; Neumark, 2013). The models suggest positive impacts of worker subsidies on labour supply for those who are unemployed. This situation is explained by the increased opportunity cost of leisure. However, the EITC could reduce the number of hours worked for individuals who are part of the workforce.

Expansions of the EITC have been evaluated empirically. The most influential study in this area finds that this policy increased the labour supply of single women with children relative

[2] These concepts are not equivalent to a hiring credit or a wage subsidy. The former term has been defined as a subsidy to employers to hire workers (Neumark, 2013). The latter term has multiple definitions. For example, it has been broadly defined as any transfer from the government that can reduce the cost of labour and/or increase take-home pay (Almeida, Orr, & Robalino, 2014); hence it includes employer and employee subsidies. Other specific definitions also exist such as subsidies to employers to hire disadvantaged workers (Katz, 1998).

to single women without children (Eissa & Liebman, 1996). A subsequent study states that the EITC and other tax changes explain over 60% of the increase in employment rates of single mothers, relative to childless single women, between 1984 and 1996 (Meyer & Rosenbaum, 2001). Eissa and Liebman (1996) find no effects on the number of hours worked for single mothers who were in the workforce, but Eissa and Hoynes (2004) observe a decline in labour force participation of married mothers. In the United Kingdom, there is a consensus that the Working Families' Tax Credit has increased the labour force participation of lone mothers but had little overall impact on the labour force participation of parents in couples (Brewer & Browne, 2006).

Regarding cash for grades programmes, these types of incentives should provide a price effect that makes specific behaviours more attractive. These effects are expected to increase students' effort and performance (Fryer, 2011; Gneezy, Meier, & Rey-Biel, 2011; Kremer, Miguel, & Thornton, 2009). However, from the psychological perspective, cash for grades programmes could undermine intrinsic motivation, negatively affecting the desired behaviours, especially after the reward is removed. However, this argument is contested, with scholars both in favour of it (Deci, Koestner, & Ryan, 1999; 2001) and against (Cameron, 2001).

The empirical evidence regarding cash for grades programmes is mixed. In developed countries some positive and statistically significant results exist. For example, Bettinger (2012) finds an effect on maths but he observes no impact on three other subjects. Angrist and Lavy (2009) find statistically significant results for girls but not for boys. In other cases, mostly no effects have been observed (Fryer, 2011; Riccio et al., 2013). In developing countries, effects near 0.20σ on test scores have been found (Behrman, Parker, Todd, & Wolpin, 2015; Kremer et al., 2009).

Formally, the literature tends not to label worker subsidies and cash for grades as CCTs. The term is usually reserved for the design commonly observed in developing countries. However, there are many similarities among all these cash transfers. Given this resemblance and for convenience, I will use the term CCT flexibly. In most cases, I will interpret it narrowly as the design typically found in developing nations. However, in some specific contexts I will use a holistic meaning, encompassing all the cash transfers programmes described in this section.

1.2 Targeting of Conditional Cash Transfers and Proxy Means Tests

The opposite of a targeted intervention is a universal approach. In the context of a cash transfer, a "universal basic income" style-programme where each individual is a beneficiary, irrespective of their welfare, contrasts with a cash transfer targeted at the poor.[3] Targeting requires the definition of the target group (aligned with the policy goal) and the choice of a mechanism to reach this group. A targeting assessment then verifies whether the target group is reached.

Given that one of the primary goals of CCTs is to increase income among poor people, targeting is an essential part of their design. These schemes are expected to be delivered to low-income households or individuals. The more resources that are directed towards the target group (poor people), the more likely a CCT is to achieve its goal of poverty alleviation. Coady, Grosh, and Hoddinott (2004) explain that the motivation for targeting lies in the grounds of efficiency. Targeting allows achieving maximum poverty reduction (or social welfare increase) within a fixed budget.

CCT targeting evaluations show that these schemes have been far more successful in finding households or individuals who live in poverty relative to allocating the transfers at random (Maluccio, 2009; Robles, Rubio, & Stampini, 2015; Skoufias, Davis, & De la Vega, 2001). Stampini and Tornarolli (2012) show that the expansion of CCTs in Latin America led to increased inclusion of the poor. For example, by 2010, the three largest CCTs (in Colombia, Mexico and Brazil) had reached 50% to 55% of the poor. However, this increase in inclusion of the poor has been accompanied by a growing proportion of beneficiaries who are not poor.

Azevedo and Robles (2013) explain that CCTs have fared well in reaching the poor but not in identifying households that under-invest in human capital. These households are an additional target group for CCTs. However, specific health and education deprivations have barely been used in CCT allocation and few targeting evaluations exist that analyse beyond income.

CCTs are generally targeted using income-related measures. However, income or consumption

[3] A long debate exists about targeted versus universal cash transfers. For example, Besley (1990) provides an analytical framework to compare these two approaches in terms of poverty alleviation. More recently, Hanna and Olken (2018) empirically analyse the trade-offs between the two approaches using data from Indonesia and Peru.

is difficult to observe accurately in developing countries. Coady et al. (2004) identify three families of targeting methods. In the first place they highlight individual or household assessments. These are mechanisms in which someone or something (usually a government employee or a computational system) directly assesses, case by case, the eligibility of the individual or the household. Within this category we can find verified means tests, simple means tests, proxy means tests (PMTs) and community-based targeting. Secondly, there is categorical targeting. In this mechanism all households or individuals within a specific category or group of categories (such as age, gender or geographical region) are eligible. Finally, there is self-selection, where the design of a universal programme encourages poor people to apply.

Within the context of CCTs, few studies have assessed the performance of different targeting mechanisms. In Mexico, Coady (2006) finds that geographic targeting was more effective than a household PMT and targeting based on demographic composition. He also concludes that as *Progresa* expanded away from the poorest localities, the contribution of proxy means targeting substantially increased. In Indonesia, an experimental design shows that requiring beneficiaries to incur small application costs resulted in improved targeting through self-selection (Alatas et al., 2016). Also in Indonesia, for an unconditional cash transfer, a PMT shows an increased capacity to find poor households relative to community-based targeting or a hybrid mechanism (Alatas, Banerjee, Hanna, Olken, & Tobias, 2012). However, the authors additionally conclude that the targeting mechanism choice would not affect the country's poverty rate significantly.

In Latin American countries, most CCTs use geographic targeting, proxy means tests or a combination of both (Fiszbein & Schady, 2009; Stampini & Tornarolli, 2012). PMTs rely on observable characteristics of the household (such as the education of its members) to estimate their income. More formally, PMTs refer to a system where information correlated with income is used in an algorithm to approximate income itself. The formula is generally obtained through statistical analysis and tends to use data that is easily observable by public officials (Coady et al., 2004; Grosh & Baker, 1995). Beyond the continent, countries such as Armenia, Bangladesh, Cambodia, Cameroon and Rwanda have also used proxy means tests for targeting (Australian Aid, 2011; Brown, Ravallion, & Van de Walle, 2016; Coady et al., 2004).

The proxy means test score is an essential eligibility criterion for many CCTs. Households or individuals with a PMT score below a certain threshold can access these programmes. Conversely, those who have a PMT score above this threshold or fail to obtain a score are not eligible. For example, Ponce and Bedi (2010) explain that in Ecuador, the CCT *Bono de Desarrollo Humano* used the SELBEN PMT index to assign the scheme. The index scale ranged from 0 (the poorest) to 100 points (the richest). Households with a low quality of infrastructure or with members with little education were more likely to have a lower score in SELBEN relative to other households. Between 2003 and 2009, the threshold used by *Bono de Desarrollo Humano* was 50.65 points.

CCT extensive use of index' scores (such as PMTs) to determine eligibility facilitates using regression discontinuity (RD) designs to estimate their impact. RD designs are a popular approach for causal inference. The method relies on comparing units just below and above a threshold in a score (or running variable) that is used to assign an intervention. RD designs exploit the idea that variation in treatment assignment near the threshold is as good as random. Various RD designs exist for CCTs, for example in Colombia (Barrientos & Villa, 2015), Ecuador (Ponce & Bedi, 2010), Chile (Carneiro, Galasso, & Ginja, forthcoming), Jamaica (Stampini, Martinez-Cordova, Insfran, & Harris, 2018) and Cambodia (Filmer & Schady, 2011).

1.3 Why Study Conditional Cash Transfers?

A first reason to justify a thesis on the topic of CCTs is the existence of the problems that these policies are trying to overcome. The following two paragraphs briefly explain the magnitude of the problem of low income and the gaps in school enrolment in Chile and Latin America.

In 2013, with a GDP (PPP) per capita close to $22,000 USD, Chile was one of the wealthiest countries in Latin America, but one of the poorest OECD nations (OECD, 2015). In 2011 and 2013 respectively, 10.9% and 7.8% of individuals lived in poverty (Ministerio de Desarrollo Social, 2015). The median monthly household income per capita reached only $146,555 CLP in 2011 ($279.6 USD using November 30th 2011 exchange rate) leaving approximately 39% of the population between this income level and the urban poverty line, which reached $61,366 CLP ($117.1 USD). In 2013, 7.9% of adolescents between 15 and 19 years old, about 110,867 individuals, were not attending any school and had not finished their secondary studies. Since 2010, the percentage of children that drop out each year from secondary studies has been near 3.0% (Opazo, Ormazabal, & Crespo, 2015), while dropouts in primary education are low.

Among Latin American countries social indicators have improved over time but poverty and school dropout remain relevant challenges. For example, Vakis, Rigolini, and Lucchetti (2016) show that extreme poverty (defined as those living with less than $2.5 USD PPP a day) declined from 24.5% to 12.3% between 2003 and 2012. Additionally, the percentage classified as moderate poor also decreased from 41.6% to 25.3% during this period. Regarding the completion of secondary education, Bassi, Busso, and Muñoz (2015) show that the regional average graduation rate (among those one year older than the school finishing age but younger than 27 years old) increased from 0.31 in the early 1990s to 0.54 in the late 2000s.

An additional reason to devote this study to CCTs is the current state of these policies in terms of their outreach. In Latin America, the number of CCT beneficiaries overtook the poor population in 2006. The massive rise in CCT recipients on the continent has raised a debate in the literature about whether these schemes have gone too far (Stampini & Tornarolli, 2012). As the relationship between CCT recipients and the population in poverty is not equal in each nation, the coverage of CCTs in many countries has exceeded the population living in poverty. In this context, finding new households living in poverty has become harder for CCTs. Additionally, their effectiveness (regarding poverty reduction) could be reduced by increased targeting errors. These problems are expected to grow as it is unlikely that governments will cut cash transfers and as poverty rates decline.

Some social policy questions naturally arise from the previous analysis, such as: i) should CCTs be reduced in scope, or at least their growth stopped, for countries that have substantially reduced poverty rates? ii) are CCTs working for the non-poor or for those beneficiaries who barely qualify? iii) should the targeting of these programmes be expanded to other CCT target groups, beyond the poor, such as students at risk of dropping out of school or children with low health care services utilisation? and, iv) should developing countries shift from CCTs to employment subsidies or cash for grades schemes, which are rarely present in these nations?

No single study can provide definitive answers to all these substantial questions. The focus of this thesis lies in the analysis of three specific questions located at the intersection of conditional cash transfers, targeting mechanisms of CCTs and outcomes in primary and secondary education using large and rich administrative datasets from Chile. In relation to the targeting of CCTs, the thesis focuses especially on proxy means tests. Chile is an ideal setting in which to conduct research related to PMTs, given that the country has been a pioneer in the use of this mechanism (Australian Aid, 2011; Fiszbein & Schady, 2009; Grosh & Baker, 1995).

1.4 The Primary and Secondary Education System in Chile

Educational outcomes play an essential role in my thesis. Therefore, this section explains in detail the Chilean educational institutional context for the primary and secondary levels.[4]

The right to education is stated in Article 19 of the Chilean Constitution, which establishes that the purpose of education is the development of human beings in different stages during their lives and that parents must educate their children. The Constitution additionally specifies that the State must give special protection to this right. In 2003, a constitutional reform introduced two significant changes. First, it establishes that both primary and secondary school are compulsory. As in the original Constitution, this new version forces the State to finance a free system at these levels to guarantee its access to the whole population. Second, the reform states that the traditional secondary school system can be provided until the age of 21.

Other regulations define schooling length and age requirements. These regulations establish that: i) the traditional primary school system is comprised of eight grades (first to eighth) and lasts for at least eight years, ii) children have to be at least 6 years old by March 31st to be able to start the first grade, but iii) educational institutions´ directors are allowed to accept five-year-olds whose birth date is no later than June 30th, iv) the traditional secondary school system is comprised of four grades (ninth to twelfth) and lasts for at least four years, and v) the maximum age at which an adolescent can start the traditional secondary school system is 16 years old.

An important feature of the Chilean school primary and secondary systems is that students are not automatically promoted to the following grade. At the end of an academic year students are classified into one of the following categories: i) progress, ii) failure or iii) withdrawal. The first group is comprised of students who have completed the academic year and meet the requirements to progress to the following grade. The second group of students are those who have completed the academic year but do not meet the requirements to progress. The third group of students abandoned their studies during the academic year. Thus, because the primary

[4] The information from this section comes from multiple laws and decrees dictated by the Chilean Ministry of Education. All these documents have been published by the *Biblioteca del Congreso Nacional de Chile* (1980, 1997, 1999, 2003, 2007, 2009, 2011, 2012, 2015). These documents can be found at https://www.bcn.cl/.

and secondary levels are compulsory in Chile, the second and third groups of students are expected to repeat the grade at which they did not progress in the next academic year.

To explain the requirements to progress in detail it is necessary to describe some characteristics of the Chilean grading system and attendance measurement. For grading, the country uses a scale from 1.0 to 7.0 in every educational level (pre-school, primary, secondary and higher education). This scale is used for specific courses, such as Mathematics or Arts, and the average grade. At the primary and secondary levels, the average grade is mostly estimated as the simple average of the individual course grades and is reported to the central government using only one decimal place. Attendance is measured daily at the primary and secondary levels and is reported to the Ministry of Education monthly and yearly. The scale used is a percentage, from 0% to 100%, without utilising decimal places. From fourth to twelfth grade, the requirements to progress to the next grade are to reach an attendance of 85% and a minimum average grade.[5]

There is no curriculum differentiation from first to eighth grade. All primary schools are expected to implement a standard curriculum. The secondary system is mainly divided into two groups. Science and humanities schools had roughly 70% of the total enrolments in 2014, while technical-professional schools accounted for nearly 30%. In both types of schools, it is expected that students will complete their secondary education as established by the Constitution.

Chilean schools can be classified into mostly three administrative categories. In 2014, private non-subsidised schools captured around 8% of total enrolments. Almost all these schools charge high fees and as a result virtually all their students come from wealthy families. Public schools are generally free and administered by local counties. These schools had approximately 39% of total enrolments in 2014. Finally, private subsidised schools accounted for nearly 52% of total enrolments in 2014. These schools may charge a small fee, but most of them fully fund their operations with a voucher provided by the State. From 2016, new legislation forced these schools to eliminate charging fees progressively. Only private non-subsidised schools do not receive State vouchers, as public and private subsidised schools rely on them.

[5] Specifically, one of the following criteria needs to be met: i) obtain a grade of 4.0 or higher in every specific course, ii) fail to obtain a grade of 4.0 in one specific course but obtain an average grade of at least 4.5, iii) fail to obtain a grade of 4.0 in two specific courses but obtain an average grade equal to or higher than 5.0. For the latter group of students in eleventh and twelfth grade it is required to obtain an average grade of at least 5.5 if Mathematics or Language is one of the two courses where the student fails to obtain a grade of 4.0.

The Chilean education system additionally recognises two other schemes. One of them is differential education for students with disabilities. Differential education is not classified by grades or levels (as in the traditional system). The other scheme is adult education, which uses the same classifications by level and grade as the traditional system. A relevant difference between the adult education system and the traditional system is that the former allows students to progress by more than one grade, generally two, in one academic year. Some of these schools (regular adult education providers) demand an 80% attendance level at classes and assess progress as in the traditional system, while others (flexible adult education providers) do not require attendance and assess progress using national examinations. In general, it is not possible for students to enter these types of schools unless they are at least 17 or 18 years old.

Table 1.1 and Table 1.2 summarise the grade length and age requirements for all the options available for education provision at the primary and secondary levels, respectively.

Table 1.1: Grade Length and Age Requirements in Primary Education

Grade	Traditional Education	Adult Education	
		Regular	Flexible
1st	1 grade per year, has to be at least 6 years old to start	4 grades per year Has to be at least 18 years old to start	
2nd	1 grade per year		
3rd	1 grade per year		
4th	1 grade per year		
5th	1 grade per year	2 grades per year Has to be at least 18 years old to start	
6th	1 grade per year		
7th	1 grade per year	2 grades per year Has to be at least 18 years old to start	
8th	1 grade per year		

Source: compilation based on official information, Chilean Ministry of Education

Table 1.2: Grade Length and Age Requirements in Secondary Education

Grade	Traditional Education (All Types of Schools)	Adult Education		
		Regular		Flexible
		Type of School		
		Science and Humanities	Technical-Professional	No Classification
9th	1 grade per year, has to be less than 17 years old to start and cannot be over 21	2 grades per year, has to be at least 17 years old to start		2 grades per year, has to be at least 18 years old to start
10th	1 grade per year, has to be no more than 21 years old			
11th	1 grade per year, has to be no more than 21 years old	2 grades per year, has to be at least 18 years old to start	1 grade per year, has to be at least 18 years old to start	2 grades per year, has to be at least 18 years old to start
12th	1 grade per year, has to be no more than 21 years old		1 grade per year, age rule not specified	

Source: compilation based on official information, Chilean Ministry of Education

The Chilean academic year starts in March and typically ends by mid-December. Currently, a Chilean student that graduates from secondary education and progressed each academic year in the traditional system can be described by the following pattern: i) entered the first grade of primary education at the age of 6 (or almost 6 if born between March and June), ii) entered the first grade of the secondary school system at the age of 14 (or almost 14), iii) entered the last grade of secondary education at the age of 17 (or almost 17) and finally, iv) completed this last grade at the age of 18 (or 17 if the student's month of birth is in January or after this month).

This traditional pattern of graduation may be affected in three ways. First, a student could have had a late start, in other words, he or she might have commenced the first grade of primary education at an age of more than six. Second, a student might not have progressed to the following grade due to failure to accomplish the attendance or academic requirements or because of withdrawal. Third, irrespective of whether a student progressed, failed or withdrew in a given academic year he or she might not have enrolled in any educational institution in the next academic year.[6]

[6] School enrolment can be measured in different ways. The approach I consistently use in this document is defined by data availability and by the methodology used by the Chilean State (Ministerio de Educación, 2013) for estimating dropout rates. Only students in the traditional system or regular adult education are classified as enrolled. Students in flexible adult education are not classified as enrolled as the Ministry of Education does not have administrative records for these students. Students in differential education are not considered in this study.

If any of these three situations occurred at some point in the pathway, the student might be able to choose between the traditional and adult systems at a later stage in their life. For those who do not start secondary education before turning 17 years old or who do not graduate from school before turning 21 years old, adult education is their only choice to start or resume their studies.

1.5 Thesis Outline

The rest of the thesis is built around three empirical chapters or papers. Each paper addresses a different type of question. The second chapter assesses how well two target groups of CCTs can be reached. The third chapter estimates the impact on educational outcomes of those who get a cash transfer. I use an RD design for this purpose. The fourth chapter elaborates on some limitations of the RD method in a setting of CCTs and proxy means tests targeting.

The second chapter contributes to the social policy targeting field. This paper analyses whether a common targeting mechanism of CCTs, a proxy means test, can identify the poor and future school dropouts effectively. Despite both being key target groups for CCTs, students at risk of dropping out are rarely considered for CCT allocation and in targeting assessments. Using simulations, I compare the targeting effectiveness of a proxy means test with other mechanisms based on a predictive model of school dropout. I build this model using machine learning algorithms, one of the first applications of this type used to predict school dropout outside a developed country.

The paper shows that using the outputs of the predictive model in conjunction with the PMT increases targeting effectiveness by identifying more students who are either poor or future dropouts. This joint targeting approach increases effectiveness in different scenarios except when the social valuation of the two target groups differs to a large extent. In these cases, the most likely optimal approach is to use solely the mechanism designed to find the target group that is valued the most. Overall, the paper shows that public officials that value these two target groups equally may improve CCT targeting by modifying the allocation rules of these programmes.

The third chapter analyses whether it is convenient to reward children for their academic performance. This paper estimates the impact of a Chilean cash for grades programme, the

Bono por Logro Escolar (BLE) in 2013, on future attendance and average grade. The cash transfer was targeted using two scores from 2012, a proxy means test and academic performance. I implement a sharp regression discontinuity design along these two running variables. I show that students marginally at each side of the two thresholds used for targeting only differed in receiving the BLE in 2013. The main causal estimates for the outcomes are not statistically significantly different from zero. If a local average effect of the BLE in 2013 exists then this is at best modest in magnitude and at least smaller than those found in the literature in developing countries, where effects near 0.20σ on test scores have been observed.

The fourth chapter contributes to the causal inference literature. This paper elaborates on two threats, administrative sorting and intermediate contamination, which have been overlooked in the RD literature in the context of three impact evaluations of CCTs on school enrolment.

Lee and Lemieux (2010) claim that “if individuals are unable to precisely manipulate the running variable then variation in treatment near the threshold is as good as random” (p283). The paper shows that variation in treatment assignment is not always as good as random even in the absence of manipulation. This can be the case when administrative procedures, beyond individuals’ control and knowledge, affect their position near the threshold non-randomly. If administrative sorting is not properly recognised it can be mistaken as manipulation. Timing also matters in RD designs. Intermediate contamination can emerge, and make units near the threshold no longer comparable, if a substantial time lag exists between the realisation of the running variable and its actual use to assign the treatment. In this type of setting, the paper highlights the value of checking variables related to this time lapse in RD falsification tests.

The final thesis chapter summarises the findings of the three empirical chapters. Additionally, it discusses the main implications and contributions of these findings, both concerning policy and methodology, for improved design and evaluation of CCTs as well as beyond CCTs.

Chapter 2 Two Become One: Improving the Targeting of Conditional Cash Transfers With a Predictive Model of School Dropout

Abstract

This paper analyses whether a common targeting mechanism of conditional cash transfers (CCTs), an income-proxy means test (PMT), can identify the poor and future school dropouts effectively. Despite both being key target groups for CCTs, students at risk of dropping out are rarely considered for CCT allocation and in targeting assessments. Using rich administrative datasets from Chile to simulate different targeting mechanisms, I compare the targeting effectiveness of a PMT with other mechanisms based on a predictive model of school dropout. I build this model using machine learning algorithms, one of their first applications for school dropout outside a developed country. I show that using the outputs of the predictive model in conjunction with the PMT increases targeting effectiveness except when the social valuation of the poor and future school dropouts differs to a large extent. Public officials that value these two target groups equally may improve CCT targeting by modifying their allocation procedures.

2.1 Introduction

Conditional cash transfers have become a favoured social policy in developing nations. Their rapid expansion, from a few countries in the late 1990s to more than 60 by 2014 (Honorati et al., 2015), demonstrates their popularity. One goal of CCTs is to increase the income of low-income households. Although CCTs have not stated their objectives identically across the globe, authors have pointed out that these schemes seek to reduce the incidence and depth of poverty (Handa & Davis, 2006) and provide a minimum consumption floor to poor households (Fiszbein & Schady, 2009).

Targeting is a crucial element in the design of CCTs. These programmes have intended to allocate their benefits primarily or “rather narrowly” to the poor (Fiszbein & Schady, 2009) (p.7). This is not unplanned, as poor households or individuals are a key target group for CCTs. Targeting is a channel through which to increase a programme’s effectiveness within a fixed budget. The more resources that are directed towards the target group (the poor), the more likely a CCT is to achieve its goal of poverty reduction. This explains why evaluations of their targeting (Maluccio, 2009; Robles et al., 2015; Skoufias et al., 2001; Stampini & Tornarolli, 2012) have focused primarily on determining whether CCTs have been given to those who live in poverty.[7]

Different targeting mechanisms exist for CCTs. Proxy means tests are one of the most common in Latin America (Fiszbein & Schady, 2009; Stampini & Tornarolli, 2012). A PMT refers to a system or situation where information correlated with income is used in a formula to proxy income. The formula is obtained through statistical analysis and tends to use data that is easily observable by public officials (Coady et al., 2004; Grosh & Baker, 1995).

Targeting low-income households or individuals makes sense not only for CCTs but for a wide range of social programmes. Correspondingly, assessing targeting mechanisms of social policies in terms of their ability to find this target group is a widespread practice. For example, Coady et al. (2004) evaluate the pro-poor targeting performance of 122 social programmes from 48 countries. Similarly, Grosh and Baker (1995) assess whether PMTs provide useful information on income to target social programmes in three countries in Latin America. Social policy targeting has been implicitly associated with finding the poor and poverty alleviation.

Since many CCTs are provided only if children or adolescents are enrolled in school, an additional purpose of most CCTs is to increase school enrolment (Handa & Davis, 2006). Hence, to maximise the likelihood of achieving this goal, CCTs need to be delivered to a differently-defined target group, namely students with the highest risk of dropping out of primary or secondary school. However, CCT targeting has been focused more on the income dimension. Thus, CCTs have rarely been assessed regarding their ability to direct their resources to those who are more likely to drop out of school. My paper addresses this gap in the CCT literature.[8]

[7] Stampini and Tornarolli (2012) provide targeting assessments for 13 countries in Latin America. The authors show that the expansion of CCTs on the continent led to increased inclusion of the poor. For example, by the year 2010 the three largest programmes (in Colombia, Mexico and Brazil) had achieved poor coverage rates near 50%. However, this was accompanied by growing levels of non-poor leakage (the proportion of CCT recipients who are not poor). On average, leakage increased by 0.46 percentage points for each additional point in poor coverage.

[8] Analysing the targeting effectiveness of CCTs in terms of reaching students at risk of dropping out of school is different from assessing the impact of CCTs on school dropout. The former evaluation assesses whether the target group is (or would be) reached by a programme. The latter assessment focuses on the (potential) effect of the programme after implementation. The literature on the impact of CCTs on school enrolment is vast, especially in Latin America. For example, positive effects of CCTs on school enrolment have been found in Colombia

Assessing the capacity of CCTs to reach potential school dropouts is very important. Without this knowledge the allocation process of CCTs could be sub-optimal from the human capital accumulation perspective. If the targeting mechanism used by a CCT is not an accurate predictor of school dropout, then some students will be given the CCT despite the fact that they would have finished their secondary education without any intervention. Conversely, other students who are at risk of leaving school will never have been the subject of the CCT. In both cases a problem of misidentification exists and its consequence is an ineffective use of resources.

Not considering potential dropouts when targeting CCTs would be less of a cause for concern if school dropout were a negligible problem. But in Latin America the graduation rate (among those one year older than the school finishing age but younger than 27) only reached 0.54 in the late 2000s (Bassi et al., 2015). Similarly, dismissing potential dropouts in CCT targeting would be less of a problem in contexts where a high degree of overlap exists between the latter group and those living in poverty. However, this is not guaranteed. For example, in Chile in 2013, only 16.1% of young school dropouts (aged 15 to 19 years old) lived in a poor household while only 12.4% of poor adolescents had dropped out of school (Opazo et al., 2015).

Targeting CCTs exclusively according to the likelihood of dropping out of school would lead to a different problem. In this case, the ability of CCTs to find the poor would be weakened. The problem of targeting both groups for a CCT is well addressed by Maluccio (2009), who states: “It is clear that not all non-poor children attend school. Such children, then, would be missed under a pure poverty-based targeting scheme, but possibly not under a targeting scheme which focused on enrolment. Conversely, many poor children already attend school. While there certainly would be overlap among the beneficiary households selected under various possible approaches they almost certainly would not yield identical groups of beneficiaries” (p.5).

This paper analyses whether a proxy means test can identify the poor and future school dropouts effectively. I evaluate the capacity of a PMT to jointly identify these two target groups

(Attanasio et al., 2010; Barrera-Osorio et al., 2011), Ecuador (Schady & Araujo, 2008), Honduras (Galiani & McEwan, 2013; Glewwe & Olinto, 2004), Mexico (Schultz, 2004) and Nicaragua (Maluccio & Flores, 2005).

relative to alternative targeting mechanisms available to public officers. I use rich administrative datasets from Chile to simulate different targeting mechanisms. The core of the alternative mechanisms I test is a predictive model of school dropout. Therefore, the paper has two complementary parts. From its first part, I derive the predictive model using a range of machine learning algorithms (MLA). In the second part, I assess the targeting effectiveness of the PMT, the predictive model and mechanisms combining both sources of information. The paper seeks to advise about the merit of using other targeting mechanisms instead of PMTs for CCTs.

My paper contributes to the body of knowledge on school dropout. The literature to date has mostly focused on the question of why students drop out rather than who will drop out. The work in the latter area is primarily confined to North America, where Bowers, Sprott, and Taff (2013) provide a comprehensive summary. Additionally, machine learning applications of dropout in educational contexts have been more salient for higher education with few papers focusing on primary and secondary schools (Knowles, 2015; Sara, Halland, Igel, & Alstrup, 2015; Sorensen, 2018). Furthermore, the few existing papers that predict school dropout in developing countries (Adelman, Haimovich, Ham, & Vazquez, 2017) have not used MLA.

My paper additionally contributes to the literature on the targeting of CCTs. There have been few attempts to assess CCT targeting that consider more dimensions than just income. A notable exception is Azevedo and Robles' (2013) evaluation in Mexico. To assess the targeting performance of a CCT, their paper presents indicators separately for each dimension. My paper offers not only unidimensional indicators but also two indicators that combine information from the target groups, which facilitates making comparisons between targeting mechanisms. My paper provides a headcount index but also a measure of social welfare to assess targeting.[9]

I compare the MLA using receiver operating characteristic (ROC) curves.[10] The most effective

[9] Some other differences exist between Azevedo and Robles' (2013) paper and mine**.** Their paper focuses on three dimensions (income, health and education) at the household level while I focus on two dimensions (income and education) for individuals in a specific age range for which these dimensions are critical. Their paper uses few variables or predictors to identify deprivation or risk in the educational dimension. Additionally, they use normative criteria (arbitrary selection of thresholds and weights to combine the predictors) for this purpose. My paper uses a larger pool of variables, which allows for predicting empirically which adolescents will drop out of school.

[10] An ROC curve presents the false positive rate and the true positive rate of all the possible results of a predictive model simultaneously. Its area under the curve measures the overall predictive performance of a model. A model

algorithm leads to an area under the curve (AUC) of 0.866. I observe this result for the model predicting school dropout at any point within two years. The most effective algorithm for predicting school dropout within one year produces an AUC of 0.893. My results are better than the ones obtained in Guatemala and Honduras (Adelman et al., 2017) and are in line with the best models tested in the United States (Knowles, 2015; Sorensen, 2018).

In my targeting assessment, there is a trade-off between using the PMT relative to using the MLA-based predictive model. When I use the PMT to target a hypothetical CCT, the targeting indicators associated with the poor improve, but the indicators related to dropouts worsen. The opposite also holds. For different fixed budgets, total leakage (the fraction of students receiving the CCT who are neither poor nor a future dropout) is minimised when I use both instruments in conjunction with each other. In other words, it is more effective to combine the predictive model and the PMT than to use them independently. However, this is not true when the social valuation of the two target groups differs to a large extent. If allocating the CCT to a poor student is four times more valuable relative to allocating it to a future dropout, or vice-versa, the likelier optimal approach is to use solely the mechanism designed to find the target group that is valued the most.

All these results have important policy implications. Firstly, the paper shows that appropriate predictive models of school dropout using administrative datasets can be available for public administrators. These models can prove useful not only for CCTs but also for further policies, such as Early Warning Systems, whose purpose is to prevent school dropout. Secondly, in contexts where public officials value finding the poor and future school dropouts equally, the paper demonstrates that the targeting of a CCT can be improved when other dimensions beyond income are considered in the design. This result highlights the importance of: i) avoiding misalignment between the policy goals, its target groups and the selection of the targeting mechanisms, and ii) developing targeting assessments that are in line with all these definitions.

In summary, the paper provides novel contributions to the social policy targeting field. Overall, the paper's findings are not only relevant for the specific Chilean case but for all developing countries that either have CCTs, wish to develop predictive models of school dropout using

that makes no predictive mistakes has an area under the curve of 1, while a model that predicts at random should achieve an area under the curve near 0.5. Further information about ROC curves is available in the next section.

administrative records or want to strengthen the targeting effectiveness of their social policies.

The paper unfolds as follows. The second section introduces the data and describes the methods I use in the development of the predictive model of school dropout and the targeting assessment. The third section presents the results of the MLA predicting school dropout. The fourth section shows the findings for the targeting assessment. The concluding section discusses the paper's main findings and comments on its main contributions and implications.

2.2 Data and Methods

This section describes in detail the data and methods. The first subsection introduces the data. The second part presents the methodological approach of the predictive model of school dropout. The third subsection elaborates on the procedures and indicators of the targeting assessment. Finally, the fourth part explains how I structure the dataset for the analysis.

2.2.1 *Data*

Most of the datasets I use were provided by the Ministry of Social Development (MSD) at my request. I combine the datasets using the individual ID number provided by the Chilean State. For privacy purposes the ID numbers were changed by the MSD using an algorithm that is unknown to me but enabled me to merge the datasets. The two most important sources of information in this research are the Ministry of Education (ME) Performance Dataset and the Social Protection File (SPF) Dataset.

Ministry of Education Performance Dataset

This dataset contains information for the entire population of students who finish an academic year in primary and secondary education. The dataset only excludes students in differential education and flexible adult education. Each yearly dataset has approximately 2,950,000 observations (one per student). I requested eight datasets (from 2009 until 2016) for this paper.

The variables available in this dataset are: i) school ID (9,500 unique values), ii) type of school (with categories such as traditional primary education, scientific-humanistic or technical-professional secondary education), iii) grade (first to twelfth), iv) academic performance, v) percentage of attendance, vi) academic end of year classification,[11] and vii) student ID.

With this information I create additional variables such as school dropout. A full explanation of this indicator is available in the next subsection. Other variables I create directly from this dataset are: i) school size, ii) academic cohort size,[12] iii) relative academic performance, iv) relative attendance,[13] v) school mobility, and vi) historical dropout rates by school.

More educational information at the school level is available from public sources. Using the variable school ID, as a key to merge, I obtain the schools': i) administrative dependency (such as public or private subsidised), ii) geographic location, iii) urban or rural status, iv) average performance in SIMCE (the national standardised test), and v) management indicators.[14]

Social Protection File Dataset

This dataset contains information for Chilean households and all their members. The dataset has a two-level structure. Each observation represents an individual (adult or child) who lives in a household. No individual can belong to more than one household. Each household has a unique ID number that allows for identifying all the individuals who belong to it.

Households voluntarily requested the SPF at the local government level. Having an SPF was essential to be eligible for multiple social policies. From January 2010, the dataset had 10,782,270 individuals (Comité de Expertos Ficha de Protección Social, 2010), approximately 63.5% of Chile's population. I use four of these datasets (from 2011 to 2014) in this research.

Some of the variables I access are: income, date of birth, proxy means test score, gender, race,

[11] At the end of an academic year students are classified into one of the following categories: i) progress, ii) failure, or iii) withdrawal. Thus, because the primary and secondary levels are compulsory in Chile, the second and third groups of students are expected to repeat the grade at which they did not progress in the next academic year.

[12] Three variables define an academic cohort: i) the school, ii) the type of education received within that school (for example traditional or adult education, scientific-humanistic or technical-professional), and iii) the grade in which the students were enrolled. Students belonging to the same cohort have these characteristics in common. Most schools have a specific orientation. However, some schools offer more than one type of education in a given grade (especially in secondary education). Students can also change streams from one academic year to the other.

[13] I create the relative academic performance and relative attendance variables by ranking all students within an academic cohort and dividing the resulting ranking of each student by their academic cohort size.

[14] These are six variables associated with improving the quality of education within each school. The first two are linked with the educational outcomes obtained by the school (such as the trend in performance). The next two factors are related to the operation of the school (for example their ability to incorporate educational innovations). The remaining variables give an account of the involvement of the school community and the degree of equity.

head of household, schooling and employment.[15] With this information I can generate additional variables for each individual such as poverty status (explained in detail in the third subsection) and number of children less than six years old within the household.

I combine the information from the Social Protection File with the Ministry of Education Performance Dataset (at the individual level) to build variables for each academic cohort of students. Some of these variables are: average household income per capita, average schooling of the head of the household and proportion of students with a proxy means test score.

2.2.2 *Methods: Predictive Model of School Dropout*

This subsection explains in detail the methodological approach I take to build the predictive model of school dropout. The first part focuses on the predictors. Secondly, the subsection describes the outcome. The third part elaborates on the characteristics of the functions I use for the predictions. The subsection concludes with the criteria I use to assess the predictions.

In general terms, the problem I address in this part of the paper is to find the best function to predict future school dropout given the information I observe from the past. Most formally:

$$Y_{it+k} = f(X'_{it}, X'_{it-1}, \dots \dots, X'_{it-\mathrm{j}}, Z'_i),$$

I need to find a function f that, given the vectors of variables X'_t (where *t* is year), X'_{t-1},….., $X'_{t-\mathrm{j}}$ and Z' available for each individual i, produces on average the most accurate prediction of the outcome Y in *t*+k. Given that the outcome, school dropout, is a dichotomous variable, this is a statistical classification problem and f is known as a classifier.

The Predictors

I include two types of predictors in the model. The first group of predictors are contained in vectors $X'_t, X'_{t-1}, \dots \dots, X'_{t-\mathrm{j}}$. Specifically, X'_t is a vector of variables that change through time for student i (such as academic performance, grade repetition, attendance and mobility). The second group of predictors are embedded within Z', a vector of variables for student *i* that do

[15] The fourth chapter (subsection 4.3.1) explains in detail the proxy means test score of the Social Protection File.

not vary through time (such as race) or where I can only use one observation (such as age).

The selection of variables I include in the model is motivated by the literature on determinants of school dropout and bounded by the availability of administrative records.

Rumberger and Lim (2008) summarise 203 studies for the United States over 25 years to identify statistically significant predictors of school dropout. Some individual characteristics of students that are relevant predictors are: i) educational performance (for example academic achievement, mobility, grade promotion and age or difference between age and expected age for the grade), ii) behaviours (such as absenteeism, deviance and employment), iii) attitudes (like goals and self-perceptions), and iv) background (for example demographics and health).

Their review also identifies institutional characteristics of students' families, schools, and communities. For example, the structure, practices, financial and human resources of students' families are singled out as predictors. Additionally, the student composition, structural characteristics, resources, processes and practices of schools are highlighted in their research.

Hunt (2008) reviews the literature on factors associated with school dropout in developing countries. The author identifies similar predictors to Rumberger and Lim (2008), but also adds other intrinsic challenges that these nations face such as migration, conflict and limited school supply.

The complete list of predictors I use is available in Appendix A. There are 50 variables in total that aim to cover all the dimensions highlighted by Rumberger and Lim (2008). Given the nature of the sources, the information is richer regarding educational performance and the characteristics of students' families, relative to other predictors such as students' attitudes towards education.

A complementary source for predictor selection is Lamote et al. (2013), who argue that predictive models of school dropout need to account for the longitudinal and hierarchical structure of the datasets. This makes perfect sense due to the relevance of educational performance, which is a time-variant variable, and schools and communities as predictors of future dropout. Accordingly, where this is feasible, all my models use three years of historical information $(X_t', X_{t-1}', X_{t-2})$ and include many variables that are at a higher level than the students.[16]

The dataset I assemble is appropriate for the task as it includes multiple strong predictors of school dropout. The dataset possesses multiple years of information on academic attainment, mobility, attendance in conjunction with information at the household level (such as years of schooling of its members and income per capita) and the school and academic cohort levels.

The Outcome

I use the Ministry of Education Performance Dataset to identify students who dropped out of school. The process involves merging different years of this dataset, and linking observations by the student ID. More precisely, I link each student in primary and secondary education who concluded their academic year *t* and did not graduate from their secondary studies with itself in years *t*+1 and/or *t*+2. Using this procedure, I identify the students that dropped out of school after year *t*.[17] Student dropout can be measured in multiple ways. I use three different measures of school dropout to verify the consistency of the results. These measures are:

- dropout_t1: The student finished the academic year *t* and then failed to enrol in *t*+1 or enrolled but withdrew before the end of the academic year *t*+1.
- dropout_t2: The student finished the academic year *t* and (disregarding what happened in *t*+1) then failed to enrol in *t*+2 or enrolled but withdrew before the end of the academic year *t*+2.
- dropout_t12: The student finished the academic year *t* and then failed to enrol in *t*+1 or enrolled but withdrew before the end of the academic year *t*+1 or failed to enrol in *t*+2 or enrolled but withdrew before the end of the academic year *t*+2.

Dropout_t12 takes a value of one if any of dropout_t1 or dropout_t2 takes a value of one. Along these lines, dropout_t12 can be interpreted as dropping out of school at any point within two years of completing an academic year. Likewise, dropout_t1 can be interpreted as leaving the school at any point within one year of finishing an academic year.

[16] A trade-off exists concerning how many years of historical information to use. Adding a year can improve the prediction of school dropout but reduces the sample size. I decide to use three years. Information on *t*–2 has some, albeit limited, predictive power (this will be shown in the next section) and this decision allows me to pool four different cohorts (I give complete details on this topic in the fourth subsection).

[17] Given how I measure school dropout, the sample only includes students that finished the academic year *t*.

The Classifier and Machine Learning Algorithms

I determine f using supervised MLA. MLA have expanded their range of users, from computer science to social sciences, such as economics. MLA are a powerful and flexible provider of quality predictions (Mullainathan & Spiess, 2017) and a helpful tool for prediction policy problems (Kleinberg, Ludwig, Mullainathan, & Obermeyer, 2015).[18] Research on topics such as recidivism, teachers' hiring and identification of vulnerable groups has used MLA.

Mullainathan and Spiess (2017) identify three essential characteristics of MLA. Firstly, these algorithms find functions that predict well out of sample or that do not overfit the data. Secondly, MLA can discover a complex structure that is not specified in advance. Finally, a subset of MLA allows researchers to manage high dimensional settings, the cases where the number of variables is larger than the observations (James, Witten, Hastie, & Tibshirani, 2013).

MLA are suitable in my case for three reasons. Firstly, in theory, an approach that maximises the predictions of an outcome outside of the sample is preferred for a prediction policy problem (such as determining which students will drop out) relative to an approach that maximises predictions within the sample. Secondly, *a priori* I ignore the structure of the function (for example the number of variables to include) or the form that achieves the best prediction of school dropout. Using MLA expands the likelihood of finding the best model because some MLA consider interactions and polynomials while other MLA directly address the challenge of variable selection. Finally, with MLA I can better manage the number of parameters to include in the dataset. Although I do not face a high-dimensionality problem, reducing the numbers of predictors (by not directly including higher order terms) facilitates the calculations.

To obtain predictions that work well out of sample, machine learning uses a training dataset and a test dataset. The models must be estimated in the former dataset and assessed with the latter. MLA aim to avoid overfitting, in other words, MLA seek to optimise their predictions in the test dataset (out of sample) rather than in the training dataset (in-sample). To do so, each

[18] The machine learning literature has focused mainly on the problem of prediction and not on capturing the relationship between the predictors and the outcome. Initially, MLA were not designed to obtain deep structural parameters or for causal inference (Nichols, 2018). However, there is emerging literature that connects MLA with causal inference for policy (Abadie, Athey, Imbens, & Wooldridge, 2014; Athey & Imbens, 2015a, 2015b).

algorithm first tries to determine its optimal level of complexity in the training dataset. The specific indicators of model complexity vary by algorithm, but in general terms these are called regularisers. The less regularisation there is, the better the model will predict in-sample (Mullainathan & Spiess, 2017). These parameters of model complexity can be viewed as variables that can be tuned with the purpose of producing optimal predictions in the test dataset (Varian, 2014).

The last process is known as empirical tuning. It consists of fitting the algorithm in one part of the training dataset and then determining the optimal value of the regulariser by assessing its prediction performance in another part of the training dataset (Mullainathan & Spiess, 2017). Van der Vaart, Dudoit, and Van der Laan (2006) show that the effectiveness of the procedure is increased if the training dataset is subdivided into multiple subsamples or folds. This is known as cross-validation, with 5 or 10-fold being the most adopted practices (Mullainathan & Spiess, 2017). In this type of cross-validation, the regulariser with the best average performance is chosen.

Machine learning algorithms vary regarding the flexibility they can offer to find the best f. Shrinkage methods such as lasso and elastic nets are the most restrictive as they can only generate linear functions (no interactions between the predictors or other higher order terms). These algorithms are less flexible than ordinary least squares as there is a penalty for every regression coefficient that is different from zero. This penalty leads to the coefficients of the linear regression being shrunken towards zero relative to least squares (James et al., 2013). Generalised additive models (GAMs) expand the range of shapes to estimate f from linear to more complex approaches, for example some non-linear relationships (James et al., 2013). In practice, GAMs fit a non-linear function separately for each predictor and then add all these functions. As the model is additive, interactions between the predictors are not considered.

Approaches based on trees admit interactions by stratifying the predictor space into some regions (McBride & Nichols, 2016). For example, if only two predictors of school dropout are available (age and attendance), a classification tree algorithm can be as follows: a dropout is predicted only if a student is older than 17 years old and has an attendance of lower than 70%. Methods such as random forest and boosting are the result of the combination of multiple trees.

Finally, a highly flexible approach is support vector machines. In broad terms, in a classification problem this algorithm aims to find a hyperplane separating the two classes. If this hyperplane cannot be found a kernel trick (Theodoridis & Koutroumbas, 2009) is applied. The feature space of the problem is expanded, and a new hyperplane is fitted in this transformed space. This process may produce non-linear class boundaries in the original predictors' space.

James et al. (2013) claim that no single algorithm is superior to all the others in every possible context. Thus, I try multiple MLA. For simplicity, the paper presents results for only six of them: i) elastic nets (glmnet), ii) generalised additive models (gam), iii) gradient boosting (gbm), iv) lasso, v) support vector machines (svm), and vi) random forest (rf). These six MLA use the same inputs, which are the 50 predictors I describe in Appendix A.

I implement the MLA in the software R using the Caret Package. Kuhn (2008) is a precious source for this purpose. I utilise 10-fold cross-validation. I use two test datasets. The first test dataset is useful to conduct out of sample validation while I use the second test dataset to assess the quality of the predictions over time. I explain in detail the design of the training dataset and the two test datasets in the fourth subsection. In respect to the treatment of the predictors, I convert categorical variables into dummies and the Caret Package carries out standardisation on all the predictors before executing each machine learning algorithm.

The Criterion to Select the Best *f*

Statistical classification problems have only four possible outcomes for dropout prediction. A model either: i) correctly predicts a dropout, ii) incorrectly predicts a dropout, iii) fails to predict a dropout, or iv) correctly predicts a non-dropout. More generally, these processes result in four categories that are labelled true positives, false positives, false negatives and true negatives. This can be summed up in a confusion matrix like the one I present in Table 2.1.

Table 2.1: Confusion Matrix or Contingency Table of a Classifier of Student Dropout

True Class	Predicted Class		Total
	Not a Dropout	Dropout	
Not a Dropout	True Negatives (TN)	False Positives (FP)	Non-Dropouts (ND)
Dropout	False Negatives (FN)	True Positives (TP)	Dropouts (D)
Total	Negatives (N)	Positives (P)	All Population (T)

Multiple indicators derived from combinations of the nine shaded cells of Table 2.1 have been used to report the quality of predictions. Within studies on dropout prediction there is no standard metric that facilitates comparisons (Bowers et al., 2013). Following these authors, I provide true positive rates, false positive rates and accuracy. Table 2.2 describes these three indicators in detail. The right-hand column of Table 2.2 uses information from six shaded cells of Table 2.1. An exhaustive list of this family of indicators is available in Appendix B.

Table 2.2: Indicators Used in the Predictive Model of School Dropout

Name	Formula
True Positive Rate or Sensitivity	True Positives (TP) / Dropouts (D)
False Positive Rate or 1–Specificity	False Positives (FP) / Non-Dropouts (ND)
Accuracy	[True Negatives(TN)+True Positives(TP)] / Total(T)

A perfect classifier would achieve a true positive rate of one, with all dropouts predicted as such, and a false positive rate of zero, or no incorrect predictions of dropouts. No classifier achieves this performance though. In practice, dropout prediction models intend to maximise the true positive rate (or sensitivity) and minimise the false positive rate (or 1–specificity). Nonetheless, there is a trade-off between these two indicators. As a predictive model classifies more observations as dropouts, the true positive rate increases but so does the false positive rate.

Receiver operating characteristics curves summarise this trade-off. An ROC curve simultaneously displays the false positive rate and the true positive rate given by a classifier (James et al., 2013). While the former indicator goes on the horizontal axis, the latter is located on the vertical axis. All possible outputs or scenarios provided by a classifier are represented in an ROC curve. Given this feature, the area under the ROC curve provides a measure of the overall predictive performance of a classifier. The AUC scale ranges from zero to one as the true positive rate and the false positive rate. The better the classifier is, the closer its AUC will be to 1. Conversely, a classifier making predictions at random has an expected AUC of 0.5.

The AUC is a useful indicator to compare the overall performance of multiple predictive models. Models with a higher AUC are, on average, better in statistical classification relative to models with a lower AUC. A model with an ROC curve that is on top of other curves all the way along the horizontal axis is unambiguously a better classifier in every possible scenario.

I use the AUC estimates to select the best f. I calculate these in the first test dataset (for out of sample validation) and for the three measures of school dropout introduced at the beginning of this subsection. The advantages of using the AUC are twofold. In the first instance, the performance of the MLA predicting school dropout can be compared graphically. Secondly, the AUC integrates in one value all the potential classification outputs of each algorithm. This feature frees me to select an arbitrary threshold to assess the performance of the classifiers (such as to choose the algorithm with the highest true positive rate when the false positive rate reaches 0.20).

Additionally, I provide true positive rates, false positive rates and accuracy for two specific scenarios. I force the MLA to classify 10% and 30% of students as future dropouts. These indicators help to establish comparisons with the outputs obtained by other scholars.

2.2.3 *Methods: Targeting Assessment*

After identifying the best performing algorithm, I can use two indexes to target a hypothetical CCT. The first is the income-proxy means test score from the Social Protection File. I derive the second from the outputs of the best machine learning algorithm (f). Each of these outputs represents the probability that the model is observing a future school dropout.

This subsection describes the methods related to the targeting assessment of a hypothetical CCT on the poor and future school dropouts. The first part explains how I construct the poverty variable. The following two parts elaborate on the indicators I use to assess targeting: total leakage and leaked welfare. The final part discusses the policy alternatives I utilise.

Poverty

Poverty status is not directly available from the Social Protection File dataset. However, it is possible to build an estimate of poverty status by using household structure and income (in the Social Protection File the most relevant sources of income are labour and pensions). There are many approaches to constructing this variable. I use total household income over its members. Using income per capita is consistent with the "traditional" methodology used in Chile to measure poverty. I define a student as poor if he or she is part of the poorest fifth regarding income per capita in the sample. I choose this poverty line considering the poverty rate was approximately 20% for the population analysed in my paper in one year of the assessment. This threshold is higher than the official poverty line because the sample in my study is not representative of the whole population. This is explained in more detail in the next subsection.

Total Leakage

The poverty targeting literature has offered multiple indicators of targeting effectiveness. For example, the AUC in ROC analyses are available (Baulch, 2002; Wodon, 1997). Nonetheless, one of the more common approaches to assessing the targeting effectiveness of transfers consists of providing undercoverage and leakage rates (Coady et al., 2004). The undercoverage rate is the proportion of poor households or individuals not receiving the programme. The leakage rate is the fraction of the non-poor among those who are receiving the programme.

Two common limitations are associated with using these two rates (Coady & Skoufias, 2004). The first is that they disregard distributional information; for example giving a transfer to someone in the highest 1% of income counts the same as giving it to someone marginally over the poverty line. The second shortcoming is that the size of the transfer is irrelevant. It does not make a difference whether a poor household receives a minuscule transfer or an amount that lifts it over the poverty line. One of the preferred ways to address this limitation has been to assess targeting based on the impact on poverty (Grosh & Baker, 1995; Skoufias et al., 2001).

On the one hand, using leakage and undercoverage rates restricts the depth of the analysis in the dimension of poverty. On the other hand, it facilitates establishing comparisons between targeting indicators of poverty and school dropout. Furthermore, it facilitates combining future school dropouts and the poor into one indicator. Thus, despite the limitations of leakage and undercoverage rates, I opt to use these types of indicators to assess the performance of targeting mechanisms. Five of the indicators I use in the paper are presented in Table 2.3.

Table 2.3: Indicators Used in the Targeting Assessment

Name	Formula
Poor Undercoverage	# poor not receiving CCT / # poor
Non-Poor Leakage	# non-poor receiving CCT / # receiving CCT
Dropout Undercoverage	# future dropouts not receiving CCT / # future dropouts
Non-Dropout Leakage	# non-dropouts receiving CCT / # receiving CCT
Total Leakage	# non-dropouts & non-poor with CCT / # receiving CCT

Total leakage, defined as the proportion of non-dropouts and non-poor receiving the CCT after the simulation, can be interpreted as the inclusion error. This can be better appreciated from Table 2.4. Students (potential recipients of a CCT) can be part of one of four classes. Either they are poor and will drop out of school, they are poor but will not drop out, they are not poor but will drop out of school, or they are not poor and will not drop out. Targeting is unsuccessful when a CCT is given to the fourth type of students because no target group is reached.

Table 2.4: Successful Targeting and Targeting Errors in a Context of Two Target Groups

True Class	Hypothetical CCT Recipient	
	No	Yes
Non-Poor & Non-Dropout	Successful Targeting	Inclusion Error
Non-Poor & Dropout	Exclusion Error	Successful Targeting
Poor & Non-Dropout	Exclusion Error	Successful Targeting
Poor & Dropout	Exclusion Error	Successful Targeting

The selection of total leakage as the first main indicator of my analysis is also justified on theoretical grounds. One minus leakage can be equivalent to the distributional characteristic (DC), a benefit-cost statistic used to compare the welfare impact of transfers with a common budget (Coady & Skoufias, 2004). The authors show that the DC λ for any given scheme j is:

$$\lambda_j = \sum_h \beta^h \, \theta^h,$$

where β^h is the social valuation (welfare weight) of extra income to household h and θ^h represents the share of the total programme budget received by household h.

An advantage of the DC is that welfare weights are made explicit and it generalises from simpler to more complex cases (Coady et al., 2004). When the size of the transfer is identical for each household and the social valuation of extra income is equal to one for a poor household and zero otherwise the DC indicator is equivalent to one minus the leakage rate, as shown:

$$\lambda_j = \sum_h \beta^h \, \theta = \theta \sum_h \beta^h = \frac{\sum_h \beta^h}{\#\,Recipients} = \frac{\#\,Poor\ Recipients}{\#\,Recipients}$$

$$= 1 - \frac{\#\,Non\ Poor\ Recipients}{\#\,Recipients} = 1 - Leakage$$

Under some additional assumptions, when the size of a CCT is identical for each individual and when the social valuation of income is equal to one for any CCT recipient who is either poor or a future dropout and zero otherwise, the DC indicator is equivalent to one minus total leakage.

Total leakage is the cornerstone indicator that I use in my research to compare the targeting performance of alternative instruments. The indicator has the major advantage of allowing for the integration of two important target groups for CCTs. Additionally, the logic behind this indicator is useful for other parts of the assessment, when I focus on social welfare and targeting costs.

Peyre Dutrey (2007) criticises the use of leakage in targeting assessments because this indicator does not account for individuals who are excluded. In other words, undercoverage is not considered. However, leakage and undercoverage rates are related. If coverage increases (and undercoverage decreases), leakage is likely to increase. Therefore, rather than intending to find the optimal rate of undercoverage and leakage, my targeting assessment is done for different coverage levels of a hypothetical CCT, more precisely for three budget allocations for a CCT. I explain this aspect of the paper in detail at the end of this subsection. Overall, within a fixed budget and coverage rate, the targeting mechanism with the lowest value of total leakage is optimal.

Leaked Welfare

I also analyse whether the findings of the targeting assessment hold when I change the social valuation of the target groups. Up to this point, I have implicitly assumed that successfully targeting a student who is poor is as socially equivalent worthwhile as correctly targeting a student that will drop out of school. I introduce four different scenarios of social valuation across the two target groups. In the first two scenarios each target group is twice as important as the other. Furthermore, in the last two scenarios the difference in valuation increases to four times the other target group. The choice of these scenarios does not have any theoretical justification, it is merely practical. Following the logic of the DC, the welfare impact of a transfer scheme *j,* which provides an equal amount for each individual *i* can be measured by the formula:

$$\lambda_j = \omega \sum_i \gamma^i = \frac{1}{\# \, Recipients} \sum_i \gamma^i = \frac{\sum_i \gamma^i}{\# \, Recipients} \ ,$$

where ω is the share of the total programme budget received by each adolescent who is a CCT recipient and γ^i is the social valuation (welfare weight) of extra income to adolescent *i*.

In order for λ_j to have minimum and maximum values of zero and one, I choose the welfare weights using the following logic. A hypothetical CCT recipient i who is neither a future dropout nor poor receives a value of γ of zero. Conversely, each CCT recipient i who belongs to the highly valued target group receives a value of γ of one. The social valuations of each class of student in each of the four scenarios I use in the paper are presented in Table 2.5.

Table 2.5: Social Valuation (Welfare Weights γ) in Different Scenarios

	Social Valuation Scenarios			
True Class	The Poor are Twice as Important as Dropouts	The Poor are Four Times More Important than Dropouts	Dropouts are Twice as Important as the Poor	Dropouts are Four Times More Important than the Poor
Non-Poor & Non-Dropout	0	0	0	0
Non-Poor & Dropout	0.5	0.25	1	1
Poor & Non-Dropout	1	1	0.5	0.25
Poor & Dropout	1	1	1	1

The targeting mechanism j that provides the highest λ_j maximises welfare. Given the weights I use, the last statement can be rephrased as follows: the targeting mechanism j that provides the lowest 1–λ_j maximises welfare (for any given budget). This last indicator is the focus of the welfare assessment. For simplicity, I refer to it as “leaked welfare”. More formally:

$$Leaked\ Welfare = 1\text{–}\lambda_j = 1 - \frac{\sum_i \gamma^i}{\#\ Recipients}$$

Budget Available, CCT Coverage and Targeting Mechanisms

CCTs are not universal schemes. Stampini and Tornarolli (2012) show the coverage of CCTs varied by year and country in Latin America. Consequently, I repeat my targeting assessment for different levels of coverage for a hypothetical CCT. Given that in my study the transfer size remains unchanged, an increase or decrease in the CCT programme budget only affects its coverage. For this reason, I repeat my targeting assessment for different budget scenarios available for a hypothetical CCT. I assume that there are no administrative costs in the first instance. I analyse three CCT coverage or budget scenarios. In the first case, the budget allows for reaching only 5% of the students in the sample. In the second and third scenarios the budget allows for reaching 20% and 40% of the sample, respectively. These three cases aim to recreate real policy environments: i) a narrowly targeted CCT, ii) a CCT whose coverage is in line with the population living in poverty (such as in Chile), and iii) a broadly targeted CCT.

I begin by looking at the targeting performance separately for each instrument. First, I only use the proxy means test score of the Social Protection File. Second, I use the predictions derived from the best f. The assessment continues with two combined mechanisms. I target a hypothetical CCT assigning the first 25% of the budget available using the PMT score, and then this percentage ascends to 75%. For example, when the budget allows for reaching 20% of the students in the sample with a CCT and I allocate this using a combined approach, such as 75% is assigned first using the SPF, the procedure works as follows. In the first place, I select the 15% with the lowest PMT scores among the sample and I assign them the CCT. Consequently, I choose the remaining 5% of the students by observing the highest likelihood of dropping out among those not selected in the first step.

2.2.4 *Sample and Dataset Structure*

The sample excludes students below seventh grade in year *t*, younger than 12 years old by June of year *t* and older than 21 years old by March of year *t*+1. I apply these restrictions considering that student dropout in Chile is a cause for concern mainly in secondary school and that 21 years old is the maximum age to stay enrolled in traditional secondary education.[19]

Another crucial characteristic of the sample is that it includes only adolescents that had an entry in the SPF. Thus, this is not a representative sample of the population, as those households with the highest earnings were less likely to request an SPF. These features do not favour making inferences about the whole student body. However, this is not problematic if the findings are linked to a subset of the entire population: those students with an SPF. Given that this subset is more likely to be recipients of social programmes, the findings of this study remain relevant.

[19] In the Chilean educational system, students are in theory expected to graduate from secondary education at the age of 18. However, grade repetition and school dropout can delay graduation from secondary studies.

In practice, 26,3% of adolescents did not make it to my final sample due to lacking an SPF.[20]

To undertake the machine learning algorithms and the targeting assessment I structure the dataset based on four year-cohorts *t* (*t* = 2011, 2012, 2013, 2014), using information from *t*–2, *t*–1, *t*, *t*+1 and *t*+2 for each individual in the cohort. Hence, each cohort on its own is a panel dataset. I pool these four cohorts to obtain the "full dataset". As a result, this dataset contains observations from eight years (from 2009 until 2016). This is explained in detail in Table 2.6.

Table 2.6: Dataset Structure

Cohort	Academic and School Information (From the Ministry of Education)	SPF Info	Dropout Information (From the Ministry of Education)
2011	2009 (*t*–2), 2010 (*t*–1) & 2011 (*t*)	2011 (*t*)	2012 (*t*+1) and/or 2013 (*t*+2)
2012	2010 (*t*–2), 2011 (*t*–1) & 2012 (*t*)	2012 (*t*)	2013 (*t*+1) and/or 2014 (*t*+2)
2013	2011 (*t*–2), 2012 (*t*–1) & 2013 (*t*)	2013 (*t*)	2014 (*t*+1) and/or 2015 (*t*+2)
2014	2012 (*t*–2), 2013 (*t*–1) & 2014 (*t*)	2014 (*t*)	2015 (*t*+1) and/or 2016 (*t*+2)

I divide the "full dataset" into two parts. The "old" subset contains cohorts 2011, 2012 and 2013. I partition the "old" subset using random assignment into a training dataset and a test dataset. Each observation in the "old" subset has a 0.75 probability of ending up in the training dataset. In this last dataset the MLA are trained. I test the algorithms and implement the targeting assessment in the test dataset. The "new" subset contains the 2014 cohort, which I use to assess the quality of the predictions of school dropout over time. This process is called out of time validation and its results are not shown in the body of the paper but in an appendix.

2.3 Results: Predictive Model of School Dropout

The first subsection provides summary statistics of school dropout and for multiple variables included in the model. The second part focuses on the results of the MLA predicting school dropout. I provide ROC curves, their AUC, true positive rates, false positive rates and accuracy for three measures of school dropout. This subsection additionally analyses which variables of

[20] Students without a Social Protection File were more likely to attend a private non-subsidised school and less likely to be enrolled in a public school or secondary technical-professional education relative to their peers with a Social Protection File. The former group was also more likely to be enrolled in schools with higher performance in SIMCE (the national standardised test) and less likely to attend schools in rural areas. This information provides evidence that students from higher-income households are not well represented in the sample as in Chile these households are more likely to prefer private schools, scientific-humanistic education, enrol their children in schools with higher levels of academic performance (measured by SIMCE), and to live in urban areas. However, given that the percentage of students attending rural schools and technical-professional education is not negligible is also feasible that some poor households did not have access to a Social Protection File.

the model are the ones that mostly explain the variation in school dropout.

2.3.1 *Summary Statistics*

Table 2.7 provides summary statistics for some individual-level variables. Panel A presents the dropout rates in years *t*+1 and/or *t*+2. Panel B describes academic variables for years *t* and *t*–1. Panel C introduces variables related to the schools where adolescents were enrolled in year *t*. Panel D presents the information provided by the SPF in year *t*. The first four columns of the table summarise the mean values of each variable for each cohort. The last four columns provide the mean, standard deviation, minimum and maximum values for cohorts 2011, 2012 and 2013. These three cohorts comprise the "old" subset, as explained in the last subsection.

The average dropout rates for years *t*+1 and *t*+2 are 0.06 and 0.09 respectively for the 2011-2013 period. Within this time range 11 out of 100 adolescents dropped out either in year *t*+1 or *t*+2. All measures of dropout declined annually from 2011 to 2014: i) from 0.07 to 0.05 for adolescents dropping out in year *t*+1, ii) from 0.10 to 0.07 for adolescents dropping out in year *t*+2, and iii) from 0.12 to 0.09 for adolescents dropping out either in year *t*+1 or in year *t*+2. The last two rows of Panel A illustrate the dynamics of dropout. On average, for 2011, 2012 and 2013 cohorts, 65 out of 100 adolescents who dropped out in year *t*+1 did not return to school in year *t*+2. Among those who were dropouts in year *t*+2*,* only 48 out of 100 adolescents dropped out in year *t*+1. Along these lines, 52 out of 100 dropped out exactly in year *t*+2.

Regarding their academic information in year *t*, between 2011 and 2013: i) adolescents had an average grade of 5.30, ii) their attendance was 89.7%, iii) 9 out of 10 students were promoted to the next grade, and iv) their mobility (the rate of students switching school between *t*–1 and *t*) was 0.24.[21] The average grade and the rate of students promoted marginally increased from 2011 to 2014.

Between 2011 and 2013, 4 out of 10 adolescents attended traditional primary education in year *t* while 35% and 21% of adolescents were enlisted in traditional secondary education, in scientific-humanistic (SH) and technical-professional (TP) schools respectively. Within this

[21] Approximately three out of five cases of mobility in my sample are explained by students switching schools between eighth and ninth grade (the transition between primary and secondary education in Chile). The mobility rate among ninth graders in year *t* reaches 0.66 in total and is as high as 0.81 among students in public schools.

period, 49% and 46% of adolescents were enrolled in private subsidised and public schools.

According to the information provided by the SPF, between 2011 and 2013, on average: i) adolescents were 15.39 years old by the end of the academic year *t*, ii) half of the students were males, and iii) 9 out of 100 were labelled as from indigenous backgrounds. Concerning the head of their households: i) 45% were females, ii) 59% lived with a partner, iii) 41% were employed and contributing to social security, and iv) their average schooling was 9.59 years. Between 2011 and 2013, the average number of people within each student household reached 4.26 with these divided among 2.15 rooms. The average monthly income per capita reached $61,012 CLP (equivalent to $116.5 USD at the December 30th, 2013 exchange rate).

Table 2.7: Summary Statistics

Variables	Year-Cohort							
	2011	2012	2013	2014	2011-2013			
	Mean	Mean	Mean	Mean	Mean	Std. Dev.	Min.	Max.
Panel A: Drop Out Information								
Dropout Year *t*+1	0.07	0.06	0.06	0.05	0.06	0.25	0	1
Dropout Year *t*+2	0.10	0.09	0.08	0.07	0.09	0.28	0	1
Dropout Year *t*+1 or *t*+2	0.12	0.11	0.10	0.09	0.11	0.31	0	1
Dropout Year *t*+2 (if Dropout Year *t*+1=1)	0.67	0.64	0.63	0.62	0.65	0.48	0	1
Dropout Year *t*+1 (if Dropout Year *t*+2=1)	0.49	0.48	0.47	0.47	0.48	0.50	0	1
Panel B: Academic Information								
Average Grade Year *t*	5.27	5.31	5.32	5.36	5.30	0.68	1	7
Attendance (%) Year *t*	88.1	90.9	90.3	90.4	89.7	11.6	1	100
Promoted Year *t*	0.88	0.91	0.91	0.92	0.90	0.31	0	1
Mobility Year *t*	0.25	0.24	0.24	0.23	0.24	0.43	0	1
Average Grade Year *t*–1	5.41	5.37	5.40	5.41	5.39	0.63	1	7
Attendance (%) Year *t*–1	92.8	90.1	92.1	91.5	91.7	9.3	1	100
Promoted Year *t*–1	0.94	0.92	0.94	0.94	0.93	0.26	0	1
Mobility Year *t*–1	0.24	0.24	0.23	0.23	0.23	0.42	0	1

Table 2.7 (continued): Summary Statistics

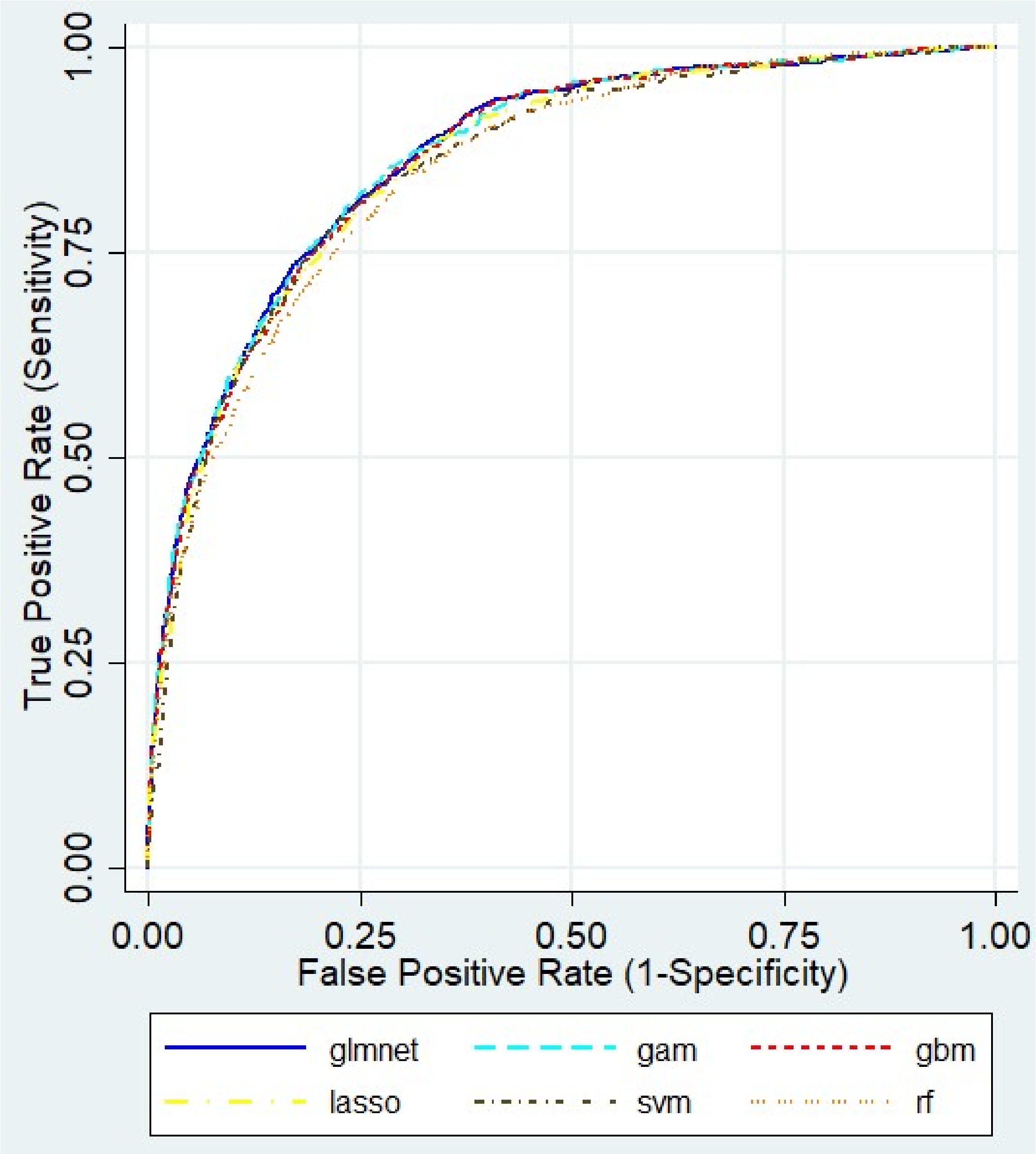
	Year Cohort							
Variables	2011	2012	2013	2014	2011-2013			
	Mean	Mean	Mean	Mean	Mean	Std. Dev.	Min.	Max.
Panel C: School Information (Year t)								
Primary Traditional School	0.40	0.41	0.41	0.41	0.40	0.49	0	1
Secondary Traditional SH School	0.34	0.35	0.36	0.36	0.35	0.48	0	1
Secondary Traditional TP School	0.23	0.21	0.20	0.20	0.21	0.41	0	1
Public School	0.48	0.46	0.45	0.44	0.46	0.50	0	1
Private Subsidised School	0.48	0.50	0.50	0.51	0.49	0.50	0	1
Number of Students in the Academic Cohort	117.1	110.8	108.3	106.1	112.1	103.2	1	1324
Rural School	0.08	0.08	0.07	0.07	0.07	0.26	0	1
Dropout Rate in Previous Academic Cohort	0.07	0.07	0.06	0.06	0.07	0.10	0	1
School Language Score SIMCE	249.8	251.5	251.1	248.7	250.8	26.3	142	345
School Maths Score SIMCE	248.7	249.6	253.7	255.5	250.7	32.7	138	381
Panel D: Social Protection File Information								
Age (Years) at the End of Academic Year *t*	15.39	15.39	15.39	15.41	15.39	1.65	12.50	20.67
Male	0.50	0.50	0.50	0.50	0.50	0.50	0	1
Indigenous Background	0.08	0.09	0.09	0.09	0.09	0.28	0	1
Household Number of Rooms	2.18	2.15	2.11	2.10	2.15	0.93	0	63
Head of Household Lives with a Partner	0.60	0.59	0.58	0.57	0.59	0.49	0	1
Head of Household (HH) is Female	0.44	0.45	0.46	0.48	0.45	0.50	0	1
HH Years of Schooling	9.46	9.62	9.70	9.83	9.59	3.41	0	24
HH Employed and with Social Security	0.41	0.41	0.41	0.41	0.41	0.49	0	1
Household Size	4.30	4.24	4.24	4.21	4.26	1.45	2	34
Household Income per Capita ($CLP)	58,095.0	61,848.8	63,169.2	70,188.6	61,011.9	53,274.3	0	5,533,636
Social Protection File PMT Score	7,537.2	7,384.2	7,193.6	7,025.8	7,373.0	3,733.3	2,072	15,625
Number of Observations	960,514	930,225	937,843	968,584	2,828,582			

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

2.3.2 *Results for Models Predicting School Dropout*

Figure 2.1 presents the ROC curves for six MLA predicting dropout_t12. A solid line plots the elastic net algorithm (glmnet). The generalised additive (gam) and boosted trees (gbm) models are plotted by long and short dashes, respectively. Dots and dashes plot the other shrinkage algorithm (lasso) and support vector machines (svm), long dashes in the former case and short dashes for the latter. The random forest model (rf) is plotted by dots only.

Figure 2.1: ROC Curve for Models Predicting School Dropout in Year *t*+1 or *t*+2



Source: own calculations using administrative datasets, Chilean ME & MSD

According to Figure 2.1, the curves of the six models are close to each other and no single one is above or below the rest along the whole horizontal axis. This suggests that the six MLA have minor differences in terms of the area under the ROC curve. A closer look at the graph shows that the solid line (representing the glmnet model) has a higher degree of convexity. This curve tends to be above all the other curves in a broad range of false positive rates. Conversely, the random forest curve (comprised exclusively by dots) is below the others in some sections of the graph (for example where the true positive rate lies between 0.50 and 0.75).

Figure 2.2 and Figure 2.3 focus on the ROC curves predicting dropout_t1 and dropout_t2. The patterns of the lines follow the same logic as for Figure 2.1. On the one hand, both figures show that the solid curves (glmnet) are predominantly above the other curves. However, the short-dashed curves (gbm) and the long-dashed curves (gam) closely follow and even overpass the solid curves along some parts of the horizontal axis. On the other hand, the svm algorithm (the short dashed and dotted curves) is unambiguously the worst performer in these assessments.

Table 2.8 presents the area under the ROC curve for each of the six MLA. The first column of the table focuses on adolescents who dropped out in either of the two years after *t*. The last two columns of the table address the predictions of dropouts in years *t*+1 and *t*+2*,* respectively.

Table 2.8: Area Under the ROC Curve for Models Predicting School Dropout

Machine Learning Algorithms	School Dropout Measures		
	Dropout in Year *t*+1 or *t*+2	Dropout in Year *t*+1	Dropout in Year *t*+2
glmnet	0.866	0.893	0.857
gam	0.865	0.892	0.854
gbm	0.863	0.891	0.851
lasso	0.858	0.885	0.845
svm	0.853	0.843	0.803
rf	0.849	0.875	0.844

Source: own calculations using administrative datasets, Chilean ME & MSD

The elastic net algorithm has the largest AUC in all three cases. In the test dataset, glmnet reaches 0.866 for dropout_t12, 0.893 for dropout_t1 and 0.857 for dropout_t2. Generalised additive models (gam) reach the second highest area under the curve in all three cases. The third highest AUC out of the sample is provided by the boosted trees algorithm (gbm). More generally, for these three algorithms the AUC is above 0.860 in the classification of dropout within two years, 0.890 in the classification of dropout after one year and over 0.850 in the second-year dropout classification. Conversely, random forest (rf) and support vector machines (svm) algorithms have the two worst performances in all three measures of school dropout.

Table 2.9 helps the reader to understand how the performance of these models translates into targeting effectiveness. The table presents the true positive rate, the false positive rate and the accuracy of the MLA. I need to set a common threshold to be able to compare the MLA among these three measures. The first three columns provide these indicators for a scenario where I classify as future dropouts 10% of adolescents with the highest probability of being a future dropout. In the last three columns the scenario considered is 30% of adolescents being classified as school dropouts. Appendix C shows the results of Table 2.8 and Table 2.9 in the second test dataset, the one that assesses the predictions over time (out of time validation).

The left part of Panel A in Table 2.9 shows that the glmnet algorithm has the best performance. This model finds future dropouts at a rate of 477 out of 1000 in the scenario where 10% of adolescents are classified as dropouts. Additionally, its false positive rate is 0.053. In other words, non-dropouts are incorrectly classified as dropouts at a rate of 53 cases out of 1000. Finally, this algorithm successfully classifies 89.5% of the students. The second and third best performing models in the first scenario are gam and gbm, consistently with the AUC ranking. The true positive rates in these cases are 0.474 and 0.471, respectively. The false positive rate and the accuracy indicators are the same for both algorithms, reaching 0.054 and 0.894 respectively. The two algorithms with the lower targeting performance are rf and svm. The first of these algorithms finds future dropouts at a rate of 438 out of 1000 and misclassifies non-dropouts at a rate of 57 cases out of 1000. These results are consistent with the ROC curves.

In the right-hand part of Panel A, where 30% of adolescents are classified as dropouts, the best performance belongs to the gam algorithm. In this context, dropouts are found at a rate of 810 out of 1000. However, the false positive rate and the accuracy of the model weakens. The first indicator reaches 0.237 while the second reaches 0.768. The algorithm based on elastic nets (glmnet) has the second-best performance after gam among the true positive rate and accuracy.

From the left part of Panel B, gam has the best performance in the classification of first-year dropouts in the scenario where 10% of adolescents are classified as dropouts. In this model, future dropouts are identified at a rate of 584 out of 1000, while its false positive rate is 0.066. For the glmnet model the true positive rate is 0.567 and the algorithm incorrectly classifies non-dropouts at a rate of 67 out of 1000. The worst performing model (svm) in this exercise has a true positive rate of 0.494 and a false positive rate of 0.072. The last three columns of Panel B present the results for the second scenario, where 30% of adolescents are classified as dropouts after the first year. Boosted trees (gbm), glmnet and gam achieve the three highest values of sensitivity. These models reach the second, third and fourth lowest false positive rates and are among the four highest-ranked accuracies.

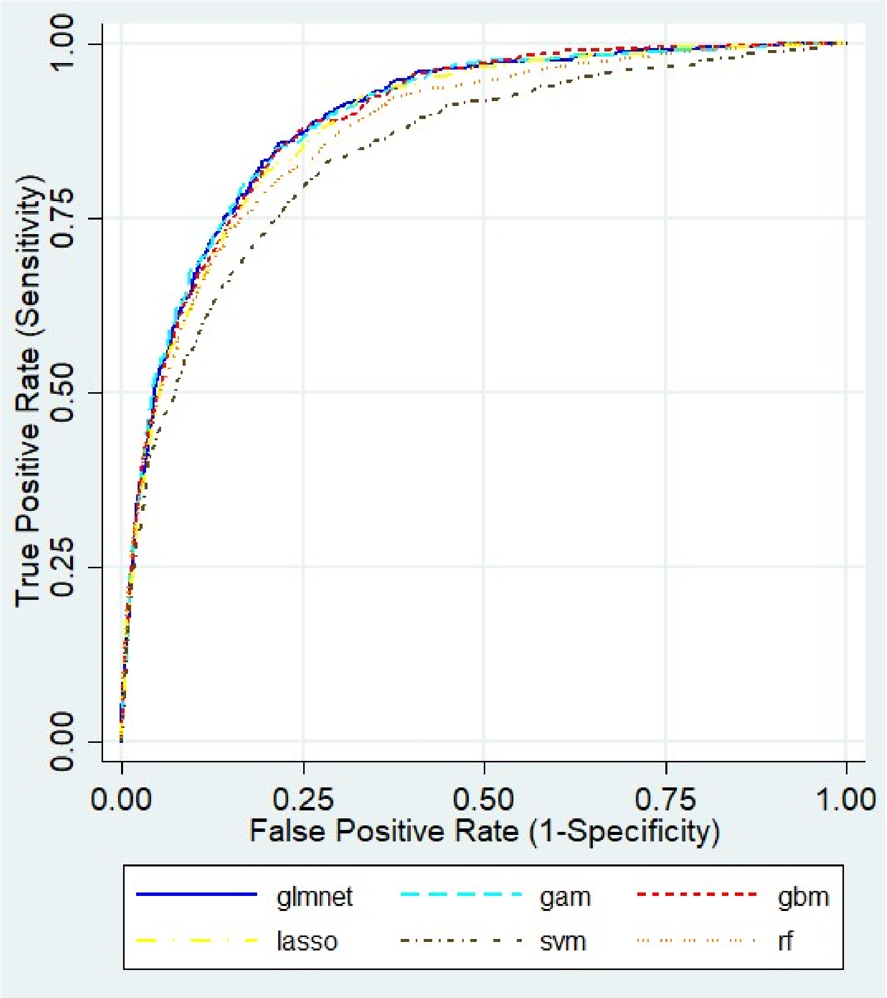
Panel C presents the sensitivity, false positive rate and accuracy for each algorithm predicting dropout in year *t*+2. In the first scenario, the random forest, elastic net and generalised additive models are the highest performers. In the second scenario, the boosted trees algorithm replaces the random forest algorithm among the group with the highest true positive rate.

Figure 2.4 and Figure 2.5 show the variable importance in the prediction of school dropout for glmnet and gbm. 13 variables can be found among the 20 most important in both algorithms.[22] Among the 5 most important in either of the models I find: age, average grade in years *t* and *t*–1, attendance in year *t*, relative average grade and attendance in year *t*, the grade (seventh to twelfth) at which the student is in year *t* and the previous average rate of dropout in the school. Income per capita plays a minor role in helping school dropout prediction in these two models.

Overall, differences in performance across the models exist but these are small in magnitude. In general, glmnet, gam and gbm are the top performers, while svm shows the worst results. The best MLA produce adequate predictions of school dropout. Regarding the true positive and false positive rates, my results are better or in the same region as 107 out of the 110 dropout flags analysed by Bowers et al. (2013). The results provided by glmnet are better than the ones obtained in Guatemala and Honduras (Adelman et al., 2017). The accuracy levels shown in the left part of Table 2.9, around 90%, are equivalent to the results obtained by the best performing MLA tested in North Carolina (Sorensen, 2018). My AUC findings are in line with the best-performing models of school dropout tested in Wisconsin (Knowles, 2015), where most of the algorithms have an AUC of between 0.860 and 0.870. However, these results are below the areas under the curve of 0.948 and 0.965 observed in Denmark (Sara et al., 2015). The policy implications of these results are discussed in the concluding section of the paper.

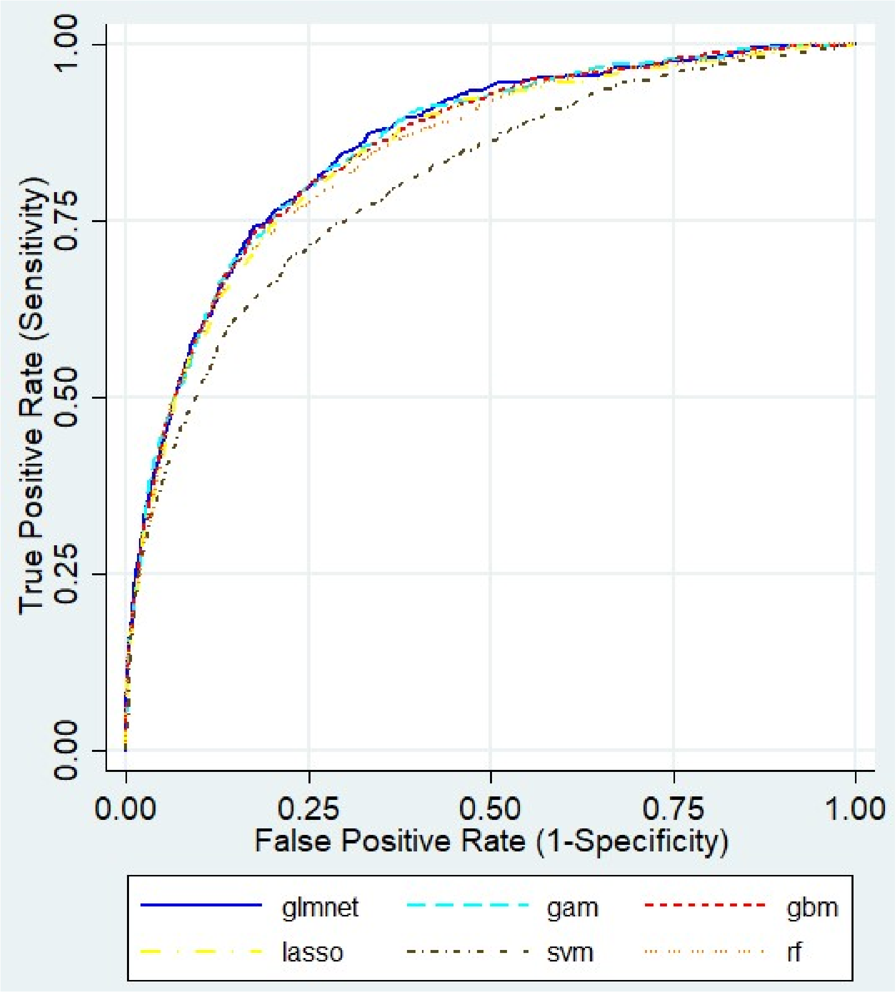
[22] In the Caret Package the maximum value of variable importance is 100. The procedure to estimate the variable importance varies by approach. For example, algorithms based on trees require permuting predictors to assess their accuracy while linear models utilise the absolute value of the *t*-statistic of each coefficient in the regression.

Figure 2.2: ROC Curve for Models Predicting School Dropout in Year *t*+1



Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Figure 2.3: ROC Curve for Models Predicting School Dropout in Year *t*+2



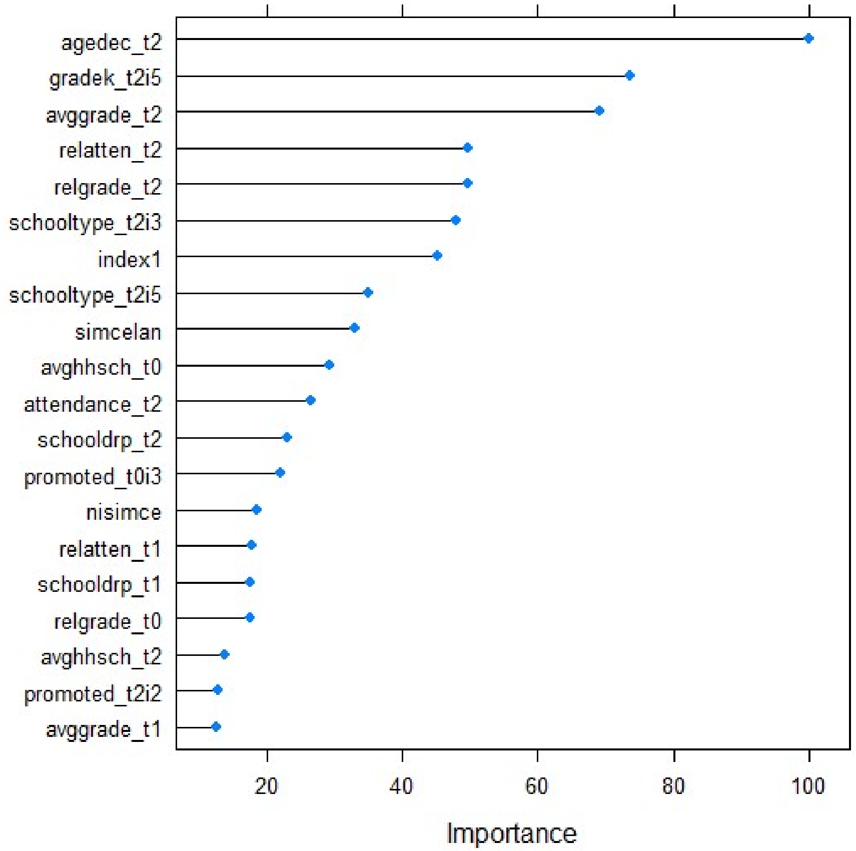
Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Table 2.9: True Positive Rate (Sensitivity), False Positive Rate (1–Specificity) and Accuracy for Models Predicting School Dropout

Machine Learning Algorithms	Scenario 1: 10% of Adolescents Classified as Dropouts			Scenario 2: 30% of Adolescents Classified as Dropouts		
	True Positive Rate (Sensitivity)	False Positive Rate (1–Specificity)	Accuracy	True Positive Rate (Sensitivity)	False Positive Rate (1–Specificity)	Accuracy
Panel A: Dropout in Year t+1 or t+2						
glmnet	0.477	0.053	0.895	0.803	0.238	0.767
gam	0.474	0.054	0.894	0.810	0.237	0.768
gbm	0.471	0.054	0.894	0.794	0.239	0.765
lasso	0.461	0.055	0.891	0.789	0.239	0.764
svm	0.449	0.057	0.889	0.801	0.238	0.766
rf	0.438	0.057	0.887	0.770	0.236	0.765
Panel B: Dropout in Year t+1						
glmnet	0.567	0.067	0.909	0.879	0.259	0.750
gam	0.584	0.066	0.911	0.874	0.260	0.749
gbm	0.561	0.068	0.908	0.881	0.259	0.750
lasso	0.561	0.068	0.908	0.861	0.261	0.747
svm	0.494	0.072	0.899	0.807	0.264	0.740
rf	0.539	0.069	0.905	0.833	0.252	0.754
Panel C: Dropout in Year t+2						
glmnet	0.483	0.064	0.897	0.800	0.252	0.752
gam	0.486	0.063	0.898	0.800	0.252	0.752
gbm	0.481	0.064	0.897	0.799	0.253	0.752
lasso	0.476	0.064	0.896	0.794	0.253	0.751
svm	0.434	0.068	0.889	0.718	0.260	0.738
rf	0.491	0.063	0.899	0.773	0.250	0.752

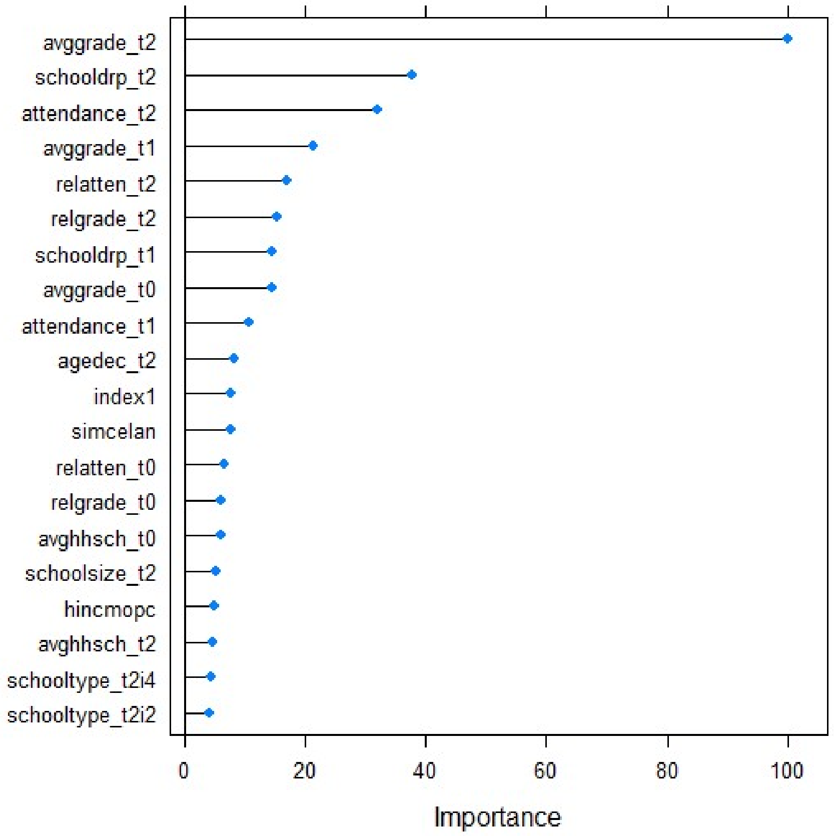
Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Figure 2.4: Variable Importance for glmnet Algorithm



Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Figure 2.5: Variable Importance for gbm Algorithm



Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

2.4 Results: Targeting Assessment

This section presents the results of the targeting assessment. The first part provides summary statistics. These statistics describe the relationship between poverty and school dropout and between each targeting mechanism and the last two outcomes. The second subsection presents the results for total leakage. The concluding part introduces the results for leaked welfare.

2.4.1 *Summary Statistics*

Table 2.10 provides bivariate summary statistics between targeting mechanisms (organised in quintile groups) and the outcomes of the targeting assessment. The outcomes, poverty status and school dropout, are presented in the columns. I offer both the mean value and the relative frequency. As poverty status and school dropout are dichotomous variables (they are either zero or one), the mean value can be interpreted as the proportion of poor adolescents and school dropouts within each quintile group, respectively. Contrarily, the relative frequency describes the distribution of poor adolescents and future school dropouts among the quintile groups.

Panel A describes the relationship between income quintile groups, poverty and school dropout in the test dataset. The rows of Panel A present the five income groups. The first group represents the bottom fifth of adolescents regarding household income per capita. As I define poverty by being in the first household income per capita quintile group of the sample, Panel A additionally partly reveals the relationship between the two outcomes of this assessment.

Panels B and C of Table 2.10 follow a similar logic to Panel A. I present the mean value and the relative frequency of poor adolescents and future school dropouts by quintile groups. Panel B focuses on the quintile groups of PMT scores in the Social Protection File. I assign adolescents in the bottom fifth of SPF scores to the first quintile group. In contrast, the quintile groups in Panel C are related to the predictions of the best performing algorithm in the previous section (glmnet). I assign adolescents with the 20% highest probability of dropping out of school to the first quintile group. Conversely, I allocate students with the lowest probability of dropping out (or a higher probability of remaining at school) to the fifth quintile group.

The measure of school dropout I present is dropout_12. Consistently, Panel C uses the quintile groups from the glmnet model predicting adolescents that leave school in years t+1 or t+2.

Summary statistics for the other two measures of school dropout are available in Appendix D.

Panel A of Table 2.10 shows that a negative correlation exists between household income per capita and dropping out. The proportion of adolescents who leave school at any time within two years declines from the first income quintile group to the fifth. On average: i) 15 out of 100 adolescents left school from the 20% in the lowest income group in the sample, and ii) only 7 out of 100 adolescents dropped out of school when they belonged to the top 20% in terms of income in the sample. Regarding the relative distribution of future school dropouts among the income per capita quintile groups, 27.97% of adolescents who dropped out belonged to the first income quintile group in the sample. Only 12.13% of school dropouts were from the 20% with the highest income in the sample. Accordingly, it is possible to learn that there is not a big overlap between poor adolescents and future school dropouts in my sample.

Given these results, it is likely that a targeting instrument designed to find one specific group (such as the poor) will have a lower capacity to identify the other group (school dropouts).

Table 2.10: Mean and Relative Frequency of Poor and School Dropout by Quintile Groups

Quintile Groups	Poor		Dropout *t*+1 or *t*+2	
	Mean	Relative Frequency (%)	Mean	Relative Frequency (%)
Panel A: By Quintile Groups of Income per Capita				
1	1	100	0.15	27.97
2	0	0	0.13	23.05
3	0	0	0.11	19.81
4	0	0	0.09	17.04
5	0	0	0.07	12.13
Total	0.20	100	0.11	100
Panel B: By Quintile Groups of SPF Scores				
1	0.44	44.03	0.14	25.59
2	0.32	32.28	0.13	23.38
3	0.16	16.29	0.12	20.94
4	0.06	5.93	0.10	17.38
5	0.01	1.47	0.07	12.71
Total	0.20	100	0.11	100
Panel C: By Quintile Groups of the Predictive Model of School Dropout (glmnet)				
1	0.29	28.81	0.38	69.63
2	0.24	24.06	0.10	18.76
3	0.20	20.18	0.04	7.46
4	0.16	16.39	0.02	3.11
5	0.11	10.56	0.01	1.05
Total	0.20	100	0.11	100

Source: own calculations using administrative datasets, Chilean ME & MSD

As in Panel A, Panel B shows a negative correlation between SPF scores and leaving school. To illustrate, 14 out of 100 students in the bottom 20% of SPF scores dropped out but only 7 out of 100 did so among those in the fifth quintile group of SPF scores. Also, there is an inverse relationship between PMT scores and poverty. For example, 44 out of 100 adolescents in the first quintile group of the SPF are poor, while only 1 out of 100 students within the 20% highest PMT scores belongs to the 20% with the lowest income in the sample. Regarding relative frequencies, 25.59% of those adolescents who dropped out belong to the first SPF quintile group and only 12.71% of dropouts are from the fifth quintile group of SPF scores.

The SPF score is a better tool for finding poor adolescents than for finding future dropouts. Among the first quintile group of PMT scores 44.03% of poor adolescents can be found. Only 25.59% of future dropouts are located in the lowest quintile group of the PMT scores. These findings are not explained by problems in the SPF model, but rather by the low overlap between poverty and school dropouts in my sample. In fact, the relative frequencies of school dropout by quintile groups shown in Panel B are similar in magnitude to the ones presented in Panel A.

Panel C shows that the predictive model is a more effective tool to find future dropouts than poor adolescents. 69.63% of school dropouts can be found in the first quintile group of the predictive model. Conversely, only 28.81% of the poor are distributed among the 20% with the highest likelihood of dropping out in my sample. Regarding absolute values, there are more future dropouts than poor students in the first quintile group of the predictive model. The latter case occurs despite the population of future dropouts being smaller relative to poor adolescents.

I extract two key findings from Panels B and C. Firstly, to find poor adolescents the SPF score is better equipped (relative to machine learning outputs). In other words, using the PMT is more income progressive than using the predictive model of school dropout. Within the first quintile group of the SPF 44 out of 100 students are poor while in the first quintile group of the predictive model only 29 out of 100 students are poor. Secondly, to find future dropouts the PMT is less effective. Among the first quintile group of SPF scores 14 out of 100 adolescents dropped out. In contrast, in the first quintile group of the predictive model 38 out of 100 adolescents left school. Thus, prioritising the use of SPF scores to target a CCT increases the effectiveness in terms of finding the poor but decreases the capacity to find future dropouts.

2.4.2 *Targeting Assessment: Total Leakage*

This subsection presents the central results of the targeting assessment. For simplicity, the evaluation focuses on one measure of school dropout. This is the indicator that captures whether an adolescent dropped out in year *t*+1 or year *t*+2 (dropout_t12). Thus, I use the outputs of the best machine learning algorithm predicting dropout_t12 as a targeting mechanism (glmnet). The results for the other two measures of school dropout are available in an appendix.

Table 2.11 shows the results for the first two (out of four) targeting mechanisms. The left side of the table represents a mechanism based solely on the proxy means test score of the Social Protection File. Conversely, the right-hand side of the table reproduces the results for a mechanism based exclusively on the outputs of the machine learning algorithm.

Table 2.11: Targeting Indicators by Independent Approach and Budget Available

Targeting Based 100% in SPF Score		Targeting Based 100% in Predictive Model	
Panel A: The Budget Allows a CCT to Reach 5% of Adolescents			
Poor Undercoverage	0.867	Poor Undercoverage	0.922
Non-Poor Leakage	0.470	Non-Poor Leakage	0.689
Dropout Undercoverage	0.934	Dropout Undercoverage	0.696
Non-Dropout Leakage	0.855	Non-Dropout Leakage	0.329
Total Leakage	0.412	Total Leakage	0.232
Panel B: The Budget Allows a CCT to Reach 20% of Adolescents			
Poor Undercoverage	0.560	Poor Undercoverage	0.712
Non-Poor Leakage	0.560	Non-Poor Leakage	0.712
Dropout Undercoverage	0.744	Dropout Undercoverage	0.304
Non-Dropout Leakage	0.859	Non-Dropout Leakage	0.615
Total Leakage	0.493	Total Leakage	0.444
Panel C: The Budget Allows a CCT to Reach 40% of Adolescents			
Poor Undercoverage	0.237	Poor Undercoverage	0.471
Non-Poor Leakage	0.618	Non-Poor Leakage	0.736
Dropout Undercoverage	0.510	Dropout Undercoverage	0.116
Non-Dropout Leakage	0.865	Non-Dropout Leakage	0.756
Total Leakage	0.544	Total Leakage	0.563

Source: own calculations using administrative datasets, Chilean ME & MSD

Table 2.11 shows that a trade-off exists between finding the poor and future dropouts. Under any budget scenario, poor undercoverage and non-poor leakage increase when switching from SPF scores to the predictive model. For example, when the budget allows for providing the CCT to 5% of adolescents in the sample, poor undercoverage is 0.867 while non-poor leakage is 0.470 if I only use the SPF for targeting. If I use the predictive model, these indicators increase to 0.922 and 0.689, respectively. Conversely, undercoverage of dropouts and leakage of non-dropouts decrease when the output of glmnet replaces the PMT. To illustrate, when the budget allows for providing the CCT to 20% of students in the sample, dropout undercoverage is 0.744 and non-dropout leakage is 0.859 when targeting is based on the SPF. If I use the predictive model for targeting these indicators drop to 0.304 and 0.615, respectively.

An additional trade-off is related to expenditure. The more of the budget that is spent, the lower undercoverage becomes among both target groups. However, leakage rates also increase. Another interesting finding is that the optimal targeting mechanism depends on the budget available. In the first two budget scenarios (in Panel A and B) total leakage is higher if the SPF score is used (relative to the predictive model) for targeting. However, when the budget allows the CCT to reach 40% of adolescents in the sample, total leakage is lower if the PMT is used.

Table 2.12 switches from independent to combined mechanisms for targeting. As explained in subsection 2.2.3, I test the targeting performance of two mechanisms that use information from both sources. The left side of the table presents the results for the mechanism where I first allocate 25% of the budget based upon the PMT score (and the remaining 75% is given based on the machine learning algorithm). Conversely, the right-hand side of the table presents the setting where I apportion the first 75% of the budget through the SPF.

Table 2.12: Targeting Indicators by Combined Approach and Budget Available

Targeting Based 25% in SPF Score		Targeting Based 75% in SPF Score	
Panel A: The Budget Allows a CCT to Reach 5% of Adolescents			
Poor Undercoverage	0.882	Poor Undercoverage	0.852
Non-Poor Leakage	0.530	Non-Poor Leakage	0.408
Dropout Undercoverage	0.740	Dropout Undercoverage	0.854
Non-Dropout Leakage	0.426	Non-Dropout Leakage	0.679
Total Leakage	0.158	Total Leakage	0.232
Panel B: The Budget Allows a CCT to Reach 20% of Adolescents			
Poor Undercoverage	0.665	Poor Undercoverage	0.609
Non-Poor Leakage	0.665	Non-Poor Leakage	0.609
Dropout Undercoverage	0.349	Dropout Undercoverage	0.524
Non-Dropout Leakage	0.640	Non-Dropout Leakage	0.737
Total Leakage	0.419	Total Leakage	0.442
Panel C: The Budget Allows a CCT to Reach 40% of Adolescents			
Poor Undercoverage	0.441	Poor Undercoverage	0.298
Non-Poor Leakage	0.720	Non-Poor Leakage	0.649
Dropout Undercoverage	0.138	Dropout Undercoverage	0.243
Non-Dropout Leakage	0.762	Non-Dropout Leakage	0.791
Total Leakage	0.553	Total Leakage	0.508

Source: own calculations using administrative datasets, Chilean ME & MSD

Similar conclusions can be obtained from Table 2.12 as from Table 2.11. In the first instance, there is a trade-off associated with the selection of the mechanisms. To assign a higher fraction of the budget based on the SPF translates into lower undercoverage for the poor and non-poor leakage but a greater lack of coverage for future dropouts and non-dropout leakage. Secondly, when the budget increases, so do all the leakage rates, yet undercoverage for both target groups decreases. Thirdly, the mechanism with the lowest total leakage depends on the budget at disposal.

The last two tables facilitate the comparisons within each targeting approach, but not across them. Table 2.13 summarises the total leakage indicator for the two independent mechanisms and the two combined mechanisms (from Table 2.11 and Table 2.12). Within a fixed budget, Table 2.13 helps the reader to identify the targeting mechanism with the lowest total leakage.

Table 2.13: Total Leakage by Targeting Mechanism and Budget Available

Targeting Mechanisms	The Budget Allows a CCT to Reach x% of Adolescents		
	x=5%	x=20%	x=40%
0% SPF; 100% Model	0.232	0.444	0.563
25% SPF; 75% Model	0.158	0.419	0.553
75% SPF; 25% Model	0.232	0.442	0.508
100% SPF; 0% Model	0.412	0.493	0.544

Source: own calculations using administrative datasets, Chilean ME & MSD

A combined approach is more effective in finding the poor or future dropouts relative to an independent approach. For example, in the context where the budget allows for reaching 5% of the sample, the mechanism that uses 25% of the SPF and 75% of the predictive model provides the lowest level of total leakage. In this example, only 15.8% of students who are assigned the hypothetical CCT are neither poor nor dropouts. In the other two budget scenarios, a mechanism that uses both sources of information also provides the optimal solution. In the second case, each combined mechanism performs better in the simulations than the independent mechanisms. When the budget increases to 40%, the optimal mechanism within the alternatives I analyse is to allocate the first 75% of the resources using the PMT score.

Appendix E discusses the robustness of the results in Table 2.13. I use multiple alternative specifications. First, I change the definitions of the poverty line and income. Second, I modify the measure of school dropout. Third, I use an alternative combined approach. This consists of a single composite score derived from weighting both instruments (the SPF and the predictive model) and assigning the hypothetical CCT using this new index. Finally, I replace the best machine learning algorithm (glmnet) with two others: boosted trees (gbm) and lasso.

Overall, Appendix E shows that my findings are robust to alternative specifications. A targeting mechanism that uses the PMT score in conjunction with the predictive model minimises total leakage (relative to independent mechanisms) in every scenario. This finding does not change depending on the budget, poverty line, income definition, dropout measure or algorithm I select.

In practice, changing the targeting mechanism of a CCT from a PMT to a mechanism that additionally requires using a predictive model of school dropout implies new targeting costs. Appendix F explores in detail whether the conclusions of this part of the assessment hold after adding administrative costs. A targeting approach that relies on both sources of information remains more effective than an independent approach. This holds for all combinations of fixed and variables costs added to targeting mechanisms that incorporate the predictive model.

2.4.3 *Targeting Assessment: Leaked Welfare*

Table 2.14 presents the results of the targeting assessment when the social valuation of the target groups differs. Panel A describes two cases where society places greater value on targeting a CCT at a poor adolescent rather than at a future dropout. Panel B does the opposite, in this case finding a future school dropout is twice or four times more important than finding a poor student.

The measure in Table 2.14 is not comparable to the indicator I provide in Table 2.13. Unlike total leakage, leaked welfare is affected by differences in the social valuation of target groups. For example, where the poor are twice more important than future dropouts, leaked welfare is: i) zero if all hypothetical recipients of a CCT are poor, ii) one if all beneficiaries are non-poor and non-future dropouts, and iii) 0.5 if all potential recipients are future dropouts but non-poor. Where finding future school dropouts is four times more important than finding the poor, leaked welfare is: i) zero if all recipients are future school dropouts, ii) one if they are neither poor nor future dropouts, and iii) 0.75 if they are all poor but not future school dropouts.

Table 2.14: Leaked Welfare by Social Valuation of Target Groups

Panel A: Higher Social Valuation of the Poor Relative to Dropouts

Targeting Mechanisms	The Poor are Twice More Important than Dropouts			The Poor are Four Times More Important than Dropouts		
	The Budget Allows a CCT to Reach x% of Adolescents					
	x=5%	x=20%	x=40%	x=5%	x=20%	x=40%
0% SPF; 100% Model	0.460	0.578	0.649	0.575	0.645	0.692
25% SPF; 75% Model	0.344	0.542	0.637	0.437	0.603	0.678
75% SPF; 25% Model	0.320	0.525	0.578	0.364	0.567	0.614
100% SPF; 0% Model	0.441	0.526	0.581	0.456	0.543	0.600

Panel B: Higher Social Valuation of Dropouts Relative to the Poor

Targeting Mechanisms	Dropouts are Twice More Important than the Poor			Dropouts are Four Times More Important than the Poor		
	The Budget Allows a CCT to Reach x% of Adolescents					
	x=5%	x=20%	x=40%	x=5%	x=20%	x=40%
0% SPF; 100% Model	0.280	0.529	0.659	0.304	0.572	0.707
25% SPF; 75% Model	0.292	0.529	0.657	0.359	0.585	0.709
75% SPF; 25% Model	0.456	0.589	0.649	0.567	0.663	0.720
100% SPF; 0% Model	0.633	0.676	0.704	0.744	0.767	0.784

Source: own calculations using administrative datasets, Chilean ME & MSD

Panel A of Table 2.14 shows that when the poor are valued more highly than future dropouts it is beneficial to make extended use of the Social Protection File to select beneficiaries. The left side of the Panel shows that the combined mechanism that assigns the first 75% of the budget using the SPF provides the lowest leaked welfare. On the right-hand side of Panel A, where the poor are four times more important, the optimal mechanism in two out of three scenarios is to use the PMT score exclusively.

Panel B of Table 2.14 demonstrates that relying exclusively on the predictive model is mostly the optimal mechanism when future dropouts are valued more highly than the poor. On the right-hand side of Panel B, where dropouts are four times more important than the poor, not using the SPF minimises leaked welfare in all three budget scenarios. According to the left side of Panel B, a combined mechanism is only superior in the context of a large budget available for a CCT.

Overall, the leaked welfare measure I provide in this subsection improves our understanding of the targeting performance of different mechanisms. When the social valuation of the target groups differs to a large extent, the preferred mechanism is the one designed to find the target group that is most socially valued. When the welfare weight γ assigned to a future dropout is much higher than that of a poor student, using solely the predictive model is the optimal mechanism to maximise welfare. Conversely, when finding a poor adolescent has a much higher social valuation γ than finding a future dropout, prioritising the PMT mostly provides higher levels of welfare.

2.5 Conclusion

This paper has analysed whether a PMT and alternative targeting mechanisms based on a predictive model of school dropout built with MLA are effective tools to reach the poor and future school dropouts. Its primary motivation has been the improvement of the targeting design and evaluations of CCTs. Overall, the paper provides novel contributions to the social policy targeting field. Its findings are not only relevant for Chile but for all developing countries that have CCTs, wish to develop predictive models of school dropout using administrative records, or wish to strengthen the targeting effectiveness of their social policies.

A first distinctive contribution of my paper is the predictive model of school dropout. The literature is extensive on the topic of determinants but less so on predictions. The core of this research comes from developed countries, especially the United States. My paper is one of the first, along with Adelman et al. (2017), to use large administrative datasets outside a developed nation to study this topic. Furthermore, there are not many applications of MLA for school dropout. The most effective MLA produce results that are in line with the related literature (Adelman et al., 2017; Knowles, 2015; Sorensen, 2018) and that are better than most of the dropout flags analysed in Bowers et al.'s (2013) summary. The best model in predicting school dropout at any point within two years reaches an area under the ROC curve of 0.866 in my test dataset.

These results show that appropriate predictive models of school dropout using administrative datasets are at hand for public officials. Naturally, the selection of variables is restricted by the availability of administrative records, as the models I implement rely solely on information that is currently available from the Chilean government. No variables are provided by costly surveys. This finding has policy implications beyond CCTs, more generally, for every policy that defines students at risk of dropping out of school as their target group. For example, Early Warning Systems can improve their impact by strengthening their ability to find those more likely to drop out of school. In contexts where countries improve their administrative records these lessons deserve attention.

Future research could complement what has been advanced in this paper. For example, longitudinal and multilevel models could be tested for the predictive part. In fact, my approach to find f does not precisely match that of Lamote et al. (2013), which is in the category of longitudinal multilevel modelling. Longitudinal growth models have provided the most accurate predictions on school dropout (Bowers et al., 2013). Additionally, future research projects could consider the case of bringing back to school those who are outside of it when a CCT is implemented.

Another potential direction for further research would be to improve the capabilities of any predictive model by adding new variables. For example, in Chile it is well documented that pregnancy and motherhood are relevant drivers of school dropout (Opazo et al., 2015). Young mothers can be identified through the Civil Register administrative datasets and added to the predictive model. Additionally, the frequency of some predictors I use in my model could be enhanced. For example, the Chilean Ministry of Education has monthly attendance records at the individual level. This information could be useful if attendance levels in the last months of an academic year are a stronger predictor of future dropout than attendance when an academic year starts. The variable absences in the last month is one of the features used by the top-performing algorithm in predicting high school dropout in the literature (Sara et al., 2015).

Another distinctive contribution of my paper is the emphasis on a double vision. Despite having multiple target groups, CCTs have been primarily targeted towards low-income households or individuals. In my targeting assessment, both the poor and future dropouts count. In the paper I offer targeting indicators that combine information on these two key target groups of CCTs. Few papers, beyond Azevedo and Robles (2013), have analysed the quality of CCT targeting by considering more dimensions than just income. My paper and theirs are alike in the sense that both offer a multidimensional targeting approach that fosters the notion that more than one target group and more than one targeting criterion should exist for CCT design and assessment.

The results of the assessment show that a trade-off exists between using the PMT relative to the predictive model. Using the PMT for targeting, instead of the predictive model, is more income progressive, as poor undercoverage and non-poor leakage are reduced. However, future dropout undercoverage and non-dropout leakage increase. This trade-off is explained by the low level of overlap between poverty and future school dropout in Chile. Generally, it is more effective to use these two mechanisms in conjunction rather to use them independently. For different fixed budgets, the proportion receiving the CCT who are neither poor nor a future dropout is minimised with a combined approach. These results hold after considering administrative costs.

These results are partly in line with the findings of Azevedo and Robles (2013). These authors find that their multidimensional targeting approach is better suited to identifying beneficiaries with higher rates of school non-attendance and child labour. This is comparable to my results. Using a combined approach reduces future dropout undercoverage relative to using only a PMT. However, the authors also find that their model identifies the income monetary poor as well as the mechanism used by the CCT. Therefore, in their case no trade-off exists between the two targeting mechanisms assessed. Using multidimensional targeting is always superior.

Another key finding of my paper is that the use of a combined approach is not necessarily more effective when the social valuation of the two target groups differs to a large extent. In the cases where allocating a CCT to a poor adolescent is four times more valuable relative to a future dropout, and vice-versa, it is common to observe that using only one instrument is the optimum.

Regarding policy implications, my paper advances the idea that the targeting of CCTs can be improved when other dimensions beyond income are considered. This finding invites policymakers to broaden the targeting design of CCTs by adding the human capital accumulation dimension. Achieving a better balance among target groups in CCT allocation could also help to enrich and diversify the targeting assessment of these schemes, where a unidimensional outlook has prevailed (Maluccio, 2009; Robles et al., 2015; Skoufias et al., 2001; Stampini & Tornarolli, 2012). An essential and implicit takeaway from the paper is that effective targeting depends on consistency. Targeting design must follow the goals of the policy and its consequential definition of the target groups. If a cash transfer has multiple purposes and target groups, then unidimensional targeting may not be the most effective design for this programme.

The latter conclusion does not necessarily hold if public officials strongly prioritise finding the poor over future dropouts. In this case maintaining the status quo, which is targeting CCTs on the basis of income, is appropriate. Alternatively, policy designers should evaluate the cost-effectiveness of adopting a new targeting mechanism for CCTs. A first step in this sense would be to: i) estimate the costs of developing and implementing a new targeting mechanism, ii) estimate the gains in targeting effectiveness, and iii) compare these with the default scenario.

A potential innovation for CCTs can be to opt for flexible designs, adapting itself to the characteristics of their households and adolescents. After the PMT, or another related targeting mechanism, identifies the poor and after the predictive model classifies every student who is likely to drop out, a tailored CCT design could follow to maximise the likelihood of an impact.

For example, households who are poor and have adolescents that are likely to drop out of school can receive an increased monthly amount for each adolescent who has a high risk of dropping out as informed by the predictive model. This design would be appropriate if economic needs within the household are one of the primary causes explaining the risk of a future dropout. The current design of CCTs can be maintained for poor households without adolescents at risk of dropping out of school. Finally, an alternative CCT design can be implemented for adolescents that are highly likely to drop out of school but are not poor. In this case a direct transfer to these adolescents (not their parents), given in few but large instalments conditional on progressing each grade or graduating (not small monthly payments) might have an increased positive effect.

The results of the paper also contribute to enriching the theoretical literature that seeks to minimise poverty or maximise social welfare (Coady & Skoufias, 2004; De Wachter & Galiani, 2006; Glewwe, 1992; Ravallion & Chao, 1989) when these models are applied to CCTs. As in the case of CCT allocation design and evaluation, moving from considering only one dimension in these theoretical models towards multiple dimensions seems desirable. For example, in welfare maximisation models it might be necessary to consider not only the utility provided by the transfer through the income dimension but also by preventing future dropout. In other cases, it might be necessary to include the elasticity of school dropout to extra income. Finally, regarding poverty minimisation problems it may be useful to incorporate future poverty alleviation explained by increased schooling in addition to current poverty alleviation due to the transfer.

Using the framework of social welfare models can enrich the discussion of what effective targeting is. In theory, CCTs should be prioritised towards the groups that would be most impacted. These are the poorest among the poor and adolescents that would drop out of school but would not because of the CCT. For example, a CCT might have an increased impact for an adolescent whose household needs money relative to a peer with little motivation to continue studying because of low school quality. In practice though, targeting CCTs using these criteria requires not only a flawless measurement of the degrees of poverty but also a perfect understanding of the causes of potential school dropout for each adolescent.

Building upon this paper, future research could strengthen my social welfare analysis. One limitation of my targeting assessment is that I use only undercoverage and leakage rates. For example, I make no distinction between those at the bottom of the distribution and those who are marginally poor. I have assumed that the social valuation of finding any poor is the same. A similar shortcoming exists in the case of dropouts. This analysis could also be enriched if the size of the transfers differs, as higher transfers increase the likelihood of obtaining the desired effects.

Additional angles for future research related to this paper are to: i) include more dimensions than education (such as health: include children who are not attending preventive check-ups as a target group), ii) consider other stages in the educational cycle (such as entry to pre-school), iii) model take-up, not everyone that is eligible for the CCT would end up accessing it, and iv) use new predictive models or means tests instead of the PMT used in this paper.

CCTs continue to be a relevant social policy across the globe. Their goals of poverty alleviation and human capital accumulation remain valid in multiple countries. This paper has intended to collaborate in their improved design and assessment concerning targeting. In Chile, a country where administrative datasets are large and rich, using a PMT in conjunction with a predictive model of school dropout allows for finding more adolescents who are either poor or future school dropouts. Public officials that value these two target groups equally, may find opportunities for increased targeting effectiveness by modifying the allocation rules of CCTs.

Appendix A. Predictors Included in Machine Learning Algorithms

Table 2.15: Predictors of School Dropout Included in Machine Learning Algorithms

Average Grade Year *t*	Urban/Rural
Attendance (%) Year *t*	Same Gender School
Relative Grade Year *t*	Age (Years) at the End of Academic Year *t*
Relative Attendance Year *t*	Male
Promoted Year *t*	Indigenous Background
Mobility Year *t*	Household Number of Rooms
Average Grade Year *t*–1	Head of Household (HH) is Female
Attendance (%) Year *t*–1	Head of Household Lives with a Partner
Relative Grade Year *t*–1	HH Years of Schooling
Relative Attendance Year *t*–1	HH Employed and with Social Security
Promoted Year *t*–1	Relationship with HH
Mobility Year *t*–1	Household Ownership
Average Grade Year *t*–2	Number of Less Than Six Years Olds
Attendance (%) Year *t*–2	Household Income per Capita ($CLP)
Relative Grade Year *t*–2	Average HH Years of Schooling in the Academic Cohort Year *t*
Relative Attendance Year *t*–2	Average HH Years of Schooling in the Academic Cohort Year *t*–1
Promoted Year *t*–2	Average HH Years of Schooling in the Academic Cohort Year *t*–2
Number of Students in the Cohort Year *t*	Index 1 Management of Schools
School Size Year *t*	Index 2 Management of Schools
School Dropout Rate Year *t*	Index 3 Management of Schools
School Dropout Rate Year *t*–1	Index 4 Management of Schools
School Type	Index 5 Management of Schools
Grade	Index 6 Management of Schools
School Administrative Dependency	School SIMCE in Mathematics
School Region	School SIMCE in Language

Appendix B. Indicators Derived from a Classifier of School Dropout

Table 2.16: Examples of Indicators Derived from a Classifier of School Dropout

Formula	Names
FN / D	False Negative Rate, Miss Rate, Type II Error Rate
TP / D	True Positive Rate, Sensitivity, Recall, Power, 1–Type II Error Rate
TN / ND	True Negative Rate, Specificity
FP / ND	False Positive Rate, 1–Specificity, Type I Error Rate, False Alarm Rate
FP / P	False Discovery Rate
TP / P	Positive Predicted Value, Precision, 1–False Discovery Rate
FN / N	False Omission Rate
TN / N	Negative Predicted Value
(TN+TP) / T	Accuracy

Table 2.16 shows examples of indicators for a classifier of school dropout. These indicators, which are derived from the nine shaded cells in Table 2.1, have various names depending on the discipline. Some can be compared to the ones commonly found in the social policy targeting literature. In the context of poor targeting, undercoverage is the proportion of poor households or individuals who do not receive an intervention targeted at the poor (Coady et al., 2004). This indicator is comparable to the false negative rate. In Table 2.16 this rate is equivalent to the proportion of future dropouts who are not predicted as such. In the context of poor targeting, the false negative rate can be calculated using the number of the poor who are incorrectly predicted as non-poor in the numerator and the total number of the poor in the denominator.

Leakage is defined as the proportion of non-poor households or individuals that receive a programme among all its beneficiaries (Coady et al., 2004). This indicator can be compared to the false discovery rate. In Table 2.16 this indicator is equivalent to the proportion of adolescents incorrectly classified as future dropouts among all those predicted as future dropouts. In the context of poor targeting, the false discovery rate would be estimated using the number of non-poor who access the programme (those who are incorrectly predicted as poor) in the numerator and all the recipients (all those predicted as poor) in the denominator.

Undercoverage and leakage are usually referred to in the social policy targeting literature as the exclusion and inclusion errors, respectively. However, these last two terms have also been used to explain conceptually false negatives and false positives. Given the potential misunderstanding that could arise from using such a range of different names, I mostly use the terms undercoverage and leakage in the targeting assessment.

Appendix C. Out of Time Validation for the Predictive Model

This appendix contains the results of the predictive model of school dropout in the second test dataset (year-cohort 2014). I use this dataset to assess the quality of the predictions over time. Table 2.17 shows the AUC for six MLA (similarly than Table 2.8). Table 2.18 presents the true positive rate, false positive rate and accuracy for two different scenarios (as in Table 2.9).

Table 2.17: Area Under the ROC Curve for Models Predicting School Dropout

Machine Learning Algorithms	School Dropout Measures		
	Dropout in Year *t*+1 or *t*+2	Dropout in Year *t*+1	Dropout in Year *t*+2
glmnet	0.878	0.906	0.876
gam	0.875	0.904	0.873
gbm	0.870	0.901	0.870
lasso	0.872	0.899	0.868
svm	0.866	0.862	0.832
rf	0.860	0.886	0.861

Source: own calculations using administrative datasets, Chilean ME & MSD

The elastic net algorithm (glmnet) also prevails in this dataset. The AUC in the second test dataset reaches 0.878 for dropout_t12 and 0.906 for dropout_t1. The generalised additive models (gam) reach the second highest area under the curve in all three measures. The third highest AUC is provided by the boosted trees algorithm (gbm) or lasso. Conversely, the random forest (rf) and support vector machines (svm) algorithms have the worst performances in all measures.

From Table 2.18 we can distinguish a similar pattern to the results in Table 2.17. For example, in the case of dropout_t12, the glmnet algorithm repeats the best performance, finding future dropouts at a rate of 531 out of 1000 in the setting where 10% of adolescents are classified as future school dropouts. In this context, the elastic net algorithm incorrectly classifies non-dropouts at a rate of 56 cases out of 1000 and correctly classifies 90.6% of adolescents. Like in Table 2.17, gam is among the best performers (in many cases the highest), while svm and rf are the worst performing algorithms in respect to the true positive rate and accuracy.

Overall, the results in this appendix show similar tendencies as in the body of the paper. The relative performance of the MLA does not vary between the two test datasets. Hence, the results in the body of the paper are not sensitive to using information only from the most recent cohort.

Table 2.18: True Positive Rate (Sensitivity), False Positive Rate (1–Specificity) and Accuracy for Models Predicting School Dropout

Machine Learning Algorithms	Scenario 1: 10% of Adolescents Classified as Dropouts			Scenario 2: 30% of Adolescents Classified as Dropouts		
	True Positive Rate (Sensitivity)	False Positive Rate (1–Specificity)	Accuracy	True Positive Rate (Sensitivity)	False Positive Rate (1–Specificity)	Accuracy
Panel A: Dropout in Year t+1 or t+2						
glmnet	0.531	0.056	0.906	0.829	0.246	0.761
gam	0.530	0.056	0.906	0.834	0.246	0.762
gbm	0.521	0.057	0.904	0.824	0.247	0.760
lasso	0.512	0.058	0.902	0.832	0.246	0.761
svm	0.512	0.058	0.902	0.818	0.247	0.759
rf	0.506	0.058	0.902	0.806	0.245	0.760
Panel B: Dropout in Year t+1						
glmnet	0.648	0.067	0.917	0.904	0.264	0.745
gam	0.662	0.066	0.918	0.899	0.264	0.745
gbm	0.635	0.068	0.915	0.882	0.265	0.743
lasso	0.635	0.068	0.915	0.887	0.265	0.744
svm	0.568	0.072	0.908	0.808	0.270	0.735
rf	0.622	0.069	0.914	0.871	0.262	0.746
Panel C: Dropout in Year t+2						
glmnet	0.556	0.063	0.908	0.845	0.256	0.751
gam	0.546	0.064	0.907	0.849	0.256	0.752
gbm	0.541	0.065	0.906	0.842	0.256	0.751
lasso	0.532	0.065	0.905	0.837	0.257	0.750
svm	0.510	0.067	0.902	0.757	0.263	0.738
rf	0.531	0.065	0.905	0.811	0.250	0.754

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Appendix D. Summary Statistics for Other Measures of School Dropout

Table 2.19 provides bivariate summary statistics. Its structure is like Table 2.10. The rows contain the different quintile groups of the targeting mechanisms. While Panel A focuses on the quintile groups of household income per capita, Panel B and Panel C concentrate on SPF scores and the predictive model of school dropout, respectively. The columns show the mean value and relative frequency of school dropout for two measures (dropout_t1 and dropout_t2).

Table 2.19: Mean Values and Relative Frequency for School Dropout in Years *t*+1 and *t*+2

Quintile Groups	Dropout *t*+1		Dropout *t*+2	
	Mean	Rel. Freq. (%)	Mean	Rel. Freq. (%)
Panel A: By Quintile Groups of Income per Capita				
1	0.09	27.74	0.13	28.84
2	0.07	22.89	0.10	23.46
3	0.07	19.98	0.09	19.63
4	0.06	17.16	0.07	16.61
5	0.04	12.23	0.05	11.45
Total	0.07	100	0.09	100
Panel B: By Quintile Groups of SPF Scores				
1	0.08	25.39	0.11	25.94
2	0.08	23.30	0.10	23.79
3	0.07	20.91	0.09	20.97
4	0.06	17.57	0.08	17.21
5	0.04	12.83	0.05	12.09
Total	0.07	100	0.09	100
Panel C: By Quintile Groups of the Predictive Model of School Dropout (glmnet)				
1	0.25	77.92	0.31	71.78
2	0.05	14.59	0.08	17.49
3	0.02	4.96	0.03	6.84
4	0.01	1.87	0.01	2.86
5	0.00	0.66	0.00	1.03
Total	0.06	100	0.09	100

Source: own calculations using administrative datasets, Chilean ME & MSD

Similar patterns as in Table 2.10 are observed in Table 2.19. Among them, there is a negative correlation between income per capita or SPF scores and school dropout. Table 2.19 also confirms that the predictive model outperforms the SPF in terms of identifying future school dropouts. There are minor differences in the relative frequency of school dropouts in Panels A and B. However, there are substantive differences in the relative frequency of school dropouts in Panel C. Other new findings emerge from Table 2.19. A higher relative frequency of school dropouts can be found in the first quintile group of the predictive model when the measure of school dropout considered in the analysis is dropout_t1 instead of dropout_t2.

Appendix E. Sensitivity Analysis for Total Leakage Targeting Assessment

I present four distinct types of sensitivity analyses in this appendix. In the first type, I change the methodological approach to measure poverty. On the left side of Table 2.20 I use a higher poverty line. Applying a higher poverty line has the effect of reducing leakage (as more adolescents are classified as poor, it is more likely that poor students will be found among the recipients). In all cases at least one combined mechanism is more effective than the best independent mechanism.

The right-hand part of Table 2.20 presents the results for an alternative definition of income. In this case I replace income per capita with household income over an index of needs. This index considers economies of scale of households with multiple members. The outcomes of the targeting assessment are almost unresponsive to these modifications. I observe minor changes in total leakage for the mechanisms that rely primarily on the Social Protection File.

Table 2.20: Total Leakage: Sensitivity Analysis by Poverty Line and Income Definition

Targeting Mechanisms	Using a Higher Poverty Line			Using a Needs Index		
	The Budget Allows a CCT to Reach x% of Adolescents					
	x=5%	x=20%	x=40%	x=5%	x=20%	x=40%
0% SPF; 100% Model	0.166	0.321	0.418	0.232	0.444	0.563
25% SPF; 75% Model	0.113	0.272	0.406	0.158	0.417	0.552
75% SPF; 25% Model	0.059	0.295	0.338	0.224	0.440	0.507
100% SPF; 0% Model	0.159	0.312	0.347	0.403	0.490	0.543

Source: own calculations using administrative datasets, Chilean ME & MSD

The targeting assessment in the body of the paper considers dropout_t12. The next sensitivity analysis assesses whether the results vary when I use dropout_t1 and dropout_t2 in the prediction of the MLA and as an outcome in the assessment. Table 2.21 presents these results.

Table 2.21: Total Leakage: Sensitivity Analysis by Measure of School Dropout

Targeting Mechanisms	Dropout in Year *t*+1			Dropout in Year *t*+2		
	The Budget Allows a CCT to Reach x% of Adolescents					
	x=5%	x=20%	x=40%	x=5%	x=20%	x=40%
0% SPF; 100% Model	0.344	0.542	0.635	0.295	0.486	0.592
25% SPF; 75% Model	0.240	0.504	0.621	0.209	0.457	0.583
75% SPF; 25% Model	0.272	0.494	0.560	0.257	0.471	0.534
100% SPF; 0% Model	0.439	0.523	0.575	0.427	0.510	0.561

Source: own calculations using administrative datasets, Chilean ME & MSD

Compared to Table 2.13 total leakage increases in every case. This is an expected result because dropout_t12 corresponds to a higher population relative to dropout_t1 and dropout_t2. Accordingly, each targeting mechanism has higher difficulties in finding future dropouts. Despite these changes, a combined approach remains more effective relative to an independent approach. For example, as shown in the first column, assigning 75% of the resources through the predictive model reduces total leakage to 0.240. Conversely, if only the predictive model had been used total leakage would have reached 0.344 while using only the SPF would have made this indicator 0.439.

In the third type of sensitivity analysis I test a new version of the combined approach. Table 2.22 focuses on these modifications. The new mechanism uses a composite score that I create through the combination of the PMT and the predictions of the glmnet algorithm. I assign different weights to each instrument. The left side of the table presents results that are comparable to Table 2.13. For example, the fourth row in Table 2.22 is the same as in Table 2.13. This can be explained by the fact that using a composite score relying 100% upon the SPF is equivalent to distributing the budget using the SPF exclusively. In the second and third rows total leakage increases relative to Table 2.13. Moreover, composite scores are not always more effective than an independent mechanism. In the scenario with the lowest budget using only the predictive model produces a total leakage of 0.232. Total leakage for the two composite indexes I test reaches 0.234 and 0.329, respectively.

The right-hand side of Table 2.22 shows the results with an additional change. In this case, total leakage corresponds to the proportion of students who are not simultaneously poor and future school dropouts. Under this definition the composite scores I present are more effective than the independent approach. This result illustrates the relevance of using a targeting mechanism that is consistent with the definition of the target group(s) of a CCT. When finding the poor or school dropouts matters, using a mechanism that takes the best information available from both sources produces better results than allocation through a composite index. When the target group is adolescents who are both poor and future dropouts, a composite index is better suited.

Table 2.22: Total Leakage: Sensitivity Analysis Using a Composite Score

Targeting Mechanisms	Target: Poor or Dropout			Target: Poor and Dropouts		
(% of Weight in	The Budget Allows a CCT to Reach x% of Adolescents					
Composite Index)	x=5%	x=20%	x=40%	x=5%	x=20%	x=40%
0% SPF; 100% Model	0.232	0.444	0.563	0.786	0.883	0.929
25% SPF; 75% Model	0.234	0.400	0.533	0.733	0.874	0.927
75% SPF; 25% Model	0.329	0.443	0.523	0.825	0.900	0.932
100% SPF; 0% Model	0.412	0.493	0.544	0.914	0.926	0.939

Source: own calculations using administrative datasets, Chilean ME & MSD

In the last type of sensitivity analysis, I exchange glmnet for two other MLA. The left side of Table 2.23 focuses on one of the best performing models: boosted trees (gbm). The right-hand side gives the results for lasso, a model that is never the better or the worst predictor in the previous section. The results from gbm are like the ones in Table 2.13. This is consistent with the equivalent performances of glmnet and gbm as predictors of school dropout. In the case of lasso, total leakage is slightly higher in all contexts relative to glmnet. Despite the latter, the two combined mechanisms are generally more effective than every independent mechanism.

Table 2.23: Total Leakage: Sensitivity Analysis by Machine Learning Algorithm

	Boosted Trees			Lasso		
Targeting Mechanisms	The Budget Allows a CCT to Reach x% of Adolescents					
	x=5%	x=20%	x=40%	x=5%	x=20%	x=40%
0% SPF; 100% Model	0.233	0.450	0.566	0.259	0.467	0.574
25% SPF; 75% Model	0.156	0.424	0.555	0.176	0.437	0.563
75% SPF; 25% Model	0.230	0.443	0.509	0.236	0.449	0.513
100% SPF; 0% Model	0.412	0.493	0.544	0.412	0.493	0.544

Source: own calculations using administrative datasets, Chilean ME & MSD

Appendix F. Targeting Assessment Including Administrative Costs

The paper has relied on an unrealistic assumption, the inexistence of targeting costs. The social policy targeting literature has identified different families of costs for targeted programmes, among them administrative, incentive, private, social and political (Besley & Kanbur, 1990). Accounting for all these costs is beyond the scope of this paper; however I do consider administrative costs. Within the context of implementing targeted transfers, Coady et al. (2004) associate administrative costs with expenses related to collecting information or building a poverty map.

Given three fixed budgets, this appendix assesses what proportion of the budget goes to students that are non-poor and non-dropouts or is spent on administrative costs. I consider two types of administrative costs. These are a fixed cost of using the predictive model and a variable cost per student selected through the predictive model. The logic behind this design is that using the model is associated with fixed costs such as organising the administrative information and running the model (which is independent of the number of students selected with the instrument) and variable costs (such as outreach through channels other than the ones used by the PMT). No costs are associated with using the PMT score. This assumption is justified on the basis that generally these PMTs are country-level instruments that would not see their cost-structure affected when one programme, out of many, changes its targeting design.

Most formally, for each targeting mechanism I estimate the “leaked budget” as follows:

$$Leaked\ Budget = \frac{Budget\ Received\ by\ Non\text{-}Poor\ \&\ Non\text{-}Dropouts + Admin.Costs}{Total\ CCT\ Budget}$$

The leaked budget indicator can range between zero and one. The targeting mechanism with the lowest value is preferred in this exercise. Table 2.24 presents the results for the leaked budget indicator. Panel A and Panel B differ by the fixed cost related to implementing a targeting mechanism that includes the predictive model. The left and right-hand sides of each Panel are differentiated from each other by the variable cost of reaching each student with the information from the predictive model. I analyse all these scenarios for the usual three hypothetical budgets.

Table 2.24: Leaked Budget by Administrative (Fixed and Variable) Costs

Panel A: Fixed Cost of Implementing the Model is 0.5% of the Cost of Universal Coverage						
	Variable Cost is 2%			Variable Cost is 4%		
Targeting Mechanisms	Budget Allows to Reach x% of Adolescents (if no adm. costs)					
	x=5%	x=20%	x=40%	x=5%	x=20%	x=40%
0% SPF; 100% Model	0.310	0.461	0.571	0.322	0.469	0.577
25% SPF; 75% Model	0.237	0.435	0.560	0.246	0.441	0.564
75% SPF; 25% Model	0.319	0.459	0.514	0.321	0.462	0.515
100% SPF; 0% Model	0.412	0.493	0.544	0.412	0.493	0.544
Panel B: Fixed Cost of Implementing the Model is 1.0% of the Cost of Universal Coverage						
	Variable Cost is 2%			Variable Cost is 4%		
Targeting Mechanisms	Budget Allows to Reach x% of Adolescents (if no adm. costs)					
	x=5%	x=20%	x=40%	x=5%	x=20%	x=40%
0% SPF; 100% Model	0.377	0.470	0.575	0.388	0.478	0.580
25% SPF; 75% Model	0.307	0.445	0.564	0.315	0.451	0.568
75% SPF; 25% Model	0.406	0.474	0.518	0.406	0.476	0.519
100% SPF; 0% Model	0.412	0.493	0.544	0.412	0.493	0.544

Source: own calculations using administrative datasets, Chilean ME & MSD

When no administrative costs exist and when the transfers are equal for every recipient the leaked budget is equivalent to the total leakage. In fact, the results for the rows in Table 2.24 where I only use the SPF are the same as in Table 2.13 (because no fixed and variable costs are associated with this option). However, for all the rest of the rows the results are higher relative to Table 2.13. The addition of administrative costs explains this. These costs reduce the amount of resources that can be directed towards the beneficiaries. Logically, the higher these administrative costs are, the higher the proportion of leaked budget is.

Table 2.24 provides one key finding. A combined approach remains predominant relative to an independent approach. This is true despite the addition of administrative costs when using the predictive model. When the budget of the programme allows for reaching 5% or 20% of the sample, assigning 25% of the budget with the SPF and 75% with the predictive model is the optimal mechanism. This holds for all the combinations of fixed and variable costs that I consider. In the case where the budget allows for reaching 40% of students in the sample, selecting 75% of the recipients first with the SPF and the rest with the predictive model minimises the leaked budget.

Chapter 3 Cash for Grades or Money for Nothing? Evidence from Regression Discontinuity Designs

Abstract

This paper estimates the impact of a Chilean cash for grades programme, the *Bono por Logro Escolar* (BLE) in 2013, on future educational outcomes. The cash transfer was targeted using two scores from 2012, an income index and academic performance. I implement a sharp regression discontinuity design along these two running variables. I show that students marginally at each side of the two thresholds used only differed in receiving the BLE in 2013. The main causal estimates for the outcomes are not statistically significantly different from zero.

3.1 Introduction

Should we pay children to learn? Few people are indifferent in regard to this controversial question. Finding a response to this question is particularly relevant for the United States where, by 2008, at least twelve states had adopted schemes where primary and secondary school students were rewarded with money for obtaining good grades or test scores or passing exams (Toppo, 2008).

Over the last eleven years, the public debate around this issue has been noticeable. Several articles have been published in the United States' mainstream newspapers (Calefati, 2008; Guttenplan, 2011; Higgins, 2015; Roberts, Becker, & Ibanga, 2008). The discussion seemed to reach its peak in 2010. During that spring the question: should schools bribe kids? and the corresponding reply "it can work, if it is done right" made it to the cover of Time Magazine (Ripley, 2010), while in the autumn of the same year, the Annual Poll of the Public's Attitudes Toward the Public Schools revealed that 76% of United States' adults opposed the idea of school districts paying small amounts of money to students to read books or get good grades (Bushaw & Lopez, 2010).[23]

[23] Educational philosophers have not been able to agree on this topic either. To illustrate, Sidorkin (2007, 2009) argues that moral reasons exist for paying students. He suggests that, at least for low-income students, most of the economic value of their education is received by the society, not themselves. Accordingly, for these students, schooling does not represent a consumer good but a form of labour that deserves fair compensation. Conversely,

How cash for grades programmes work is also a matter of active deliberation. Kremer et al. (2009), Gneezy et al. (2011) and Fryer (2011) offer syntheses about how these policies operate. On the one hand, incentives provide a price effect that makes specific behaviours more attractive. These economic effects are expected to increase individuals' effort and performance. On the other hand, a psychological effect potentially exists. This could undermine intrinsic motivation, negatively affecting the desired behaviours, especially after the reward is removed.[24]

The evidence regarding the effectiveness of cash for grades schemes is mixed. In Ohio, an effect of 0.15σ is estimated for maths but no impact is observed in three other subjects (Bettinger, 2012). In Kenya, an effect of 0.19σ in test scores and other positive externalities have been found (Kremer et al., 2009). In Israel, the results are positive and statistically significant only for girls (Angrist & Lavy, 2009). In New York, mostly no effects have been observed (Fryer, 2011; Riccio et al., 2013). In Mexico, effects from 0.17σ to 0.30σ in maths test scores have been estimated but these are partly explained by students' cheating (Behrman et al., 2015). In lab experiments, characterised by immediate but lower rewards, some positive effects have been found in Chicago (Levitt, List, Neckermann, & Sadoff, 2012) and India (Hirshleifer, 2017).[25]

My paper adds to this body of knowledge. Specifically, I assess the effect of a cash for grades intervention in Chile on subsequent educational outcomes. The *Bono por Logro Escolar* consists of a one-off cash transfer of up to $50,000 CLP ($100 USD approximately) given to high achieving students from fifth to twelfth grade. The programme intends to provide an incentive or reward for students' effort and overall academic achievement. My assessment is mostly done for the first version of the BLE, which was implemented in the middle of the 2013 academic year (and used information from 2012 to determine eligibility). Using large and rich administrative datasets, the outcomes I analyse are attendance and average grade for the academic years 2013

Warnick (2017) questions cash for grades programmes on ethical grounds. His main argument is that cash incentives are uniquely corruptive of key educational aims and values. Specifically, he claims that these schemes hinder the development of students' self-control, promote advancing private versus public ends and additionally reinforce a perception of students as market actors.

[24] This last argument is quite contested in psychology. After conducting a meta-analysis, a group of authors conclude that, generally, extrinsic rewards are not harmful to motivation (Cameron, 2001; Cameron, Banko, & Pierce, 2001; Cameron & Pierce, 1994). However, Deci, Koestner and Ryan (1999; 2001) and Kohn (1999) have claimed for substantial weakening effects after reassessing evidence on this topic.

[25] I provide more information about the design and results of all these interventions in Appendix G. All these evaluations are randomised control trials. In all cases, randomisation did not occur at the individual level.

and 2014.

I use a sharp regression discontinuity (RD) design to recover the causal estimates. To be eligible for the BLE, students needed to belong to the poorest 30% of the population and be in the top 30% in terms of performance within their cohort. I compare students just below and above each of these two thresholds. Students at one side or the other of each threshold only differ in whether they receive the BLE. Thus, any differences in outcomes can be attributed to the programme. This empirical strategy is suitable as the BLE was implemented retrospectively. Students did not know that information from 2012 would determine whether they would be a recipient in 2013.

Each main causal estimate is not statistically significantly different from zero for both types of outcomes. Additionally, the main causal estimates are centred around zero and their standard errors are small. As a result, each 95% confidence interval contains values of a small magnitude exclusively. To illustrate, the highest upper bound I find for any 95% confidence interval is 0.035 (0.056σ) for average grade and 0.49% (0.052σ) for attendance. In practical terms, these estimates are near a third of the minimum distance between reported grades in the country's educational system and are equivalent to a day of school attendance within an academic year.

If a local average effect of the BLE in 2013 exists this is at best modest in magnitude. Therefore, I am unable to detect it with statistical certainty. Additionally, the analysis by subgroups does not consistently show estimates that are statistically significantly different from zero. These results cannot be generalised for the entire population who received the BLE in 2013 given that the RD estimates are only valid for students near the two thresholds used to target the BLE. Among these students, any potential impact of the BLE in 2013 would have been at least smaller than those found in developing countries, where effects of at least 0.17σ on test scores have been observed.

A possible reason for these results is that the programme was not very salient for the population. If children and adolescents were unaware of the implementation of the BLE then it would not be expected to observe changes in their behaviour. An alternative explanation is that students and their families were aware of the cash transfer but unresponsive to its size. Another potential cause of these results is that the BLE provided two types of effects that cancelled each other out overall.

The rest of the paper is structured as follows. The second section briefly describes the design and implementation of the *Bono por Logro Escolar*. The third section introduces the data sources, provides summary statistics and explains the empirical strategy. The fourth section presents the results of the impact assessment. The last section discusses the results and main implications.

3.2 Programme Description

The design of the *Bono por Logro Escolar* attempted to provide a cash transfer for high achieving students in primary and secondary schools belonging to the poorest segments of the population. Only students enrolled between the fifth and twelfth grades and not older than 24 years old in year *t*–1 were eligible to receive the cash transfer in year *t*. According to the Chilean educational system this meant that, conditional on age, only students in the last four years of primary school (fifth through eighth grade) and all students in secondary education were eligible in principle.[26]

In addition, to be eligible for the programme in year *t* students needed to belong to the poorest 30% of the population. The PFSE index measured this in year *t*–1.[27] In practice, the threshold used in this relative income index was 98 points. Accordingly, to be eligible for the BLE in 2013, students needed to have a PFSE score in the year 2012 equal to or lower than 98 points.

The academic performance requisite needed to receive the BLE in year *t* depended on students' average grade in year *t*–1. Students needed to be in the top 30% of their cohort to be eligible for the BLE.[28] Within each cohort the students were ranked. The number one was assigned to the student with the highest average grade. The average grade in Chile ranges from a minimum of one to a maximum of seven, is generated within each school, and is reported to the central level using only one decimal place. Ties were allowed in the ranking. For example, if two students had the same average grade and this grade was the highest in their cohort both received

[26] The Decree number 24 of the Chilean Ministry of Social Development, published in June 2013, regulates general aspects of the cash transfer (Biblioteca del Congreso Nacional de Chile, 2013).

[27] The PFSE index score was the result of the combination of a proxy means test and a means test. The former variable was the Social Protection File score. This proxy means test score provided the relative position of Chilean households regarding income by needs. Not having a Social Protection File score translated automatically in not being eligible for the BLE. The other variable in the PFSE formula was income per capita. The Ministry of Social Development built this variable from multiple administrative datasets.

[28] Three variables define an academic cohort: i) the school, ii) the type of education provided by that school (for example traditional or adult education, scientific-humanistic or technical-professional), and iii) the grade in which the students were enrolled. Students belonging to the same cohort have these three characteristics in common. Most schools have a specific orientation. However, some schools offer more than one type of education in a given grade (especially in secondary education). Students can also change streams from one academic year to the other.

a value of one in the rankings. Also, in this scenario, the student or students with the second highest average grade received the third position. To be eligible for the BLE in year *t* the proportion between the student ranking and his or her cohort size in year *t*–1, the relative ranking, had to be no higher than 0.3.

In 2013, the BLE provided a lump sum of $50,000 CLP ($100 USD approximately) for students in the top 15% of highest performance and $30,000 CLP (nearly $60 USD) for students within the 15% and 30% range. The size of the cohort needed to be at least seven students for this last rule to hold. If the cohort size was between two and six, only the first-ranked student in the cohort was eligible.

The Ministry of Social Development first delivered the cash transfer in July 2013. Given that the Chilean academic year starts in March and ends in December, the BLE in 2013 was implemented in the middle of the academic year. Thereafter, the BLE has been given once per year with payments occurring between September and November. When it first started in 2013, the programme was implemented retrospectively. The rules regarding eligibility were established after December 2012, which marked the conclusion of the 2012 academic year.

Until 2014, the BLE was paid in cash. Beneficiaries needed to collect their payments in person from local government agencies. From 2015, the BLE progressively adopted bank transfers. The BLE is paid to the student if they are at least 18 years old. Otherwise, a member of the student’s household, most likely the mother, receives the payment. In the Chilean educational system, students are, in theory, expected to graduate from their secondary studies at the age of 18. Hence, the majority of BLE payments are not received by the students but by another household member.

The conception of the BLE is related to the *Seguridades y Oportunidades* law approved in May 2012. Within this large piece of legislation, the *Bono por Esfuerzo* (Cash Payment for Effort) was created. This law establishes that the State can provide conditional cash transfers in diverse areas of social policies to foster social achievements. The *Bono por Logro Escolar* (Cash Payment for Student Achievement) was created in 2013 following this logic. Consequently, the public discourse from politicians and public managers about the goals of the BLE has been that the programme is a tool to appreciate, incentivise or reward students' achievement and effort. This logic has been unaltered despite successive changes of coalitions in government.

I assess whether receiving the BLE in 2013 impacted attendance and academic performance in 2013 and 2014. I analyse the programme in 2013 because I can guarantee that students were not aware that their academic performance in 2012 would have an impact on whether they received the BLE in 2013. This is not necessarily true for the programme in 2014 (which was implemented using information from the academic year 2013) and for its later versions.

The research questions respond to the design, implementation and goals of the BLE but also the literature on cash for grades. Students who received the BLE in 2013 were expected to increase their effort (which, in this paper, is measured through attendance) and future academic performance to receive it again in the future. Additionally, the cash transfers could have had an effect through investments leading to better educational outcomes. For a household with one student and one adult member earning the minimum wage (which reached $210,000 CLP or nearly $410 USD in August 2013), the $50,000 or $30,000 CLP BLE cash transfer could have accounted for 23,8% or 14,3% of the household income of a given month, respectively.

From the psychological standpoint, a common concern related to these types of programmes is that they could have an adverse effect through decreased motivation after the incentive is withdrawn. Accordingly, if BLE in 2013 had been only a one off-payment it would have been improbable to observe positive effects on future attendance and average grade in 2013 or 2014. Any positive impact related to the use of the cash transfer would have been neutralised by weakened student's motivation associated with perceiving BLE as an extrinsic reward. Given that the BLE has remained through the years there is less support for an argument of this kind.

3.3 Data and Methods

This section describes the data and methods. The first subsection introduces the data. The second part of the section explains how I structure the dataset for the analysis. The third subsection provides descriptive statistics. Finally, the fourth part discusses the methodological approach I use to recover the causal estimates of the BLE, the regression discontinuity design.

3.3.1 *Data*

The Ministry of Social Development (MSD) provided most of the datasets I use in this paper.

I combine the datasets using the individual ID number provided by the Chilean State. For privacy purposes the ID numbers were changed by the MSD using an algorithm that is unknown to me. The three primary sources of information for this research are as follows:

Bono por Logro Escolar Dataset

Created since 2013 and replicated annually by the MSD, this dataset contains information for all students between the fifth and twelfth grades in year *t*–1. The dataset excludes students in flexible adult and differential education. Each dataset has approximately 1,900,000 students. I requested two versions of this dataset (the years 2013 and 2014) for this paper. Some variables available in this dataset are: i) school ID, ii) type of school (with categories such as traditional primary education, scientific-humanistic or technical-professional secondary education), iii) grade, iv) average grade, v) attendance, vi) student ID, vii) age, viii) student ranking in the cohort, ix) cohort size, and x) PFSE score. All these variables refer to year *t*–1. Another key variable in this dataset denotes whether the student was a recipient of the BLE in year *t*.

Ministry of Education (ME) Performance Dataset

This dataset is created each year by the Ministry of Education, which later shares it with the MSD. The dataset contains information for all students who finish the academic year from the first through to the twelfth grade (except for flexible adult and differential education). Each version of this dataset has approximately 2,950,000 students. I have at my disposal eight datasets (from 2009 until 2016). The variables available in this dataset are the same as in points i) to vi) of the BLE Dataset.[29] More educational information at the school level is available from public sources. Using the variable school ID, as a key to merge, I obtain the schools': i) administrative dependency (such as public or private subsidised), ii) geographic location, and iii) urban or rural status.

Social Protection File (SPF) Dataset

This dataset contains information for Chilean households and all their members. Each

[29] All these variables are from year *t*. For example, the 2013 ME Performance Dataset provides the average grade for the academic year 2013. The 2013 BLE Dataset provides the average grade for the academic year 2012.

observation represents an individual (adult or child) who lives in a household. Each household has a unique ID number that allows for the identification of all the individuals who belong to it. Households voluntarily requested the SPF at the local government level. The SPF information was essential to be eligible for multiple social policies. From January 2010, the dataset had 10,782,270 individuals (Comité de Expertos Ficha de Protección Social, 2010), approximately 63.5% of Chile's population. I use two versions of this dataset (years 2012 and 2013) in this research. The MSD administers the dataset. Some of its variables are household structure, gender and schooling. With this information I can generate variables such as household size, female head of household and years of schooling of the head of the household.

3.3.2 *Dataset Structure and Sample*

To carry out the assessment, I structure the dataset by cohorts (years *t:* 2013 and 2014). Each year-cohort uses information from years *t*–1, *t* and *t*+1. The BLE Dataset of year *t* provides the programme recipients in year *t* and, among other variables, the academic performance and PFSE scores from year *t*–1. These last two variables are useful to assess eligibility for the cash transfer. I use the information from the SPF of year *t*–1, the same year as the PFSE score, for characterisation. Finally, future average grade and attendance come from the ME Performance Datasets of years *t* and/or *t*+1. This organisation is presented in detail in Table 3.1.

Table 3.1: Dataset Structure by BLE Year-Cohorts

BLE Cohorts	Previous Academic Performance and PFSE Score	SPF Information	BLE Recipient	Future Average Grade and Attendance
Year 2013	2012	2012	2013	2013 & 2014
Year 2014	2013	2013	2014	2015

The body of the paper presents the results for the 2013 cohort. This cohort is more likely to provide valid causal estimates as students were not aware that their academic performance in 2012 would affect whether they received the BLE in 2013. I show the results for the 2014 cohort in an appendix.

The sample excludes students in the eleventh and twelfth grades in year *t*–1. This restriction limits the sample to students that would not (or were unlikely to) graduate from secondary education in years *t*–1 or *t.* Therefore, these students were more likely to have been enrolled in primary or secondary education in years *t* and *t*+1, which reduces sample attrition. Additionally, I exclude from the analysis those students whose cohort size in year *t*–1 was lower than seven.

I also exclude students who were at least 18 years old in the month of year *t* in which the BLE was paid. This action intends to restrict the sample to students that were unable to collect their payments personally. Hence, the estimates are only valid for those students whose cash transfer was collected by a household member. The effect for students aged 18 years or older, who could collect their payments on their own, could not be estimated due to low statistical power.

Another essential characteristic of the sample is that it is comprised exclusively of students that had a PFSE score, and accordingly a valid SPF score. 76.9% of the students in the BLE 2013 Dataset had a valid PFSE score. This is not a representative sample of the population of Chilean students, as households with higher earnings were less likely to request a Social Protection File.

These characteristics of the sample do not favour making inferences about the whole student body. However, this is not problematic if the main findings of this study are linked only to students from the fifth to the tenth grades with a Social Protection File in 2012 who were younger than 18 years old in 2013. Given that this subset of students is more likely to be recipients of cash for grades programmes, the relevance of the findings of this study holds.[30]

3.3.3 *Descriptive Statistics*

Table 3.2 provides descriptive statistics for the BLE in 2013. The first four columns of the table display the mean values for four subsets of students. I obtain these subsets after splitting the sample by PFSE score (below or equal to 98 points and above this threshold) and by relative ranking (equal to or lower than 0.3 and values higher than this threshold). Thus, the first column only contains students who were eligible for the BLE in 2013. The last four columns of the table give the mean, standard deviation, minimum and maximum values for the entire sample.

[30] Students without a PFSE score were less likely to attend a public school and to be enrolled in secondary technical-professional education relative to their peers with a PFSE score. The former group was also more likely to reside in the metropolitan region and urban areas. This information provides evidence that students from higher-income households are not well represented in the sample as in Chile these households are more likely to prefer private schools, scientific-humanistic education, to live in the metropolitan region, and choose schools in urban areas. However, given that the percentage of students attending rural schools and technical-professional education is not negligible is also feasible that some poor households did not have access to a PFSE score.

Panel A shows the mean for the variable BLE recipient in 2013. The mean is one in the first column and zero in the second, third and fourth columns. Every student in the sample who was eligible for the cash transfer was provided with it (if the cash transfer was collected by an adult member of his or her household).[31] Conversely, students who were ineligible for the BLE in 2013 did not have any access to the cash transfer that year. The fifth column of the table shows that approximately 14% of the students in the sample were provided with the BLE in 2013.[32]

Panel B presents descriptive statistics for the variables influencing eligibility for the BLE in 2013. The mean PFSE score in the sample is 113 points with a minimum of 24 and a maximum of 769. Among the BLE recipients, the mean PFSE score is 58.6 points. The mean value for average grade in 2012 is 5.48. For students in the highest 30% of academic performance the mean is higher than 6.00, while for students who are not in this group the mean is around 5.20. The average cohort size is 86.2 students while the average student ranking is 41.1.

Panel C shows that students in the highest 30% in terms of academic performance had a higher percentage of attendance relative to students outside of this group. Regarding the type of school attended, the poorest 30% were more likely to attend public schools and approximately seven out of ten students were enrolled in a traditional primary school. Three out of ten students attended traditional secondary schools. Among this group of students, enrolment in scientific-humanistic (SH) schools was nearly twice that in technical-professional (TP) schools. Students with a PFSE score higher than 98 points were slightly more likely to be enrolled in a secondary SH school relative to their peers with a PFSE score lower than or equal to 98 points.

From Panel D, we see that the sample mean age in 2012 is 12.53 years and that boys were less likely to be part of the highest academically achieving group than girls. Additionally, students in the poorest 30% of the population had a head of household with fewer years of schooling and who was more likely to be female compared to students who were not among the poorest 30%. In relative terms, students with a PFSE score no higher than 98 were also more likely to attend a rural school and to live outside the metropolitan (or capital) region.

[31] The take-up in 2013 did not reach 100% though. By August 2014, nearly 5% of payments were still pending.

[32] The sample contains students that finished the academic year 2012. A tiny fraction did not enrol or enrolled but withdrew in the academic year 2013 (or 2014). For example, only 1.1% of BLE recipients were in this condition in 2013. Given these low levels of school dropout among BLE recipients, I decided against analysing this outcome.

Table 3.2: Descriptive Statistics for the BLE in 2013

	Relative Ranking ≤ 0.3		Relative Ranking > 0.3		Total			
Variables	PFSE ≤ 98	PFSE > 98	PFSE ≤ 98	PFSE > 98				
	Mean	Mean	Mean	Mean	Mean	Std. Dev.	Min.	Max.
Panel A: BLE Recipient in 2013								
BLE Recipient in 2013	1	0	0	0	0.141	0.348	0	1
Panel B: BLE Eligibility Variables in 2012								
PFSE Score	58.6	165.4	57.9	161.9	113.0	67.3	24	769
Average Grade	6.04	6.09	5.18	5.23	5.48	0.55	4.0	7.0
SR: Student Ranking	12.6	13.5	51.0	57.4	41.1	50.1	1	786
CS: Cohort Size	82.9	91.6	79.5	91.2	86.2	79.3	7	786
SR/CS: Relative Ranking	0.156	0.149	0.646	0.630	0.483	0.284	0.001	1
Panel C: Attendance and School Information in 2012								
Attendance (%)	93.9	94.4	91.3	92.2	92.5	7.0	1	100
Public School	0.521	0.425	0.520	0.409	0.465	0.499	0	1
Primary Traditional Education	0.685	0.682	0.717	0.678	0.693	0.461	0	1
Secondary SH Traditional Education	0.197	0.223	0.177	0.227	0.206	0.404	0	1
Secondary TP Traditional Education	0.112	0.091	0.096	0.088	0.094	0.292	0	1
Panel D: Demographic Information								
Age in 2012 (Years)	12.44	12.39	12.58	12.60	12.53	1.82	8	16
Male	0.418	0.438	0.525	0.538	0.499	0.500	0	1
Head of Household Schooling (Years)	9.23	10.77	8.78	10.34	9.74	3.37	0	24
Household Monthly Income ($CLP)	112,418.9	272,970.4	108,020.8	256,265.9	189,150.7	194,888.1	0	15,276,972
Metropolitan Region	0.329	0.365	0.324	0.374	0.349	0.477	0	1
Rural School	0.116	0.070	0.121	0.067	0.093	0.290	0	1
Head of Household is Female	0.582	0.294	0.601	0.331	0.450	0.498	0	1
Head of Household is Employed	0.710	0.818	0.704	0.804	0.760	0.427	0	1
Household Size	4.18	4.19	4.28	4.24	4.24	1.45	1	33
Household Number of Rooms	1.95	2.23	1.93	2.20	2.08	0.93	0	63
Number of Observations	149,834	188,959	357,416	368,451	1,064,660			

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

3.3.4 *Methodological Approach*

Regression Discontinuity Designs: Sharp vs Fuzzy, One vs Multiple Running Variables

A researcher interested in identifying the causal relationship between receiving the BLE and future attendance or academic performance needs a suitable strategy for doing so. For example, a simple comparison of outcomes among those who receive the cash transfer and those who do not is likely to provide biased estimates. Restricting the comparisons to subsets of the population is not likely to work either. Within students in the highest 30% of student achievement, BLE recipients differ relative to non-recipients regarding key observable characteristics such as the type of school they attend or the schooling of their household's heads. An analogous situation happens within the poorest 30%. In this case, BLE recipients and non-recipients are not comparable regarding their previous attendance and gender, among other variables.

An alternative approach is a regression discontinuity design. In all RD designs some exogenous variation in treatment occurs around a threshold of a running variable or rating score. Two types of RD designs exist: sharp and fuzzy. In fuzzy RD designs the running variable is a relevant factor, but not the only one, in explaining treatment status. In sharp RD designs treatment is fully explained by the running variable. In the context of RD designs to evaluate the impact of cash transfers, a sharp RD design is suitable for causal inference only if all units that meet the eligibility criteria (generally those having a score below a threshold) receive the cash transfer and if all units that do not meet the criteria never receive the cash transfer.

A sharp RD design is a suitable approach in this paper. In the simplest sharp RD design only one running variable and threshold exists. Here there are two running variables. Regardless, I can use two different sharp RD designs depending on how I restrict the sample. For example, if I utilise the subset of students in the highest 30% of academic performance, I can implement a sharp RD design using the PFSE scores as a running variable as the treatment changes from zero to one at the 98-point threshold (as shown in the first two columns of Panel A in Table 3.2). Similarly, I can execute another sharp RD design if I restrict the sample to the poorest 30%.

I employ these two alternatives in the paper. Their details are explained later in this subsection.

This type of approach, with more than one running variable and where a restricted sample is used to implement an RD design, has been labelled a frontier RD estimation (Reardon & Robinson, 2012) and provides frontier-specific estimates (Wong, Steiner, & Cook, 2013). Consequently, I report two frontier-specific estimates in this paper. The first type of estimate informs us about the effect of the BLE in 2013 solely for students around the 30% of lowest income. The second type of estimate does so only for students near the 30% of highest achievement.

To obtain causal estimates, RD designs compare outcomes for units just above and below a threshold in a running variable. In a context of two running variables, the alternative to estimating two frontier-specific effects is to estimate a unique frontier-average treatment effect directly. This can be done either by estimating the discontinuity in the outcome along both thresholds simultaneously using bivariate regressions or by collapsing the two running variables into a single one (Wong et al., 2013). Either of these two approaches will imply using more complex methods and assumptions. Given that the frontier-average treatment estimate is the result of a weighted average of frontier-specific estimates (Wong et al., 2013) I opt to provide only the latter type of estimate and use standard (univariate) RD estimation strategies and assumptions.

Local Randomisation vs Continuity-Based Framework in RD Designs

In RD designs, causal inference relies on the assumption that the average outcome for units marginally at one side of the threshold must represent a valid counterfactual for the group just at the other side of the threshold (Hahn, Todd, & Van der Klaauw, 2001; Lee, 2008). Despite this common theoretical understanding, there are disparities in, and a lack of consensus about how RD designs are interpreted and implemented in practice (Cattaneo, Idrobo, & Titiunik, 2018a). Two frameworks exist for RD analysis: local randomisation and continuity-based. These two frameworks rely on different identification assumptions and, consequently, differ in their strategies for estimation and the tests they use to assess the internal validity of their estimates (Cattaneo, Titiunik, & Vazquez-Bare, 2017b; Sekhon & Titiunik, 2017). In practice, the continuity-based framework is most commonly employed (Cattaneo et al., 2018a).

Researchers who adopt the local randomisation framework use the logic of experimental designs to recover causal estimates. They take the simple average of the outcome in a small window on either side of the index' threshold. Hence, the impact estimate is equivalent to the difference in means across the cutoff point. The same approach is used to assess the quality of the randomisation. The underlying assumption of this framework is that the average potential outcomes are uncorrelated with the running variable within a small neighbourhood close to the threshold (Cattaneo, Idrobo, & Titiunik, 2018b).[33] For this assumption to hold, the treatment needs to be at least as good as randomly assigned for units within a distance w from the cutoff C.

Researchers relying on the continuity-based framework use regression at each side of the threshold to predict the limiting value of the outcome precisely at the threshold. In this framework the running variable can be associated with average potential outcomes, but this association is assumed to be smooth at the threshold. Therefore, this continuity assumption allows us to interpret any discontinuity in the conditional expectation of the outcome (as a function of the running variable) at the threshold as causal evidence of the treatment (Imbens & Lemieux, 2008). To assess the internal validity of the causal estimates, it is necessary to check whether the distributions of the pre-treatment variables show discontinuities. Repeated discontinuities in these distributions at the threshold cast doubt on the plausibility of the continuity assumption.

Within the continuity-based framework, when the estimation uses only observations near the threshold the approach is known as local polynomial or non-parametric (or a combination of these two terms). Conversely, a global or flexible parametric model uses all or most of the running variable's support. The former approach has been recommended over the latter in the recent literature (Cattaneo et al., 2018a; Cattaneo et al., 2017b) due to its increased capacity to predict boundary points, the estimate of interest in the continuity-based RD framework.

The local randomisation framework requires a stronger assumption relative to the continuity-based framework (Cattaneo et al., 2018b). For the causal estimates to be unbiased in the local randomisation framework, the average potential outcomes need to be uncorrelated with the running variable along the whole interval [C–w, C+w]. If this assumption holds, then the

[33] Each observation i has two potential outcomes. $Y_i(1)$ is the hypothetical outcome that would be observed if assigned to treatment while $Y_i(0)$ would be observed in case of being assigned to the control group. In a sharp RD context, we can only observe $Y_i(1)$ for units at one side of the threshold (while $Y_i(0)$ remains unobserved for this group) and $Y_i(0)$ for units at the other side of the threshold (with $Y_i(1)$ being unobserved for this group).

continuity assumption on which the other RD framework relies holds. However, continuity does not assure local independence between the average potential outcomes and the running variable.

A priori, I am not able to guarantee that the local randomization RD framework central assumption holds. For example, by design PFSE scores are highly correlated with income per capita, a variable that in turn may be correlated with future academic performance and attendance within the neighbourhood around the 98-point threshold. Similarly, previous academic performance is likely to be a strong predictor of future academic performance within the interval [*C*–*w*, *C*+*w*]. In any of these cases, RD estimates equivalent to differences in means will most likely be biased.

In contrast, the continuity assumption holds more plausibly in the two sharp RD applications I propose. Accordingly, the body of the paper focuses on the continuity-based RD framework. Regardless, I discuss the local randomisation RD framework and its results in an appendix.

RD Application #1: Sharp RD Design Using PFSE Score in 2012 as a Running Variable

In the first RD application, I use the PFSE score as a running variable. I implement a sharp RD design after restricting the sample to students in the highest 30% of academic performance.

Because I use a continuity-based RD framework, first I need to choose the size h of the bandwidth. The size of the bandwidth determines which observations of the income index (around the 98-point PFSE threshold) are used in the local regression. Hence this design relies only upon units within the interval [98–h, 98+h]. I choose h in a data-driven way to avoid selecting it arbitrarily. More precisely, I choose the *h* that minimises the mean squared error of the local polynomial estimator. This is the most popular approach for bandwidth selection in RD designs (Cattaneo et al., 2018a). I obtain the causal estimates from the following regression:

$$Y_i = \alpha_1 + \beta_1 I_{1i} + \theta_1 f(\Delta PFSE_i) + \gamma_1 f(\Delta PFSE_i) I_{1i} + \varepsilon_{1i},$$

where Y_i is average grade or attendance for student *i* in year *t* or *t*+1. $\Delta PFSE_i$ is the 98–PFSE score in *t*–1 for student i (distance to the 98-point PFSE threshold). I_{1i} is a binary variable that takes a value of zero if $\Delta PFSE_i$ is negative. Otherwise, I_{1i} is one. $f(x)$ is a local polynomial of x of order p. ε_{1i} is the error term, the difference between the observed and predicted values.

In this approach β_1 corresponds to the causal effect. Cattaneo et al. (2018a) recommend using a triangular kernel in the regression. Additionally, they recommend that the order p of the local polynomial for use in the estimations should be one or two. For inference, I assume that the observations are clustered by schools. To assess the internal validity, I use the same local regression but I replace Y_i with each variable of X_i', a vector of pre-treatment variables. After running the regression for each pre-treatment variable, I perform a test of joint significance. If the hypothesis of no joint significance is rejected the continuity assumption is unlikely to hold.

RD Application #2: Sharp RD Design Using Average Grade in 2012 as a Running Variable

A second candidate for running variable is the relative ranking in 2012. Although students did not know that their relative academic performance in 2012 was being used for the BLE assignment in 2013, this is not a suitable running variable. The relative ranking is a result of a two-step administrative procedure that transforms the average grade of each student. The first step transforms the average grade into a ranking. The second step transforms the latter value into a relative ranking by dividing the student ranking by the number of students in the cohort. This two-step procedure non-randomly affects the position of students near the 0.3 threshold and causes them not to be comparable around it. Accordingly, there is no valid counterfactual as potential outcomes are likely to differ for units across the 0.3 threshold in the relative ranking.[34]

In my second RD application I use the average grade in 2012 as a running variable. I implement a sharp RD design after restricting the sample to the poorest 30% of the students. The average grade in 2012 is a suitable candidate for a running variable in a sharp RD design. This variable changed the eligibility for the BLE in 2013 deterministically. Among the 30% poorest students, those whose average grade in 2012 was equal to or higher than T received the BLE in 2013, while those whose average grade was lower than T did not receive it. T represents the average

[34] Appendix H provides empirical support for this argument. First, I present differences in means in a small window around the 0.3 threshold in the relative ranking for key pre-treatment variables. I find multiple, large and statistically significant differences. Then, I provide estimates for the continuity-based framework. I find comparable results relative to the local randomisation framework. Finally, the appendix explains how the procedure affects the position of different types of students near the 0.3 relative ranking threshold systematically.

grade that determined which students were in the top 30% in terms of highest achievement in each academic cohort. Thus, various thresholds exist between cohorts. Unlike the relative ranking, the average grade is free from administrative sorting.[35] Students that differ in their average grade by a tiny fraction are not expected to differ in terms of potential outcomes and pre-treatment variables.

As there are different thresholds for multiple subgroups I implement a multi-cutoff RD design. The approach generally consists of normalising the running variable, for example assigning the value of zero to all units with a score of *T*, and then pooling all the observations (Cattaneo, Keele, Titiunik, & Vazquez-Bare, 2016a). This strategy has been used in multiple RD papers in topics such as education, poverty and health (Carneiro et al., forthcoming; De la Mata, 2012; Lindo, Sanders, & Oreopoulos, 2010; Pop-Eleches & Urquiola, 2013). Accordingly, in my second RD application I fit the following local regression (where β_2 is the causal estimate):

$$Y_i = \alpha_2 + \beta_2 I_{2i} + \theta_2 f(\Delta AG_i) + \gamma_2 f(\Delta AG_i) I_{2i} + \varepsilon_{2i},$$

where Y_i is the average grade or attendance for student *i* in year *t* or *t*+1. ΔAG_i is the average grade of student *i* in *t*–1 minus *T* (distance of average grade to the threshold). I_{2i} is a binary variable that takes a value of one if ΔAG_i is non-negative. Otherwise, the variable takes a value of zero. $f(x)$ is a local polynomial of x of order p. Finally, ε_{2i} corresponds to the error term.

I could not select the bandwidth h driven by the data on this occasion. There are not enough unique values in the running variable to implement the algorithm that estimates the optimal bandwidth. This is explained by the average grade in Chile being rounded and reported only with one decimal place (the average grade is mostly the result of a simple average of multiple courses).[36] Instead, I opt for two values of h (*h*=0.2 and *h*=0.3). Given the average grade scale, these are the minimum bandwidths from which I can fit a local regression of order p one and

[35] In the fourth chapter I elaborate in detail the concept of administrative sorting. Administrative sorting relates to procedures, beyond the control and knowledge of individuals, that affect the position of these individuals non-randomly near the threshold. Administrative sorting threatens the continuity assumption on which RD designs rely.

[36] Given that my second running variable is rounded, there is a potential risk of a rounding error in the RD estimations. One way to account for this is to follow Lee and Card (2008) and assume random deviations between the true regression function and the approximating function and estimate confidence intervals based on standard errors that are clustered by the running variable. However, Kolesár and Rothe (2018) recommend against this practice. Another approach is to follow Dong (2015), but this implies modelling the curvature of the outcomes by the running variable within the discrete values used, adding complex and untestable assumptions to the estimates.

two, respectively. The other technical aspects of the estimation (such as the type of kernel, the internal validity tests and clusterisation) are the same as in my first RD application.

RD Graphs and Running Variable Density Test

A key component of RD papers are graphs. I provide multiple types of figures to help the reader to assess the robustness of the key assumption, continuity, on which these designs rely.

The first type shows the relationship between the running variable and the outcomes. In this case a clear discontinuity at the threshold is suggestive of a treatment effect. The second type presents the relationship between the running variable and pre-treatment variables. Repeated discontinuities at the threshold cast doubt on the plausibility of the continuity assumption.

I build these two types of RD graphs following Cattaneo et al. (2018a). The running variable is shown on the horizontal axis while the outcome or pre-treatment variable is represented by the vertical axis. I calculate the average value of the vertical axis variable for a limited number of non-overlapping bins of the running variable. These values are shown by dots. I add a fourth-degree polynomial, fitted in the original data, at each side of the threshold. The polynomial represents the association between the variables in the horizontal and vertical axes. The dashed lines surrounding each polynomial represent the 95% confidence interval of the fitted function.

The third type of RD figures I provide in the paper are histograms of the running variable. A discontinuous density around the threshold usually indicates manipulation or sorting, which makes the continuity assumption on which the RD design mostly relies less likely to hold. I provide these graphs along with the results of a manipulation test for discrete running variables (Frandsen, 2017). I use this method instead of the McCrary test (McCrary, 2008) because the latter tool performs poorly when the running variable is not continuous (Frandsen, 2017).

Frandsen's test is based on smooth approximations of the running variable density close to the threshold of interest. Accordingly, the test detects deviations in the running variable density. Therefore, if the test is rejected this is interpreted as a sign of manipulation or sorting. In this sense this test is like McCrary's but Frandsen's test uses only points immediately adjacent to the threshold. If I assume that the density of the running variable is linear near the threshold (k=0) the test will detect small deviations from linearity at the threshold. This is the most rigorous criterion. If I allow any degree of curvature for the density of the running variable (k>0) the test is less likely to be rejected. I run the test using three values of k (k = 0, 0.1 and 0.2), which allows me to assess the sensitivity of the estimates to the curvature of the running variable.

3.4 Results

This section presents the results of the impact assessment. The first subsection focuses on the use of PFSE scores in 2012 as a running variable. The second part is centred around the use of average grade in 2012 as a running variable. Only continuity-based RD estimates are shown in these subsections. The third part incorporates the findings of the previous two subsections and synthesises the results of the local randomisation RD framework, the impact of the BLE in 2014 and the BLE in 2013 on subgroups (available in Appendixes I, J, and K, respectively).

3.4.1 *PFSE Score in 2012 as a Running Variable*

RD Estimates

Table 3.3 presents the RD estimates (β_1) for each outcome. The first four columns show the results for average grade in 2013 and 2014 while the last four columns do so for attendance. For each outcome, I present two estimates. Within each outcome, the first estimate uses a local quadratic regression (p = 2) while the second estimate uses a local linear regression (p = 1).

The estimates for future average grade are all close to zero and statistically insignificant. The estimates range from –0.010 to –0.003 points. Concerning the standard deviation of average grade, these local regression outputs range from –0.015 to –0.005. The estimates for attendance in 2013 and 2014 are also close to zero and not statistically significant at any level of confidence. The estimates are negative for 2013 and positive for 2014. Overall, the range goes from –0.098% to 0.108% where these results translate approximately into a fifth of a school day. On the scale of the standard deviation of attendance the estimates range from –0.011 to 0.012. Among the estimates, the lowest standard error is 0.013σ while the highest is 0.021σ. Therefore, if any RD estimate would have had an absolute value higher than 0.042σ this would have been statistically different from zero with a 95% level of confidence. Depending on the estimate I analyse, this could have been statistically significant with an absolute value as low as 0.026σ.

RD estimates are more sensitive to bandwidth selection than any other component (Cattaneo et al., 2018a). For this reason, these estimates are expected to be robust to different bandwidth sizes. Table 3.4 and Table 3.5 present again the results of the continuity-based framework for future average grade and attendance. On this occasion, I test two alternative bandwidths for each specification (column) of Table 3.3. I change the size of bandwidth *h* by 1.5 times and by half. Modifying the size of the bandwidth reveals few changes relative to the estimates in Table 3.3. Table 3.4 shows that the estimates for average grade in 2013 and 2014 remain near zero, are all statistically insignificant and mostly negative. Table 3.5 focuses on attendance. The estimates are close to zero, negative in 2013 and positive in 2014, and they all lack statistical significance.

Table 3.6 presents RD estimates for ten pre-treatment variables. None of these variables could have been affected by the BLE in 2013. Consequently, this exercise helps to assess the plausibility of the continuity assumption. I show two estimates per variable. While Panel A uses a local quadratic regression, Panel B uses a local linear regression. Overall, the conditional distributions of these variables do not show discontinuous behaviour at the threshold. I find no statistically significant coefficients for previous academic performance and attendance, age, enrolment in public or rural schools, household income and size, or the schooling and gender of the head of the household. Gender is the only variable for which I obtain a 95% statistically significant estimate. The differences in the estimated proportion of males at the threshold are 0.027 and 0.022 points. For both Panel A and Panel B, the joint tests of statistical significance of these ten variables are not rejected.

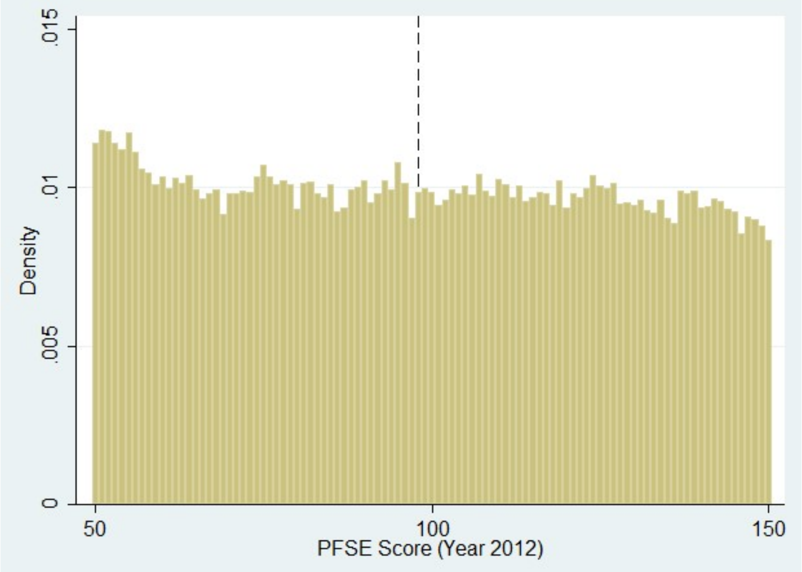
To sum up, the main causal estimates are not statistically significantly different from zero. If an effect of the BLE in 2013 exists (for students near the 30% relative income threshold), its size is likely to be no larger than a small fraction of a standard deviation and cannot be statistically detected. There is a small discontinuity in the distribution of gender, but no other pre-treatment variable shows this behaviour. In my sample, being male is not correlated with attendance; thus it is improbable that this outcome is affected. Being male is weakly negatively correlated with future average grade. This correlation may explain the negative coefficients for average grade. If this is the case the causal estimates could be slightly downwardly biased.

However, the effects are most likely small in magnitude and not statistically significant.

RD Graphs and Running Variable Density Test

Figure 3.1 presents the density of the PFSE scores among the subgroup of students in the highest 30% of academic achievement and the results of the manipulation test proposed by Frandsen (2017). No abrupt changes in the density are observed along this figure or close to the BLE threshold (vertical line). The test fails to reject the hypothesis of no difference in the expected density at each side of the threshold. The most rigorous application, with no degree of curvature allowed, has a *p*-value of 0.187. The *p*-value increases when I partly relax this restriction.

Figure 3.1: PFSE Score Density and Frandsen Manipulation Test
(Students in the Highest 30% of Academic Performance in 2012)



Manipulation test (*k*=0) *p*-value=0.187
Manipulation test (*k*=0.1) *p*-value=0.221
Manipulation test (*k*=0.2) *p*-value=0.293

Source: own calculations using administrative datasets, Chilean ME and MSD

Figure 3.2 presents a series of graphs that depict the relationship between the running variable and the outcomes of the assessment. Future average grade graphs can be found in the upper panel of the figure, while future attendance graphs are in the lower panel. In general terms, the figure shows a positive but weak association between the PFSE scores and both types of outcomes. More importantly, no graph shows a relevant discontinuity between the polynomial fitted functions at each side of the vertical line. Any detected discontinuity is small and not statistically significant as the confidence intervals of the fitted functions noticeably overlap at the threshold. This evidence is consistent with the RD estimates of Tables 3.3, 3.4 and 3.5.

Figure 3.3 shows eight graphs that illustrate the relationship between the PFSE scores in 2012 and the variables that could not have been affected by the BLE in 2013. Some variables, such as income and head of household's schooling, have an evident degree of association with the PFSE scores. Conversely, other variables have a weaker correlation with the PFSE index. From any of the eight graphs I present in Figure 3.3 it is possible to claim that discontinuous behaviour exists for a pre-treatment variable at the threshold. In all cases where a difference in the polynomial fit is observed, this is small in magnitude. Additionally, the 95% confidence intervals of the polynomial fits mostly overlap in each of these eight graphs. Although not strictly comparable with the estimates in Table 3.6, Figure 3.3 provides similar insights to this source.

Table 3.3: RD Estimates for Outcomes (Using Mean Squared Error Optimal Bandwidth)

Outcomes	average_grade2013		average_grade2014		attendance2013		attendance2014	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
RD Estimate:	-0.010	-0.008	-0.008	-0.003	-0.025	-0.098	0.108	0.106
Original Outcome	(0.012)	(0.009)	(0.013)	(0.010)	(0.146)	(0.107)	(0.159)	(0.143)
RD Estimate:	-0.015	-0.012	-0.012	-0.005	-0.003	-0.011	0.012	0.011
In Standard Deviations	(0.019)	(0.014)	(0.021)	(0.016)	(0.017)	(0.013)	(0.017)	(0.015)
Number of Observations	91,832	95,380	87,263	87,263	109,231	105,768	125,075	73,092
Bandwidth Size (*h*)	25.73	27.50	24.96	24.87	31.14	30.45	35.58	21.12
Order *p* of Local Polynomial	2	1	2	1	2	1	2	1

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Table 3.4: RD Estimates for Future Average Grade (Sensitivity Analysis to Bandwidth)

Outcomes	average_grade2013				average_grade2014			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
RD Estimate:	-0.010	-0.016	-0.005	-0.010	-0.005	-0.011	0.001	-0.011
Original Outcome	(0.010)	(0.017)	(0.008)	(0.011)	(0.011)	(0.018)	(0.009)	(0.013)
Number of Observations	136,909	45,760	144,061	49,285	128,471	41,880	128,471	41,880
Bandwidth Size (*h*)	38.59	12.86	41.25	13.75	37.43	12.48	37.31	12.44
Order *p* of Local Polynomial	2	2	1	1	2	2	1	1

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Table 3.5: RD Estimates for Future Attendance (Sensitivity Analysis to Bandwidth)

Outcomes	attendance2013				attendance2014			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
RD Estimate:	-0.090	-0.166	-0.090	-0.028	0.122	0.034	0.080	0.005
Original Outcome	(0.122)	(0.200)	(0.093)	(0.143)	(0.134)	(0.219)	(0.121)	(0.196)
Number of Observations	166,185	56,237	162,448	52,798	185,597	62,667	111,405	38,417
Bandwidth Size (*h*)	46.71	15.57	45.68	15.23	53.37	17.79	31.69	10.56
Order *p* of Local Polynomial	2	2	1	1	2	2	1	1

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

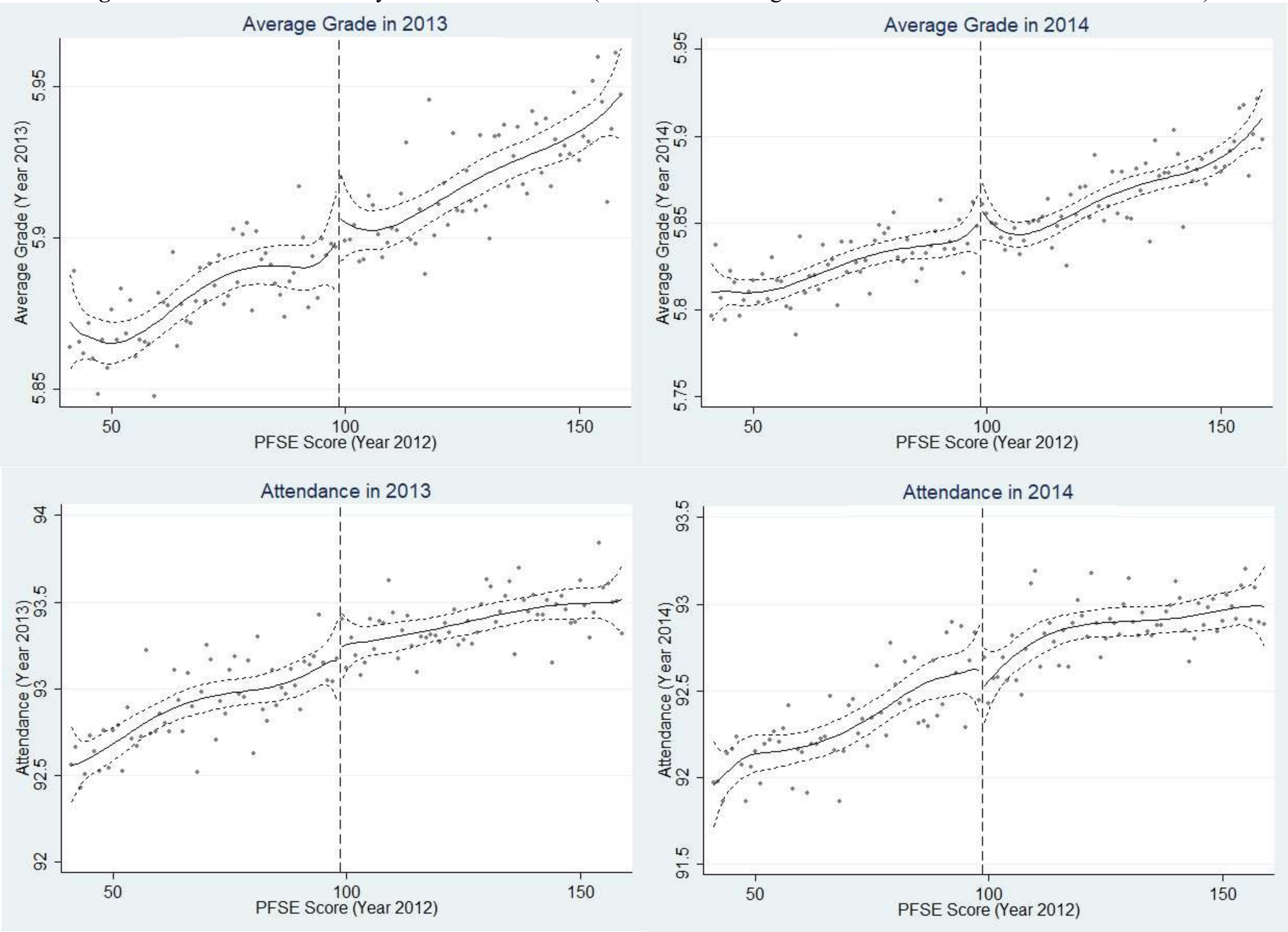
Table 3.6: RD Estimates for Pre-Treatment Variables (Using Mean Squared Error Optimal Bandwidth)

Pre-Treatment Variables (Var.)	(1) avg_grade2012	(2) attendance2012	(3) age	(4) male	(5) schoolpub	(6) schoolrural	(7) hmonthincome	(8) hsize	(9) hhschooling	(10) hhfemale
Panel A: Order p of Local Polynomial =2 (Local Quadratic Regression)										
RD Estimate:	-0.007	-0.086	0.014	0.027**	-0.007	-0.004	-1,024.1	0.014	-0.102	0.001
Original Var.	(0.008)	(0.130)	(0.043)	(0.011)	(0.015)	(0.008)	(2,743.5)	(0.029)	(0.080)	(0.010)
Number of Observations	120,566	96,218	117,115	100,039	106,700	103,351	100,039	106,652	79,028	110,005
Bandwidth Size (*h*)	34.09	26.94	32.68	28.77	29.84	28.66	28.50	31.37	22.87	32.40
Panel B: Order p of Local Polynomial =1 (Local Linear Regression)										
RD Estimate:	-0.006	-0.135	0.012	0.022**	-0.004	-0.002	-1,070.2	0.011	-0.085	-0.000
Original Var.	(0.007)	(0.099)	(0.038)	(0.010)	(0.014)	(0.007)	(2,364.1)	(0.025)	(0.064)	(0.008)
Number of Observations	85,347	92,635	85,347	65,158	74,553	81,678	61,849	72,109	61,849	79,028
Bandwidth Size (*h*)	24.20	25.86	23.84	18.58	20.98	22.84	17.81	20.55	17.95	22.77

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Figure 3.2: Future Outcomes by PFSE Score in 2012 (Students in the Highest 30% of Academic Performance in 2012)



Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Figure 3.3: Pre-Treatment Variables by PFSE Score in 2012 (Students in the Highest 30% of Academic Performance in 2012)

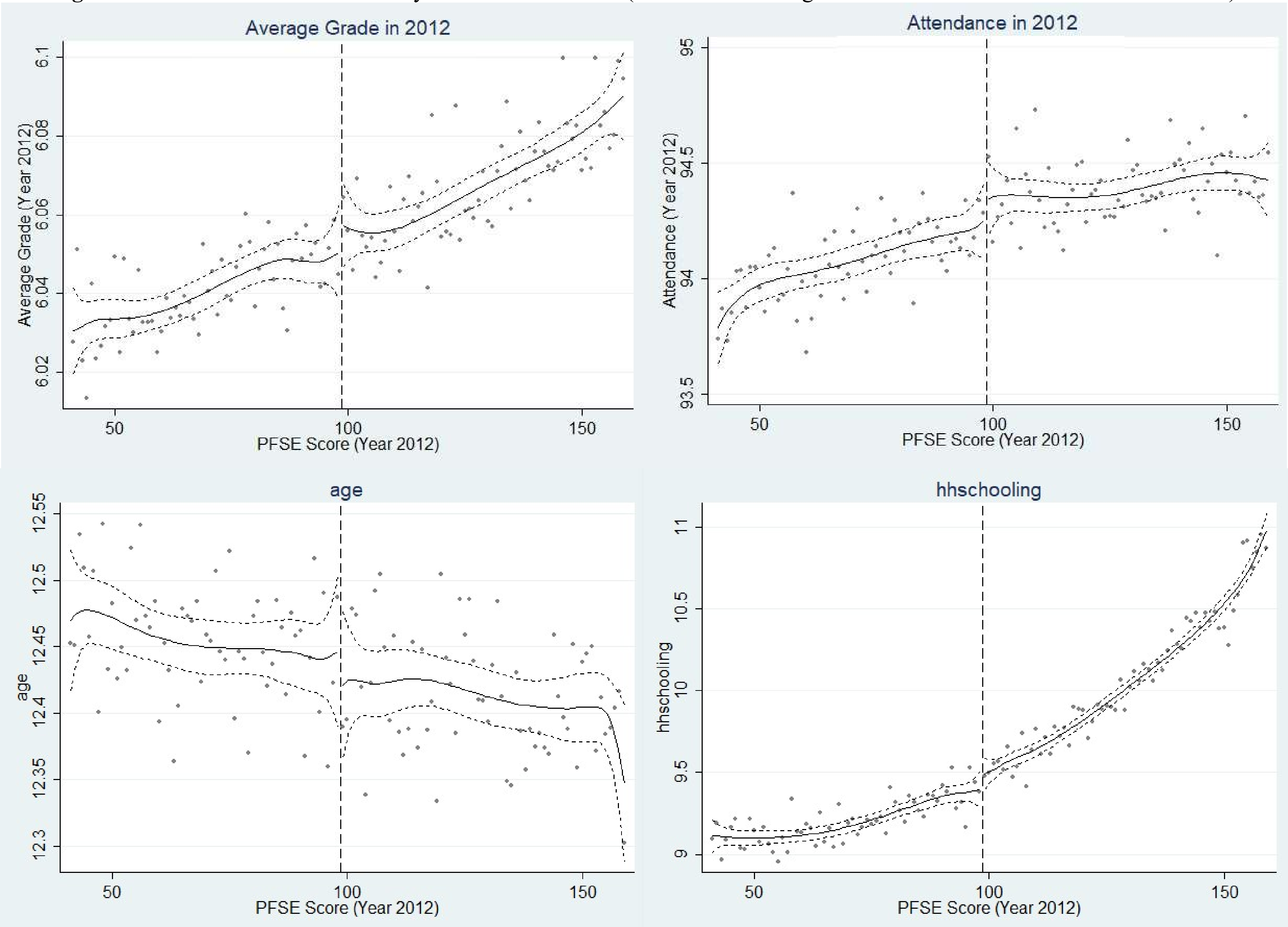
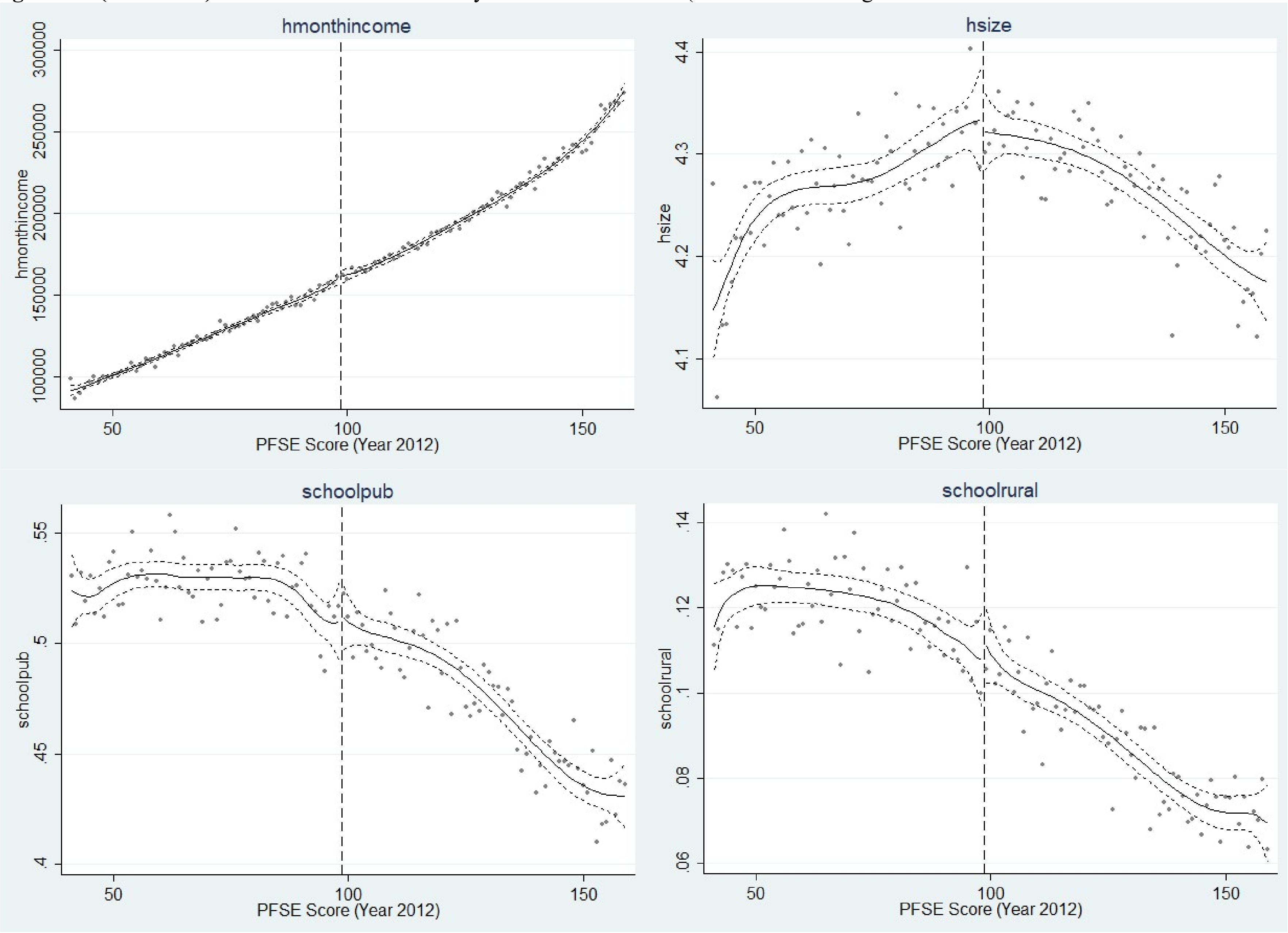


Figure 3.3 (continued): Pre-Treatment Variables by PFSE Score in 2012 (Students in the Highest 30% of Academic Performance in 2012)



Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

3.4.2 *Average Grade in 2012 as a Running Variable*

RD Estimates

Table 3.7 presents the results for future average grade and attendance. The first four columns focus on the former type of outcome, while the last four focus on the latter. The first type of RD estimate per outcome (β_2) uses a bandwidth h of 0.3 and a local quadratic regression. The second type of estimate comes from a local linear regression that uses a bandwidth h of 0.2.

All the main estimates for future average grade are small and not statistically different from zero at a 95% level of confidence. The estimates fluctuate from a minimum of –0.003 points (–0.004σ) to a maximum of 0.017 points (0.027σ). This last value is close to achieving a 95% level of statistical significance given that its standard error is 0.015σ. The range for the estimates of attendance goes from –0.043% to 0.147% (–0.005 to 0.016 of a standard deviation). Given that the highest standard error is 0.020σ, the approach would have detected statistical significance for any estimate higher than 0.040σ. Overall, these results are similar to the estimates in Table 3.3.

Table 3.8 and Table 3.9 summarise the sensitivity analysis by bandwidth choice. Overall, between the outputs of each of these tables and the estimates from Table 3.7, the differences in magnitude are positive but small. The estimates for future average grade in Table 3.8 are all close to zero (ranging from 0.002 to 0.023 points). Some estimates are statistically significant at a 95% level of confidence but overall these are not robust to alternative specifications. The estimates for attendance in 2013 and 2014 are small and lack any statistical significance. The estimates for 2013 range from 0.012% to 0.053%, and in 2014 from 0.138% to 0.189%.

Table 3.10 presents the estimates for ten pre-treatment variables. The specifications I use per variable follow a similar pattern to Table 3.7. The Panel A estimates are the result of a local quadratic regression run in a bandwidth of size 0.3. Panel B provides the results for a local linear regression run in a bandwidth of size 0.2. This approach helps to assess the likelihood of the continuity assumption to hold, as the BLE could have impacted none of these variables.

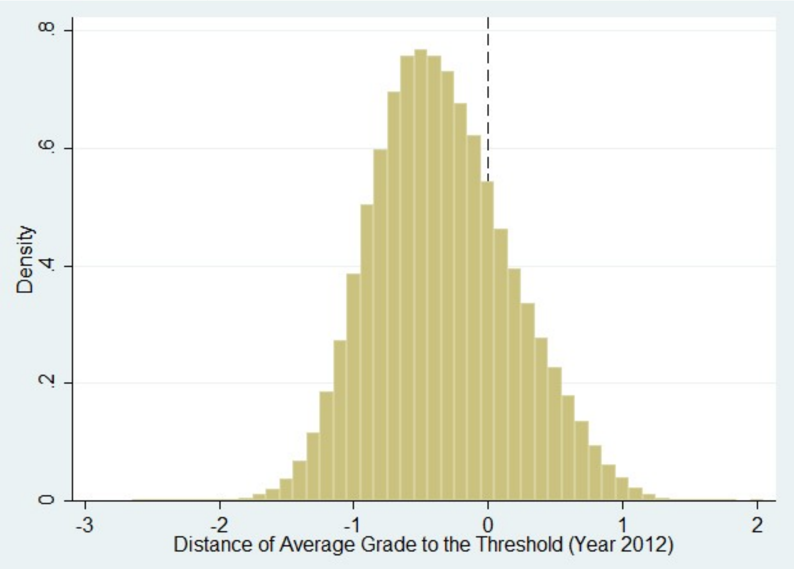
Overall, no clear discontinuities emerge in the distribution of any of these variables at the threshold. I find no significant differences from zero at a 95% level of confidence. The only estimate at a 90% level of confidence is average grade in 2012 for a local linear regression (p=1). Given that this coefficient is small and positive, if the average grade linear causal estimates in Tables 3.7 and 3.8 are not free of bias, these estimates would be slightly inflated. Accordingly, the potential unbiased estimates would be lower and less likely to be statistically significant.

The main causal estimates for both outcomes are not statistically significantly different from zero at a 95% level of confidence. I observe statistically significant coefficients for some alternative specifications of average grade, but these are not robust. The main estimates for both outcomes are near zero and have standard errors no larger than 0.020 of a standard deviation. Hence, for students around the 30% of highest achievement, the BLE in 2013 is unlikely to have caused a substantial effect.

RD Graphs and Running Variable Density Test

Figure 3.4 presents the density of the normalised average grade variable among the poorest 30% of students. Overall, the distribution of the distance of average grade to the cohort threshold T is smooth. The density decreases when approaching the zero threshold (the vertical line) from the left and increases when reaching this cutoff point from the right. The Frandsen test is not able to detect significant deviations in the expected density of the running variable. The minimum p-value of the test, for the most rigorous version of it, is 0.678.

Figure 3.4: Distance of Average Grade to the Threshold Density and Frandsen Manipulation Test (Students With PFSE ≤ 98 in 2012)



Manipulation test (*k*=0) *p*-value=0.678
Manipulation test (*k*=0.1) *p*-value=0.862
Manipulation test (*k*=0.2) *p*-value=0.989

Source: own calculations using administrative datasets, Chilean ME and MSD

Figure 3.5 illustrates the relationship between the running variable and the outcomes. The upper panel shows the graphs for average grade while the lower panel focuses on attendance. The figure displays a positive association between the running variable and the outcomes. This association is strong as the confidence intervals of the fits are narrow. If the polynomial fits at each side of the vertical line were to be extended to the threshold it would not be possible to distinguish a clear discontinuity. In other words, any estimated discontinuity in the extrapolation would most likely be small. However, it does not seem possible to discard statistically significant differences only by looking at this figure. Although it is not strictly comparable, this graphic evidence is concordant with the results in Tables 3.7, 3.8 and 3.9.

Figure 3.6 shows eight graphs that illustrate the relationship between the running variable and other variables that could not have been affected by the BLE in 2013. The relationship between the running variable and each of these eight pre-treatment variables differs notoriously. The graphs help to determine the plausibility of the continuity assumption to hold. No discontinuous relationships at the threshold are noticeable from the figure. No graph suggests an abrupt change in the projected distributions of the pre-treatment variables at the threshold. The results from Table 3.10 are consistent with Figure 3.6. Both findings provide support for the suitability of this RD approach to identify the causal effects of the *Bono por Logro Escolar* in 2013.

Table 3.7: RD Estimates for Outcomes

Outcomes	average_grade2013		average_grade2014		attendance2013		attendance2014	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
RD Estimate:	-0.003	0.006	0.010	0.017*	-0.043	0.009	0.119	0.147
Original Outcome	(0.011)	(0.009)	(0.012)	(0.009)	(0.169)	(0.126)	(0.188)	(0.141)
RD Estimate:	-0.004	0.010	0.016	0.027*	-0.005	0.001	0.013	0.016
In Standard Deviations	(0.017)	(0.014)	(0.019)	(0.015)	(0.020)	(0.015)	(0.020)	(0.015)
Number of Observations	166,638	112,023	163,219	109,783	166,638	112,023	163,219	109,783
Bandwidth Size (*h)*	0.300	0.200	0.300	0.200	0.300	0.200	0.300	0.200
Order *p* of Local Polynomial	2	1	2	1	2	1	2	1

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Table 3.8: RD Estimates for Future Average Grade (Sensitivity Analysis to Bandwidth)

Outcomes	average_grade2013				average_grade2014			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
RD Estimate:	0.004	0.002	0.016**	0.011	0.016*	0.014	0.023***	0.021**
Original Outcome	(0.009)	(0.009)	(0.007)	(0.008)	(0.009)	(0.010)	(0.008)	(0.008)
Number of Observations	269,585	219,355	219,355	166,638	263,547	214,693	214,693	163,219
Bandwidth Size (*h)*	0.500	0.400	0.400	0.300	0.500	0.400	0.400	0.300
Order *p* of Local Polynomial	2	2	1	1	2	2	1	1

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Table 3.9: RD Estimates for Future Attendance (Sensitivity Analysis to Bandwidth)

Outcomes	attendance2013				attendance2014			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
RD Estimate:	0.020	0.012	0.053	0.043	0.189	0.183	0.138	0.160
Original Outcome	(0.124)	(0.138)	(0.101)	(0.110)	(0.139)	(0.155)	(0.113)	(0.123)
Number of Observations	269,585	219,355	219,355	166,638	263,547	214,693	214,693	163,219
Bandwidth Size (*h)*	0.500	0.400	0.400	0.300	0.500	0.400	0.400	0.300
Order *p* of Local Polynomial	2	2	1	1	2	2	1	1

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

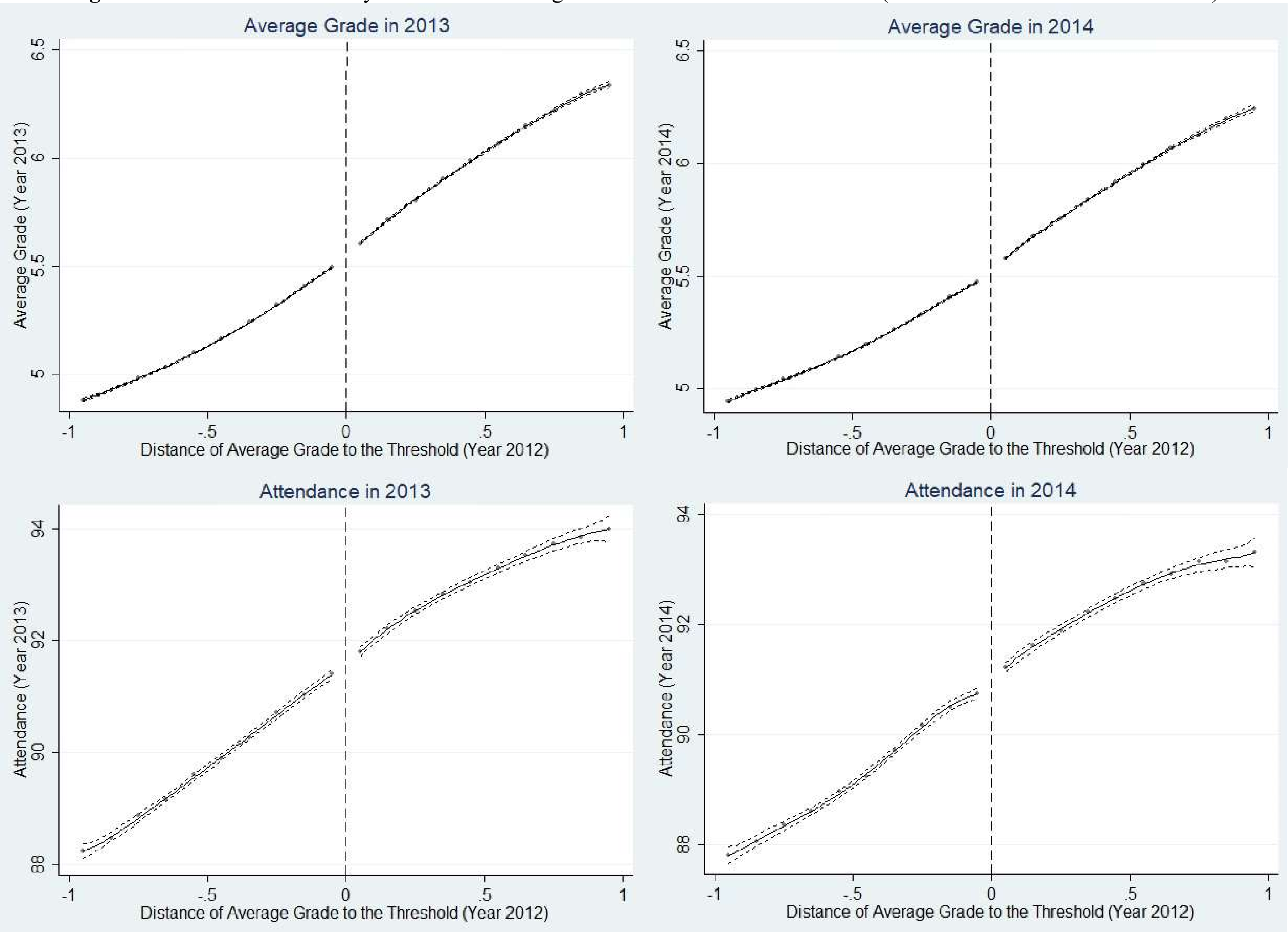
Table 3.10: RD Estimates for Pre-Treatment Variables

Pre-Treatment Variables (Var.)	(1) avg_grade2012	(2) attendance2012	(3) age	(4) male	(5) schoolpub	(6) schoolrural	(7) hmonthincome	(8) hsize	(9) hhschooling	(10) hhfemale
Panel A: Bandwidth h=0.3 & Local Quadratic Regression (p=2)										
RD Estimate:	0.002	-0.152	0.070	0.017	0.012	0.004	827.5	0.022	-0.047	-0.007
Original Var.	(0.008)	(0.146)	(0.045)	(0.011)	(0.015)	(0.009)	(2,175.9)	(0.031)	(0.070)	(0.010)
Number of Observations	169,944	169,944	169,943	164,310	169,944	169,944	164,310	164,310	164,310	164,310
Panel B: Bandwidth h=0.2 & Local Linear Regression (p=1)										
RD Estimate:	0.012*	-0.064	0.044	0.013	0.008	0.002	1,028.4	0.010	-0.000	-0.005
Original Var.	(0.007)	(0.113)	(0.037)	(0.008)	(0.014)	(0.007)	(1,525.1)	(0.022)	(0.052)	(0.008)
Number of Observations	114,153	114,153	114,152	110,369	114,153	114,153	110,369	110,369	110,369	110,369

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Figure 3.5: Future Outcomes by Distance of Average Grade to the Threshold in 2012 (Students With PFSE ≤ 98 in 2012)



Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Figure 3.6: Pre-Treatment Variables by Distance of Average Grade to the Threshold in 2012 (Students With PFSE ≤ 98 in 2012)

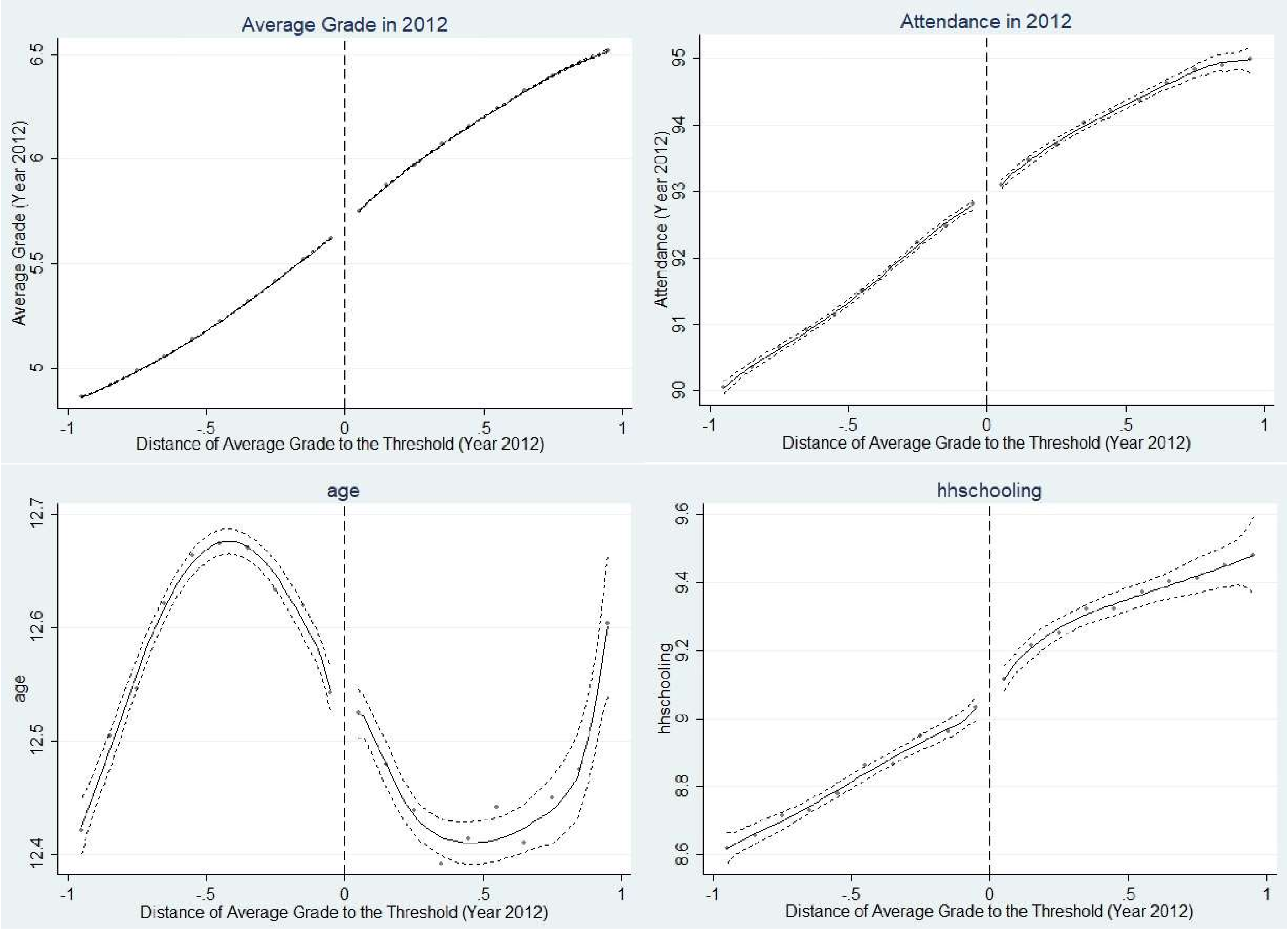
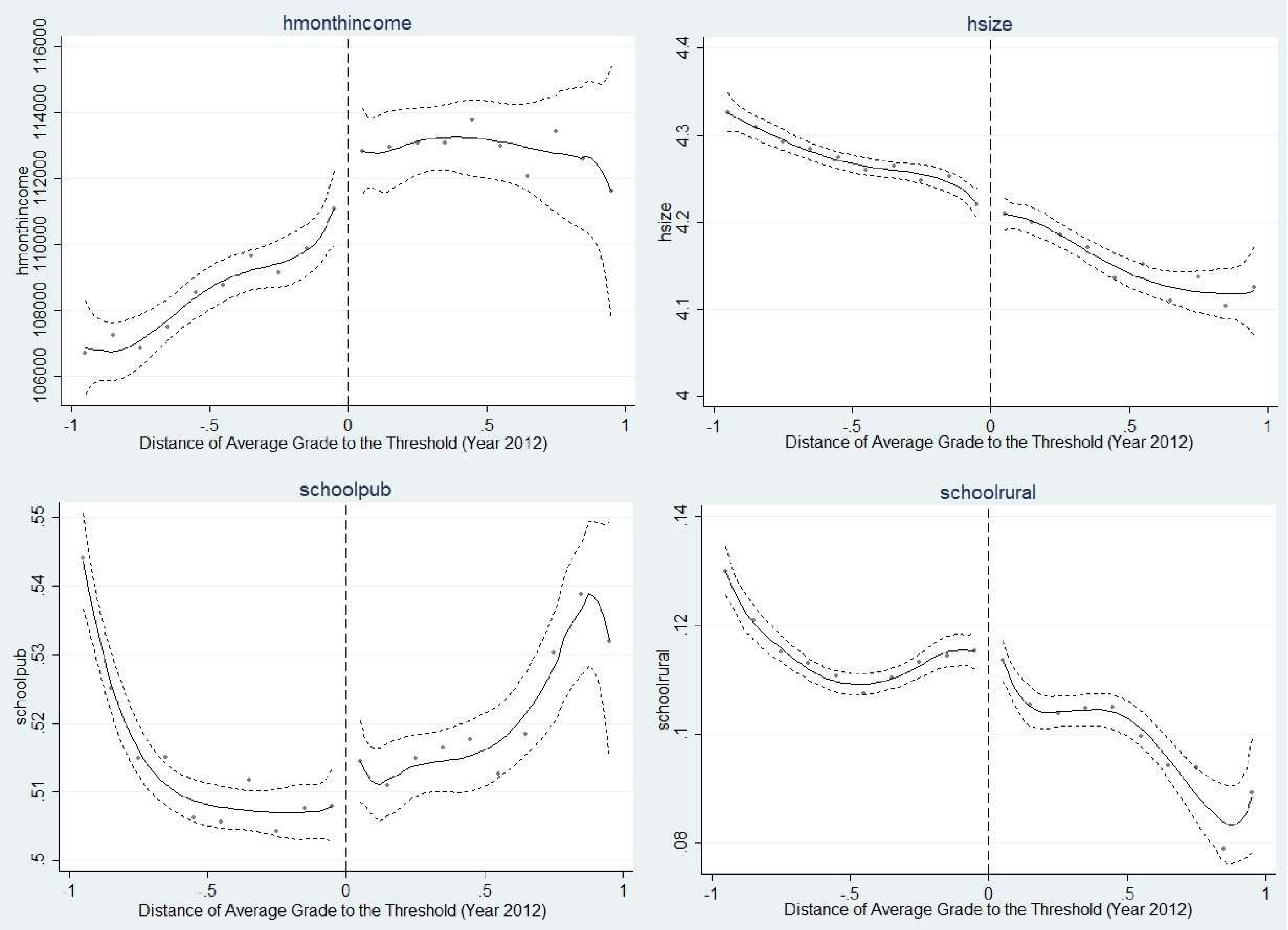


Figure 3.6 (continued): Pre-Treatment Variables by Distance of Average Grade to the Threshold in 2012 (Students with PFSE ≤ 98 in 2012)



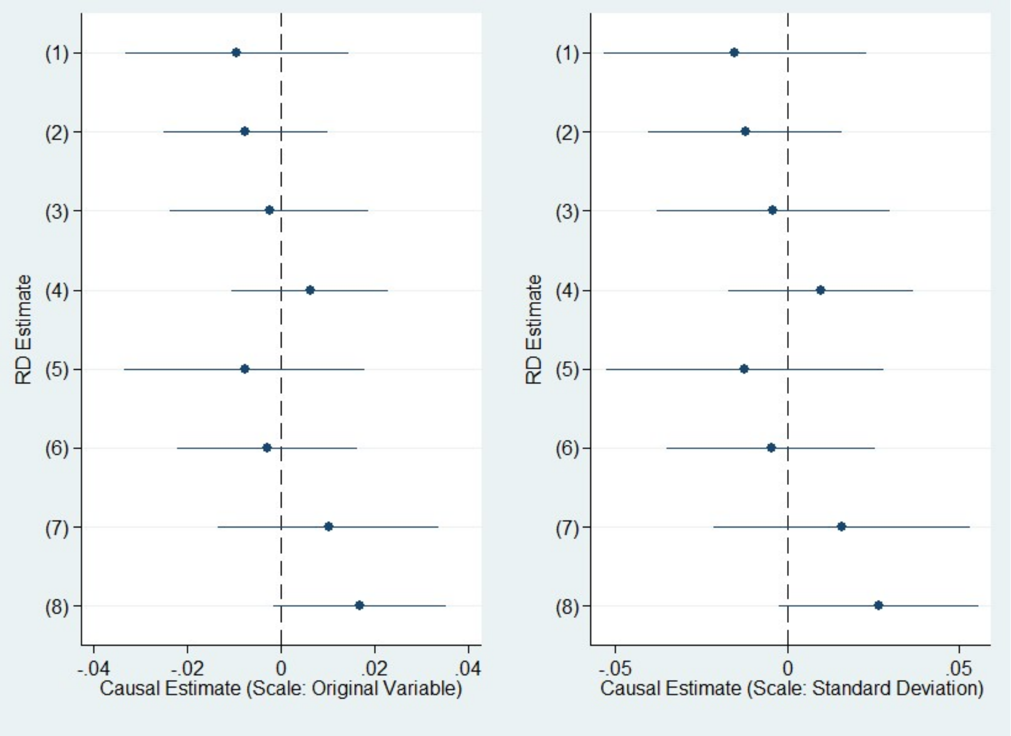
Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

3.4.3 *Summary*

I analyse different effects of receiving the BLE in 2013 over two groups of students, those around the 30% threshold of lower income and around the 30% threshold of highest academic achievement. Despite using different approaches, the value of zero is part of the 95% confidence interval in every main causal estimate. Additionally, the standard errors range from 0.013σ to 0.021σ.

Figure 3.7 summarises the main estimates for average grade. The left panel of the figure uses the outcome original scale. The right-hand panel shows the estimates in standard deviations. The main estimates vary by the year of the outcome (2013 or 2014), the running variable used (PFSE score or distance of average grade), and the order *p* of the local polynomial (one or two).

Figure 3.7: Summary of RD Point Estimates and Confidence Intervals for Average Grade



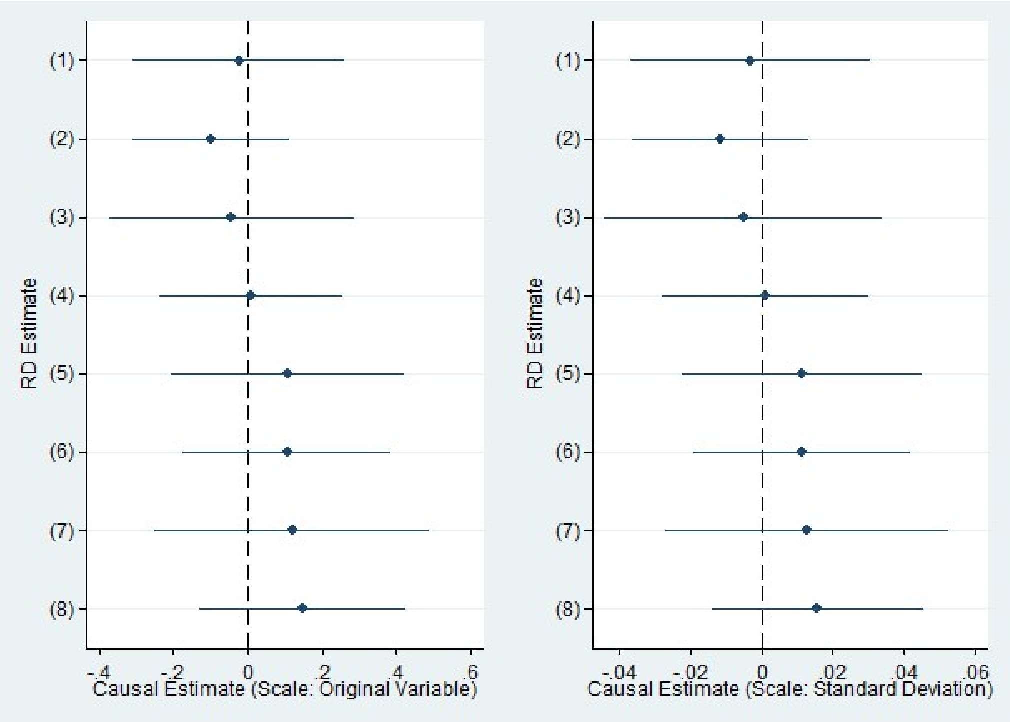
(1): Year of outcome is 2013, running variable is PFSE score and *p* is two
(2): Year of outcome is 2013, running variable is PFSE score and *p* is one
(3): Year of outcome is 2013, running variable is distance of average grade and *p* is two
(4): Year of outcome is 2013, running variable is distance of average grade and *p* is one
(5): Year of outcome is 2014, running variable is PFSE score and *p* is two
(6): Year of outcome is 2014, running variable is PFSE score and *p* is one
(7): Year of outcome is 2014, running variable is distance of average grade and *p* is two
(8): Year of outcome is 2014, running variable is distance of average grade and *p* is one

Source: own calculations using administrative datasets, Chilean ME & MSD

The most negative estimate of the 95% confidence interval lower bound is –0.034 (–0.053σ) while the highest estimate of the upper bound is 0.035 (0.056σ). These values represent nearly one third of the shortest distance between two grades in the country's educational system.

Figure 3.8 summarises the main estimates for attendance. The left and right-hand panels of the figure present the estimates using the outcome original scale and in standard deviations, respectively. The smallest estimate of the lower bound of the 95% confidence interval is – 0.37% (–0.044σ), while the largest estimate of the upper bound is 0.49% (0.052σ). In practice, the latter value is equivalent to one day of attendance at school within an academic year.

Figure 3.8: Summary of RD Point Estimates and Confidence Intervals for Attendance



(1): Year of outcome is 2013, running variable is PFSE score and p is two
(2): Year of outcome is 2013, running variable is PFSE score and p is one
(3): Year of outcome is 2013, running variable is distance of average grade and p is two
(4): Year of outcome is 2013, running variable is distance of average grade and p is one
(5): Year of outcome is 2014, running variable is PFSE score and p is two
(6): Year of outcome is 2014, running variable is PFSE score and p is one
(7): Year of outcome is 2014, running variable is distance of average grade and p is two
(8): Year of outcome is 2014, running variable is distance of average grade and p is one

Source: own calculations using administrative datasets, Chilean ME & MSD

Given these estimates, if frontier-specific average effects of the BLE in 2013 exist these are most likely to be modest in magnitude. Thus, I am unable to detect them with statistical certainty. These findings also hold after I apply the local randomisation RD framework (in Appendix I) and calculate the effect of the BLE in 2014 (in Appendix J). The RD design provides average estimates that may miss causal effects on some population subgroups. Additionally, the RD estimates are only valid for observations near the threshold, which are not the poorest of the population in this paper. Appendix K analyses the impact of the BLE in 2013 over some population subgroups. This analysis does not consistently show estimates that are statistically different from zero. Therefore, any effects of this kind are unlikely to be large and could not be captured with statistical certainty.

3.5 Conclusion

This paper contributes to the empirical literature on cash for grades impact assessments. I estimate the effect of a Chilean cash for grades programme on subsequent attendance and academic performance. Specifically, I evaluate the impact of the *Bono por Logro Escolar* in 2013. As the cash transfer in 2013 was targeted using two scores from 2012, a relative income index and average grades, it is possible to implement a sharp RD design along these two running variables. The differences in students' outcomes at the two thresholds used have a causal interpretation.

The main causal estimates are not statistically significantly different from zero for both types of outcomes. If anything, the BLE local or frontier-specific average effects are modest and as a result I am unable to detect them with statistical certainty. The size of these potential effects is smaller than those statistically significant effects of near 0.20σ found for interventions of this kind in developing countries (Behrman et al., 2015; Kremer et al., 2009). The results by subgroups do not consistently show estimates that are statistically significantly different from zero.

RD estimates are informative for the population around the thresholds but not necessarily away from them. By design my BLE impact estimates only provide information for students around the poorest 30% threshold and around the top 30% in terms of highest academic achievement. As RD estimates provide average effects for the population near the thresholds, the features of the design do not facilitate observing effects for entire subgroups. For example, we cannot learn about the effects of the BLE for those at the lower end of the income distribution who are at the median or the bottom end in terms of academic achievement. These subgroups of the population may be the ones who are more susceptible and would benefit the most from cash for grades programmes.

Given these caveats it is not possible to generalise my results for the entire population who received the BLE in 2013. Future research could overcome these limitations by introducing some degree of randomisation, allowing for obtaining estimates for the entire population that are expected to be eligible for the cash for grades intervention or entire subgroups of interest.

A possible explanation for the results is that the programme was not very salient for the targeted students. The sample I analyse were not able to collect payments on their own. These were received by an adult member of their household. Moreover, students may not have learnt about the existence of the programme when it was first implemented in 2013. If students were unaware of the implementation of the BLE then it would not be expected to observe changes in their behaviour. If this is still the case, then programme managers could implement actions that increase students' awareness of the benefits (for example, by also giving a diploma to students).

An alternative reason is that students were aware of the cash transfer but unresponsive to its $100 USD maximum size. The monthly minimum wage was $210,000 CLP (approximately $410 USD) in August 2013. If this is a likely scenario then raising the amount of the cash transfer could lead to increased effects. However, whether this is a cost-effective initiative compared to others available deserves more analysis. Another potential explanation for my results is that the BLE provided two types of effects that cancelled each other out overall. The price effect may have incentivised effort (measured by attendance) while a psychological factor could have reduced it.

All these different hypotheses deserve further exploration in future research. Unlike my paper (which addresses the question of whether the BLE worked), this research will need to focus on a different question: why is the BLE producing little effect on educational outcomes? Interviewing parents and students to find out how aware they are of the BLE implementation could prove useful. Additionally, these interviews could help to understand the causal mechanisms (or lack of them) between providing this cash transfer and subsequent improvements in effort and academic performance. Given the nature of this inquiry, this research should probably be qualitative.

Unlike the other programmes analysed in the literature, the BLE treatment assignment was not randomised at a higher level than students. Other schemes had a group of courses or schools that were part of a treatment group while the rest belonged to a control group. Conversely, in Chile within the same course it is possible to observe BLE eligible and ineligible students. Hence, in this context students may compete for access to the BLE. Students that did not receive the programme in 2013 may have observed classmates or other students accessing the cash transfer. This could have influenced awareness about the BLE and future academic performance in terms of accessing the cash transfer in the future (for example in 2014 and 2015). If this hypothesis is correct, then we would expect to observe higher estimates using the PFSE index as a running variable than for average grade (as eligibility by PFSE score is harder to modify by the student relative to academic performance). However, this is not the case and this hypothesis is unlikely to hold.

This paper provides multiple other contributions to the cash for grades evaluation literature. Beyond analysing the effect of a cash for grades intervention in a context of competition, I study the impact on an overall measure of academic performance, the average grade. This contrasts with the subject-specific criteria (for example test scores in maths, reading or writing) commonly found in the previous literature. Then, an additional potential explanation for the lack of results for the BLE is that for students it implies too much effort, or it is too hard to improve in all subjects (or to improve largely in a few subjects) to increase their average grade.

Whether to reward children according to their academic performance remains a hotly debated and unresolved topic. I observe no significant effects on educational outcomes for the first Chilean cash for grades programme. Further research and an enhanced BLE design may be needed to deliver grades for cash. Otherwise, the country risks little return on its money.

Appendix G. Summary of Cash for Grades Evaluations in Schools

Table 3.11: Summary of Cash for Grades Evaluations (Randomised Control Trials) in Primary and Secondary Education

Author(s)	Country/State	Programme Description[37]	Main Results
Kremer et al. (2009)	Kenya	Girls who performed well in academic exams had their school fees paid for two years (7th and 8th grade) and received a grant of $19.20 per year. The scholarship schools were randomly selected from two Kenyan districts. The scholarship was awarded to the highest scoring 15% of 6th grade girls in the programme schools within each district.	An overall effect of 0.19σ is found on academic exams. The results are statistically significant in one out of two districts. Positive externalities are observed among girls with low pre-test scores and for some boys.
Angrist and Lavy (2009)	Israel	Treatment assigned at the school level (among very low performing schools). The programme lasted for three years. The awards were given to high school students. A student who passed all achievement milestones (mainly exams related to obtaining a high school matriculation certificate) could obtain just under $2,400.	Positive results in certification rates for girls (on the order of 0.10 percentage points). The results are mainly driven by the group for whom the certification is "within reach". No effects on boys.
Fryer (2011)	Chicago and New York	The experiment had two cash for grades arms. In Chicago, 9th graders were paid every five weeks upon performance in five core courses. The maximum a student could have won in a year was $2,000. In New York City, 4th and 7th grade students were rewarded based on internal assessments. The maximum amount students could have made in a school-year was $250 and $500, respectively. School-based randomised control trials determined treatment.	In both states, no effects are found in maths or reading achievement tests. In Chicago, marginally significant effects (0.10 of a standard deviation) are observed for grades in the five core subjects. In New York, the effects on the interim assessments are, if anything, negative.
Bettinger (2012)	Ohio	Cash payments (as much as $100 per student) were given to students in the 3rd through to the 6th grade for scoring "proficient" or "advanced" in their (state level) standardised testing. Eligibility by randomisation at the school-grade level.	Positive effects (0.15 of a standard deviation) for maths but no impacts are observed in reading, social science and science test scores.

[37] The currency of all the cash transfers described in Table 3.11 is United States Dollars.

Table 3.11 (continued): Summary of Cash for Grades Evaluations (Randomised Control Trials) in Primary and Secondary Education

Author(s)	Country/State	Programme Description	Main Results
Levitt et al. (2012)	Chicago	Low-income children and adolescents were offered cash ($10 or $20) for an improvement in a computer test score. These tests lasted between 15 to 60 minutes. Randomisation occurred at the class or school-grade level.	An effect of approximately a tenth of a standard deviation is observed for the $20 incentive. No effects for the $10 transfer. Secondary students are more responsive to the size of the transfer relative to elementary students.
Riccio et al. (2013)	New York	Payments, available for three years, were awarded when low-income households met specific education-based conditions of children. Among these conditions were superior attendance at school and certain performance levels in standardised tests. Each child was rewarded with between $600 and $700 per year for scoring proficient or above. The selection of families or households into the programme was made randomly.	The intervention does not improve outcomes for elementary and middle school students but shows effects among high school students who are more academically prepared than their peers (who entered high school as proficient readers).
Behrman et al. (2015)	Mexico	Mexican high schools with over 40,000 students were assigned to three treatment groups and a control group at random. In one of the treatment groups the payments depended on student performance in mathematics tests in 10th to 12th grade. Payments ranged from $227 up to $1,363 (depending on the level of progress in the tests).	Effects ranging between 0.17 and 0.30 of a standard deviation on maths test scores. All the estimates are statistically significant, but these results are partly explained by students copying.
Hirshleifer (2017)	India	The intervention was carried out among school children (4th through 6th grade) as an experiment (randomisation at the classroom-level). Two tests were implemented. The cash transfer, up to $3, depended on the result of the first test. A second test was only used to measure students' learning.	A positive result (0.24 of a standard deviation relative to the control group) is found but this is not statistically significant due to low statistical power.

Appendix H. Using Relative Ranking in a Regression Discontinuity Design

If I use a local randomisation RD framework, then I should expect not to observe statistically significant differences in the pre-treatment variables across the 0.3 threshold in relative ranking. This logic follows what is commonly shown in experimental designs. I choose two tiny windows of relative ranking (*w*=0.002 and *w*=0.005) to implement this RD framework. Therefore, this approach only uses observations whose relative ranking lies within the interval [0.3–*w,* 0.3+*w*]. I obtain the RD estimates in this framework from the following regression:

$$X_i = \alpha_3 + \beta_3 I_{3i} + \varepsilon_{3i},$$

where X_i is a pre-treatment variable for student i in year *t*–1, I_{3i} is a binary variable that takes a value of one if the relative ranking for student i in year *t*–1 is equal to or lower than 0.3. Otherwise, this variable takes a value of zero. ε_{3i} represents the error term of the regression.

Table 3.12 shows β_3. Each estimate is equivalent to the difference in means across the threshold. Panel A shows that students with a relative ranking in the [0.298,0.3] interval notably differ compared to those in the]0.3,0.302] interval. The first group, relative to the second, had higher average grades and attendance levels in 2012. These differences are statistically significant at a 99% level of confidence. Additionally, students in the first group were younger, and most likely to be enrolled in primary education, in a rural school and to belong to smaller cohorts relative to their peers. Panel B shows that the differences become smaller as the window *w* increases, but many estimates remain statistically significant at a 95% level of confidence.

In a continuity-based framework, a discontinuous relationship between the relative ranking and each pre-treatment variable will cast doubt on the plausibility of the continuity assumption. The estimates for the pre-treatment variables in this framework are provided by the regression:

$$X_i = \alpha_4 + \beta_4 I_{4i} + \theta_4 f(\Delta Rel\_Rank_i) + \gamma_4 f(\Delta Rel\_Rank_i) I_{4i} + \varepsilon_{4i},$$

where $\Delta Rel_Rank_i$is 0.3-relative ranking in *t*–1 for student *i,* I_{4i} is a binary variable that takes a value of one if ΔRel_Rank_i is non-negative. Otherwise, the variable takes a value of zero. Additionally, $f(x)$ is a local polynomial function of *x* of order p and ε_{4i} models the error term.

Table 3.13 shows β_4. I present two estimates per variable, both using an ad-hoc data-driven bandwidth and small order polynomials. Multiple statistically significant estimates are observed, for example average grade and attendance in 2012, cohort size and rural school.

The evidence I present in Tables 3.12 and 3.13 suggests that the variation in BLE treatment cannot be considered as good as random near the 0.3 relative ranking threshold. There are systematic differences between students in the neighbourhood of this cutoff. These differences are very likely to affect potential outcomes and will bias the RD estimates in both frameworks.

The relative ranking is the result of a two-step transformation. In the first step the average grade is transformed into a ranking. In the second step the ranking is divided by the cohort size. This procedure defines some characteristics for observations around a relative ranking value of 0.3. Table 3.14 summarises possible values for academic cohort size in relative ranking intervals.

Observations with a relative ranking of 0.3 can only come from a cohort whose size is a multiple of ten. Observations in the [0.298, 0.3[or the]0.3, 0.302] intervals necessarily come from cohorts of at least 57 and 53 students. Expanding the observations to the [0.295, 0.3[and]0.3, 0.305] intervals reduces the minimum cohort sizes to 27 and 23. Therefore, students whose relative ranking is 0.3 come from cohorts with distinctive characteristics relative to those who are very close to this threshold. Table 3.15 provides summary statistics to prove this point.

Students whose relative ranking is 0.3 belong to cohorts whose average size is 43.63. The average cohort size near this threshold increases sharply. Moreover, other characteristics of the schools and students are substantially different. Students with a relative ranking of 0.3 are more likely to be enrolled in a primary school, to be younger and to attend a rural school. Additionally, these students had a higher academic performance and attendance in 2012.

Tables 3.14 and 3.15 provide evidence against comparability between students whose relative ranking is 0.3 and their peers just above and below this threshold. In an RD design for the BLE, students with a relative ranking of 0.3 will be pooled together with those who are slightly below this value. This will attenuate the problems but will not solve them. Overall, there are no comparable observations for students whose relative ranking is 0.3. An additional challenge is that students at each side of the threshold are almost certain to come from different cohorts. Given all this evidence, the relative ranking is not a suitable running variable for an RD design.

Table 3.12: RD Estimates for Pre-Treatment Variables in Local Randomisation Framework

Pre-Treatment Variables (Var.)	(1) av_grade2012	(2) attendance2012	(3) age	(4) schoolprimary	(5) schoolrural	(6) cohort_size	(7) hmonthincome	(8) hhschooling	(9) hhfemale
Panel A: Size of Window w=0.002									
RD Estimate:	0.094***	1.040***	-0.396***	0.142***	0.206***	-48.89***	-10,488.7*	-0.380**	-0.021
Original Var.	(0.023)	(0.359)	(0.134)	(0.039)	(0.016)	(4.07)	(5,494.6)	(0.160)	(0.025)
Number of Observations	2,348	2,348	2,348	2,348	2,348	2,348	2,271	2,271	2,271
Panel B: Size of Window w=0.005									
RD Estimate	0.046***	0.434**	-0.084	0.033	0.098***	-13.03***	-4,222.5	-0.119	-0.012
Original Var.	(0.014)	(0.213)	(0.076)	(0.020)	(0.012)	(2.40)	(3,374.4)	(0.100)	(0.015)
Number of Observations	5,082	5,082	5,082	5,082	5,082	5,082	4,917	4,917	4,917

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Table 3.13: RD Estimates for Pre-Treatment Variables in Continuity-Based Framework (Using Mean Squared Error Optimal Bandwidth)

Pre-Treatment Variables (Var.)	(1) av_grade2012	(2) attendance2012	(3) age	(4) schoolprimary	(5) schoolrural	(6) cohort_size	(7) hmonthincome	(8) hhschooling	(9) hhfemale
Panel A: Order p of Local Polynomial =2 (Local Quadratic Regression)									
RD Estimate:	0.025***	0.282**	-0.085*	0.031**	0.084***	-16.79***	-1,374.0	-0.126**	-0.025***
Original Var.	(0.010)	(0.134)	(0.051)	(0.015)	(0.011)	(2.18)	(1,987.0)	(0.064)	(0.010)
Number of Observations	97,336	143,886	109,592	97,240	78,797	61,704	126,062	127,140	117,446
Bandwidth Size	0.094	0.139	0.106	0.093	0.075	0.060	0.126	0.127	0.117
Panel B: Order p of Local Polynomial =1 (Local Linear Regression)									
RD Estimate:	0.020**	0.232*	-0.037	0.016	0.081***	-14.29***	-739.5	-0.111*	-0.020***
Original Var.	(0.008)	(0.120)	(0.036)	(0.011)	(0.010)	(1.82)	(1,449.6)	(0.057)	(0.007)
Number of Observations	70,027	87,241	101,376	87,304	48,226	42,843	110,176	76,554	102,748
Bandwidth Size	0.068	0.084	0.099	0.084	0.047	0.041	0.110	0.076	0.100

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Table 3.14: Theoretical Academic Cohort Size by Intervals of Relative Ranking

Interval of Relative Ranking	Minimum Academic Cohort Sizes	Comments
0.3	10, 20, 30, 40, 50, 60	Only academic cohorts whose size is a multiple of 10 can have observations whose relative ranking equals 0.3.
[0.298, 0.3[	57, 67, 77, 87, 97	First academic cohort size where it is theoretically possible to have observations in each of these two intervals of relative ranking is 255.
]0.3, 0.302]	53, 63, 73, 83, 93	
[0.295, 0.3[	27, 37, 44, 47, 54, 57	First academic cohort size where it is theoretically possible to have observations in each of these two intervals of relative ranking is 105.
]0.3, 0.305]	23, 33, 43, 46, 53, 56	

Table 3.15: Summary Statistics (Mean Values) by Intervals of Relative Ranking

Variables (Year 2012)	Interval of Relative Ranking				
	0.3	[0.298, 0.3[	]0.3, 0.302]	[0.295, 0.3[	]0.3, 0.305]
Cohort Size	43.63	123.35	115.79	96.01	88.12
Average Grade	5.75	5.64	5.63	5.68	5.66
Attendance (%)	93.81	92.48	92.38	92.76	92.74
Primary School	0.808	0.566	0.595	0.655	0.683
Age (Years)	12.16	12.84	12.76	12.58	12.50
Rural School	0.278	0.060	0.008	0.081	0.062

Source: own calculations using administrative datasets, Chilean ME and MSD

Appendix I. Local Randomisation Regression Discontinuity Framework

Using PFSE Scores as a Running Variable

To implement a local randomisation RD framework, first I need to choose the size w of the window. The size of this window determines which observations of the running variable I use in the estimation. Hence, this RD design relies only on observations within the interval]98–*w*, 98+*w*] of PFSE scores. I choose two values of w (*w*=1 and *w*=2). The justification behind this selection is twofold. Firstly, these values are the two minimum w available (no decimal places are available for the PFSE scores in my dataset). Secondly, these values are still likely to be small enough for the assumption on which the local experiment RD framework relies to hold. I obtain the RD impact estimates (β_5) in this framework from the following regression:[38]

$$Y_i = \alpha_5 + \beta_5 I_{5i} + \varepsilon_{5i},$$

where Y_i is the average grade or attendance for student *i* in year *t* or *t*+1. I_{5i} is a binary variable that takes a value of one if the PFSE score for student *i* in year *t*–1 is equal to or lower than 98. Otherwise, this variable takes a value of zero. The error term corresponds to ε_{5i}.

The first four columns of Table 3.16 show the results for average grade in 2013 and 2014 while the last four columns show the estimates for future attendance. The estimates for future average grade are all negative and statistically insignificant. The estimates range from –0.023 (–0.037σ) to –0.003 points (–0.005σ). The estimates for attendance in 2013 and 2014 are all close to zero and not statistically significant at any level of confidence. The estimates range from –0.244% (–0.026σ) to 0.071% (0.008σ). Overall, all these findings are similar to the ones observed in Table 3.3. For both types of outcomes, the estimates are closer to zero when *w*=2. In standard deviations, any estimate higher than 0.06σ will be statistically significant.

Table 3.17 presents the local randomisation estimates for ten pre-treatment variables. To assess the internal validity of the estimates, I use the same regression but replace Y_i for each pre-treatment variable within X_i'. These variables could not have been affected by the BLE in 2013. This step assesses the quality of the randomisation and, by extension, the internal validity of

[38] This coefficient equals the difference in means in the threshold's neighbourhood. $\beta_5 = \bar{Y}_{]98-\mathrm{w},98].} - \bar{Y}_{]98,98+\mathrm{w}]}.$

the causal estimates. The first two columns show the average grade and attendance in 2012. The third to the tenth columns present characteristics of the students, and their schools and households. Overall, the students at each side of the threshold tend not to differ in terms of these ten variables. For both Panels, the joint test of statistical significance of these ten variables is not rejected.

Why can a Local Randomisation RD Framework not be used for Average Grade?

In its original format the average grade is a continuous variable. A local experiment RD framework could be implemented using this variable in a neighbourhood around the threshold. However, previous academic performance is highly correlated with future average grades and attendance. Consequently, the critical assumption of an RD local experiment framework is only likely to hold in a very small window around the threshold. Students who differ only by a very tiny fraction in their average grade in 2012 are more likely to be comparable. Unfortunately, the average grade in Chile is rounded at the schools and only one decimal place is reported to the central level. The smallest interval of units available is the [*T*–0.1, *T*+0.1[neighbourhood.

Table 3.18 presents the local randomisation RD estimates. These are equivalent to differences in means between units in the [*T*, *T*+0.1[and [*T*–0.1, *T*[intervals of average grade in 2012. Among the poorest 30%, all students in the first group received the BLE in 2013 while the second group did not. Every estimate is statistically significant at a 99% level. These coefficients range from 0.102 to 0.106 points for average grade, and between 0.399% and 0.476% for attendance. These results cannot be interpreted as evidence of treatment effects from the BLE. The variation in treatment is not as good as random within the [*T*–0.1, *T*+0.1[neighbourhood. Table 3.19 shows that students at each side of the threshold within this neighbourhood significantly differ regarding key pre-treatment variables such as their percentage of attendance in 2012 and the years of schooling of the head of their household. By construction, the students differ in their average grade in 2012.

Potential outcomes are likely to be correlated with the running variable in the [*T*–0.1, *T*+0.1[neighbourhood. Given the characteristics of my dataset, it not advisable to implement a local randomisation framework using average grades as a running variable. The estimates from this framework will not be useful, as they will encompass both treatment effects and selection bias.

Table 3.16: RD Estimates for Outcomes in Local Randomisation Framework

Outcomes	average_grade2013		average_grade2014		attendance2013		attendance2014	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
RD Estimate:	-0.023	-0.012	-0.012	-0.003	-0.222	-0.146	-0.244	0.071
Original Outcome	(0.018)	(0.013)	(0.019)	(0.013)	(0.233)	(0.170)	(0.279)	(0.195)
RD Estimate:	-0.037	-0.020	-0.020	-0.005	-0.026	-0.017	-0.026	0.008
In Standard Deviations	(0.028)	(0.020)	(0.030)	(0.021)	(0.027)	(0.020)	(0.030)	(0.021)
Number of Observations	3,539	6,901	3,515	6,840	3,539	6,901	3,515	6,840
Size of Window *w*	1	2	1	2	1	2	1	2

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Table 3.17: RD Estimates for Pre-Treatment Variables in Local Randomisation Framework

Pre-Treatment Variables (Var.)	(1) avg_grade2012	(2) attendance2012	(3) age	(4) male	(5) schoolpub	(6) schoolrural	(7) hmonthincome	(8) hsize	(9) hhschooling	(10) hhfemale
Panel A: Size of Window w=1										
RD Estimate:	-0.019	-0.244	0.098*	0.008	-0.006	-0.006	-1,532.4	-0.014	-0.095	-0.004
Original Var.	(0.013)	(0.192)	(0.058)	(0.017)	(0.017)	(0.010)	(4,759.5)	(0.048)	(0.113)	(0.017)
Number of Observations	3,567	3,567	3,567	3,430	3,567	3,567	3,430	3,430	3,430	3,430
Panel B: Size of Window w=2										
RD Estimate:	-0.009	-0.033	0.064	0.024**	-0.003	-0.002	-1,915.9	0.002	-0.078	0.002
Original Var.	(0.009)	(0.139)	(0.043)	(0.012)	(0.012)	(0.008)	(3,227.2)	(0.034)	(0.082)	(0.012)
Number of Observations	6,959	6,959	6,959	6,722	6,959	6,959	6,722	6,722	6,722	6,722

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Table 3.18: RD Estimates for Outcomes in Local Randomisation Framework (*w*=0.1)

Outcomes	(1) average_grade2013	(2) average_grade2014	(3) attendance2013	(4) attendance2014
RD Estimates: Original Outcome	0.106***	0.102***	0.399***	0.476***
	(0.004)	(0.004)	(0.067)	(0.076)
Number of Observations	56,775	55,659	56,775	55,659

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean ME & MSD

Table 3.19: RD Estimates for Pre-Treatment Variables in Local Randomisation Framework (*w*=0.1)

Pre-Treatment Variables (Var.)	(1) average_grade2012	(2) attendance2012	(3) hmonthincome	(4) hhschooling
RD Estimates: Original Var.	0.126***	0.291***	1,716.8*	0.085***
	(0.003)	(0.055)	(899.4)	(0.027)
Number of Observations	57,806	57,806	55,969	55,969

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean ME & MSD

Appendix J. Effects of Receiving the BLE in 2014 on Educational Outcomes of 2015

The body of the paper shows no statistically significant results for the BLE in 2013. A potential explanation for these results is that students were not aware enough about the implementation of the BLE in 2013. This was less likely to be the case after one year. Students who received the BLE in 2014 could have increased their effort and performance in 2015. This appendix explores the effects of receiving the BLE in 2014. In 2014 the BLE was implemented during September, three months before the end of the academic year. Given this late implementation, I only consider outcomes in 2015 in this assessment.

In practice, the estimates for the BLE in 2014 remain not statistically different than zero. For average grade, Panel A in Table 3.20 shows that the estimates range from –0.016 to 0.004 points. The standard errors of these estimates range from 0.010 to 0.020 points. Panel B in Table 3.20 presents the results for attendance. The estimates range from –0.042% to 0.041% while the standard errors are between 0.132% and 0.279%. The findings for the BLE in 2014 are like those in 2013. I find no statistically significant effects. Consequently, if an effect exists for the BLE, it is likely to be small in magnitude and cannot be detected statistically.

Table 3.20: RD Estimates for Outcomes in 2015

RD Framework and Running Variable (RV)	Local Randomisation RV: PFSE		Continuity-Based RV: PFSE		Continuity-Based RV: Average Grade	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Estimates for Average Grade in 2015						
RD Estimates:	-0.010	-0.007	-0.016	-0.013	-0.002	0.004
	(0.020)	(0.014)	(0.012)	(0.011)	(0.013)	(0.010)
Number of Observations	3,094	6,023	103,352	67,235	133,102	89,748
Window *w* / Bandwidth *h*	1.00	2.00	33.84	21.77	0.300	0.200
Panel B: Estimates for Attendance in 2015						
RD Estimates	-0.009	-0.025	0.039	0.041	-0.042	0.021
	(0.279)	(0.197)	(0.185)	(0.132)	(0.205)	(0.152)
Number of Observations	3,094	6,023	82,467	82,467	133,102	89,748
Window *w* / Bandwidth *h*	1.00	2.00	27.40	27.14	0.300	0.200
Order *p* of Polynomial	NA	NA	2	1	2	1

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean ME & MSD

Appendix K. Effects of Receiving the BLE in 2013 for Population Subgroups

I analyse the effect of receiving the BLE in 2013 on educational outcomes in 2014 for multiple subgroups. I divide the first two groups by income. The hypothesis to test in this case is that the BLE is most likely to have an impact at the lower end of the income distribution. There is limited information on this question for cash for grades programmes, though Galiani and McEwan (2013) and Maluccio and Flores (2005) find in Honduras and Nicaragua, respectively, that conditional cash transfers have a stronger impact on school enrolment among children living in the poorest households.

Analysis by Income (PFSE Scores)

Table 3.21 shows RD estimates for average grade and attendance in 2014 using average grade in 2012 as a running variable. The first row of Table 3.21 presents the results without using income subgroups. These results are equivalent to those shown in Table 3.7. Among those with a PFSE score equal to or lower than 98 points, I divide the sample into two. The first half is meant to include the poorest (approximately the bottom 15% in terms of income) while the other represents the second most deprived group (approximately between the bottom 15% to 30% in the income distribution).

Table 3.21: RD Estimates for Outcomes in 2014 by Halves of Lowest PFSE Scores in Continuity-Based Framework (Using Average Grade in 2012 as a Running Variable)

Subgroup	Outcomes			
	Average Grade in 2014		Attendance in 2014	
Total	0.010	0.017*	0.119	0.147
	(0.012)	(0.009)	(0.188)	(0.141)
1st Poorest Half	0.020	0.023**	0.234	0.277
	(0.016)	(0.012)	(0.257)	(0.185)
2nd Poorest Half	-0.000	0.010	-0.008	0.010
	(0.016)	(0.012)	(0.256)	(0.185)
Bandwidth *h*	0.300	0.200	0.300	0.200
Order *p* of Polynomial	2	1	2	1

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean ME and MSD

Table 3.21 shows that the first group has more positive estimates than the second group. The impact estimates for average grade in 2014 are 0.020 and 0.023 for the poorest. Only one estimate is statistically significant at a 95% level of confidence. Therefore, the significance is not robust to alternative specifications. The impact estimates for attendance in 2014 are 0.234% and 0.277% for the poorest, neither of which is statistically significant. Conversely, the RD estimates for the second poorest half are close to zero for both types of outcomes.

There is not enough evidence to claim that the BLE in 2013 had a statistically significant effect in 2014 on the poorest of the population. Given the standard errors shown, 0.012 and 0.016 for average grade and 0.185% and 0.257% for attendance, if an effect of the BLE exists on the poorest then it is at most modest in size and could not be captured consistently with statistical certainty.

Analysis by Gender and Educational Level

The next two groups are boys and girls. The final two groups of students are those who were either in the fifth or sixth grade and between seventh and tenth grade in 2012, respectively. In 2014 the former group was most likely to be enrolled in primary education while the latter group was most likely to be enrolled in secondary education. These two pairs of groups are justified in terms of the cash for grades literature, where heterogeneous results have been observed.

Table 3.22 and Table 3.23 present RD estimates for average grade and attendance in 2014, respectively. The first four columns in each table provide estimates that use PFSE scores as a running variable. The first and second columns focus on the local randomisation framework, while the third and fourth rely on the continuity-based framework. The last two columns of each table provide estimates for the continuity-based framework using the distance of average grade in 2012 as a running variable. The first row of each table provides the estimates without using subgroups. These estimates are equal to those presented in Tables 3.3, 3.7 and 3.16.

From Table 3.22 it can be seen that girls have less negative estimates than boys in all the specifications and frameworks I use and that students between the seventh and tenth grades have mostly lower estimates relative to younger students. Only the local linear regressions using average grade as a running variable provide some statistically significant estimates. The estimates by subgroup mostly remain close to zero and are statistically insignificant. Overall, these estimates are not very different from those of the entire sample.

There is less consistent behaviour by subgroup for attendance in 2014. Table 3.23 shows that the RD estimates for boys are lower when I use the PFSE as a running variable but higher when the estimations rely upon previous academic performance. I observe a similar case for students in the fifth and sixth grades in 2012 relative to their peers in higher grades. Two estimates for the former subgroup show some degree of statistical significance in one type of RD framework.

Overall, the estimates by gender and educational level remain close to zero and are statistically insignificant for both outcomes. Naturally, the analysis by subgroup has wider confidence intervals given the smaller samples. The standard errors for average grade vary from a maximum of 0.029 points to a minimum of 0.012 points. Thus, the analysis by the subgroups of gender and educational level could have identified any estimate higher than 0.057 and as low as 0.024 points in average grade as statistically significant. Concerning attendance, the standard errors of the estimates range from 0.181% to 0.406%. Differences at the threshold higher than 0.800% and as low as 0.355% could have been statistically significant then.

For both types of outcomes, in the few cases where I observe statistically significant estimates, the significance is sensitive to the RD framework I use. Additionally, within each RD framework, the significance is also sensitive to the specification I utilise. Given my results and analysis, if an effect of the BLE exists by gender and educational level it is highly unlikely to be large. Any potential effect of this kind could not be captured with statistical certainty in this assessment.

Table 3.22: RD Estimates for Average Grade in 2014 (by Subgroups)

Subgroups	RD Framework and Running Variable					
	Local Randomisation RV: PFSE		Continuity-Based RV: PFSE		Continuity-Based RV: Average Grade	
	(1)	(2)	(3)	(4)	(5)	(6)
Total	-0.012	-0.003	-0.008	-0.003	0.010	0.017*
	(0.019)	(0.013)	(0.013)	(0.010)	(0.012)	(0.009)
Boys	-0.030	-0.011	-0.009	-0.010	0.010	0.017
	(0.029)	(0.021)	(0.020)	(0.015)	(0.017)	(0.013)
Girls	-0.003	0.006	-0.004	0.003	0.018	0.024**
	(0.026)	(0.018)	(0.017)	(0.012)	(0.016)	(0.012)
5th to 6th Grade in 2012	0.008	-0.002	-0.000	0.007	0.023	0.028**
	(0.029)	(0.020)	(0.019)	(0.014)	(0.017)	(0.013)
7th to 10th Grade in 2012	-0.017	-0.001	-0.011	-0.009	0.008	0.014
	(0.024)	(0.017)	(0.017)	(0.013)	(0.016)	(0.012)
Window *w* / Bandwidth *h*	1.00	2.00	24.96	24.87	0.300	0.200
Order *p* of Polynomial	NA	NA	2	1	2	1

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean ME & MSD

Table 3.23: RD Estimates for Attendance in 2014 (by Subgroups)

Subgroups	RD Framework and Running Variable					
	Local Randomisation RV: PFSE		Continuity-Based RV: PFSE		Continuity-Based RV: Average Grade	
	(1)	(2)	(3)	(4)	(5)	(6)
Total	-0.244	0.071	0.108	0.106	0.119	0.147
	(0.279)	(0.195)	(0.159)	(0.143)	(0.188)	(0.141)
Boys	-0.359	-0.049	-0.053	-0.060	0.268	0.142
	(0.391)	(0.279)	(0.236)	(0.214)	(0.270)	(0.205)
Girls	-0.164	0.101	0.178	0.177	-0.044	0.114
	(0.406)	(0.272)	(0.210)	(0.187)	(0.257)	(0.184)
5th to 6th Grade in 2012	-0.197	0.022	-0.012	-0.008	0.438*	0.388**
	(0.382)	(0.284)	(0.215)	(0.192)	(0.254)	(0.181)
7th to 10th Grade in 2012	-0.201	0.130	0.164	0.160	0.021	0.067
	(0.362)	(0.249)	(0.208)	(0.187)	(0.246)	(0.185)
Window *w* / Bandwidth *h*	1.00	2.00	35.58	21.12	0.300	0.200
Order *p* of Polynomial	NA	NA	2	1	2	1

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean ME & MSD

Chapter 4 (Don’t) Call me by Your Name: Reassessing Threats to Internal Validity in Regression Discontinuity Designs

Abstract

This paper elaborates on threats to internal validity, administrative sorting and intermediate contamination, which have been overlooked in the regression discontinuity (RD) literature. Lee and Lemieux (2010) claim that if individuals are unable to precisely manipulate the running variable then variation in treatment near the threshold is as good as random. The paper shows that variation in treatment assignment is not always as good as random even without manipulation. This can be the case when administrative procedures, beyond individuals’ control and knowledge, affect their position near the threshold non-randomly. If administrative sorting is not recognised it can be mistaken as manipulation. Timing also matters in RD designs. Intermediate contamination can emerge if a substantial time lag exists between the realisation of the running variable and its use in treatment. In this case, the paper highlights the value of checking variables related to this time lapse in RD falsification tests.

4.1 Introduction

In all RD designs some exogenous variation in treatment assignment occurs around a threshold of a running variable. Two types of RD designs exist: sharp and fuzzy. In sharp RD designs, treatment is entirely explained by the running variable. In fuzzy RD designs, the running variable is a relevant factor, but not the only one, in explaining treatment status.

In the context of RD designs to evaluate the impact of cash transfers, the running variable generally corresponds to an index and the threshold is a specific index score. The index score of each household (or individual) corresponds to their value in the running variable. The index is used as the eligibility criteria for the cash transfer. A sharp RD design is suitable for causal inference only if all households that meet the eligibility criteria (generally having a score below the threshold) receive the cash transfer and if the households that do not meet the criteria (those with a score above the threshold) never receive the cash transfer.

In RD designs, causal inference relies on the assumption that the average outcome for units marginally at one side of the threshold represents a valid counterfactual for the group just at the other side of threshold (Hahn et al., 2001; Lee, 2008). This rationale allows us to interpret any discontinuity in the conditional expectation of the outcome (as a function of the running variable) at the threshold as causal evidence of the treatment (Imbens & Lemieux, 2008).

RD designs are a popular approach for causal inference. Consequently, the methodological literature in RD designs has been developing quickly over the last decade.[39] Despite its extensive use for causal inference, practical applications of RD are partly determined by researchers' interpretation of the methodology (Sekhon & Titiunik, 2017). Researchers follow different steps in RD applications although there is a common understanding that treatment assignment is as good as random in a neighbourhood close to the threshold.

As Cattaneo et al. (2018a) explain, applied practitioners of RD designs choose between a local randomisation or a continuity-based framework. Researchers who adopt the former framework use the logic of experimental designs to recover causal estimates. They take the simple average of the outcome in a small window on either side of the index' threshold. Hence, the impact estimate is equivalent to the difference in means across the cutoff point. Conversely, researchers relying on the continuity-based framework use regression at each side of the threshold to predict the limiting value of the outcome precisely at the threshold.

Further elaboration about the relative merits, required assumptions and practical applications of these two frameworks is beyond the scope of this paper. Instead, I introduce these frameworks to support the idea that central aspects of RD designs are still subject to debate. Further developments and clarifications of the RD design represent a valuable contribution.

This paper highlights two threats to internal validity that have been relatively overlooked in the RD literature. By doing so, the paper contributes to further conceptual clarification of RD designs. The first threat, administrative sorting, has received little attention and therefore represents the most novel part of the paper. The second, intermediate contamination, reinforces a threat that has been broadly identified but can still be overlooked in RD falsification tests. I introduce these two threats empirically. I select this approach, relative to simulations, to illustrate RD design challenges in real-world contexts. In this paper, these

[39] For a detailed summary see Cattaneo and Escanciano (2017).

are three evaluations of conditional cash transfers (CCTs) on adolescents' school enrolment in Chile. The paper's findings are useful for empirical applications in both RD frameworks.

Manipulation of the running variable is one of the two main conceptual concerns in the application of RD designs (Imbens & Lemieux, 2008). The issue is that individuals with a stake might try to manipulate the running variable close to the threshold (Angrist & Pischke, 2009; Imbens & Lemieux, 2008; McCrary, 2008; Skovron & Titiunik, 2015) and sort themselves around it (Cattaneo, Jansson, & Ma, 2017; Lee & Lemieux, 2010). Manipulation threatens the plausibility of the continuity assumption on which RD designs rely.[40] Thus, the analysis of individuals' ability to precisely manipulate the running variable becomes central. This rationale explains why the running variable density test proposed by McCrary (2008) has been interpreted, as the title of his article suggests, as a test of manipulation of the running variable. Furthermore, Imbens and Lemieux (2008) later stated that: "a discontinuity (in the density) is suggestive of violations of the non-manipulation assumption" (p.27).

One point that is highlighted in Lee and Lemieux's (2010) influential paper in RD designs is that "if individuals are unable to precisely manipulate the running variable, then variation in treatment (assignment) near the threshold is as good as random" (p.283). However, my paper shows that variation in treatment assignment is not always as good as random in the absence of manipulation and that a McCrary test can fail for different reasons than individuals' induced sorting. This can be the case when administrative procedures, beyond the control and knowledge of individuals, affect their position near the threshold non-randomly, and by doing so, these actions threaten the continuity assumption of RD designs.

Sorting is not a new concept in the RD methods literature. However, the concept has been intrinsically tied to individuals' manipulation (McCrary, 2008), optimising behaviour in response to rules (Lee & Lemieux, 2010), and self-selection (Cattaneo et al., 2017). In other words, manipulation and sorting have been used as synonyms of deliberate human action to

[40] In the continuity-based RD framework, the running variable can be associated with average potential outcomes, but this association is assumed to be smooth or continuous at the threshold (Imbens & Lemieux, 2008). This is the continuity assumption. Meanwhile, the underlying assumption on which the local randomisation RD framework relies is that the average potential outcomes are uncorrelated with the running variable within a small neighbourhood close to the threshold (Cattaneo et al., 2018b). The continuity-based framework is most commonly employed in practice (Cattaneo et al., 2018a). Given this, I will use the continuity assumption concept generically throughout this document. Unless I state otherwise, the concept will refer to the vital assumption needed for any RD framework to provide unbiased causal estimates.

locate themselves across the threshold. In this sense, administrative sorting is different from the threat of manipulation.[41] Both represent a threat to the continuity assumption in RD designs. However, manipulation is deliberate while administrative sorting is not intentional.

I define then administrative sorting by three features. The first is being the result of administrative procedures that affect the position of individuals near the threshold non-randomly. The second is that these actions are beyond the control and knowledge of individuals and do not deliberately intend to locate individuals across the threshold. The third is that these procedures threaten the continuity assumption on which RD designs rely.

Administrative sorting is not the only threat preventing variation in treatment being as good as random in the absence of individuals' manipulation. Timing also matters in RD designs. If a substantial time lag exists between the realisation of the running variable and its actual use to assign the treatment of interest, then intermediate contamination could emerge. In this scenario, units near the threshold are no longer comparable before the treatment of interest occurs, and consequently, RD designs risk providing biased estimates.

The second main conceptual concern in RD designs highlighted by Imbens and Lemieux (2008) is "the possibility of other changes at the same threshold of the running variable" (p. 631). The example that Imbens and Lemieux (2008) give concerns individuals who become eligible for discounts at cultural institutions at the age of 65. In this case, using age as a running variable in an RD design to estimate the effects of discounts is problematic. Changes in attendance at cultural institutions at one side of the threshold might be explained by other factors, such as free transport, which also affect individuals who have just turned 65.

The most common guidelines given about how to conduct falsification tests in RD designs are to observe placebo (Skovron & Titiunik, 2015) or pseudo outcomes (Imbens & Lemieux, 2008), variables determined after treatment but known not to be affected by the treatment, and to test the quality of the random assignment for variables determined before treatment.[42]

[41] Camacho and Conover (2011) discuss potential cases of manipulation of a targeting system for social welfare programmes. The authors argue that the index in which the system relies on, in theory, can be manipulated in alternative ways to make individuals eligible for social benefits. Among these are respondents lying in answers that affect their index score or people in positions of power changing the answers of respondents or their final index scores. All these examples represent deliberate human actions to locate individuals across the threshold.

[42] The falsification tests of each RD framework are run in the same way as for the outcomes of interest.

For the latter group of variables, multiple definitions exist. Some authors use the concept of pre-treatment (Angrist & Pischke, 2009; Cattaneo & Escanciano, 2017) or pre-determined variables/covariates (Cattaneo et al., 2018a; Skovron & Titiunik, 2015). Other authors use the term baseline characteristics (Lee, 2008; Lee & Lemieux, 2010; McCrary, 2008), variables determined before the realisation of the running variable or randomisation.

These guidelines make sense when the realisation of the running variable and treatment occur close together (for example elections and politicians taking office or test scores and scholarship decisions). In this context the definition of baseline characteristics is equivalent to pre-treatment variables. In these cases, observing variables determined before treatment will adequately assess the quality of randomisation near the threshold. As well, assessing placebo outcomes can help to determine whether other changes at the same threshold took place during or after treatment.

However, these guidelines are less appropriate when the running variable was in place well before treatment assignment and proper treatment took place. In this case, it might be preferable to classify pre-treatment or pre-determined variables between: i) baseline characteristics, and ii) intermediate variables, those that could have been affected by the running variable, especially around the threshold of interest, before the treatment occurred.

This classification is not aimless. The first group of variables, unless any type of sorting occurred, should have a continuous or smooth distribution at the threshold at the moment of the realisation of the running variable. However, only the second group of variables could have been affected by intermediate contamination, changes happening at one side of the threshold between the realisation of the running variable and treatment. These changes are caused by the usage of the running variable and threshold within the described time frame.

Going back to the previous example, let us assume that no other changes happen when individuals turn 65 years old apart from becoming eligible for discounts at cultural institutions. In the RD falsification tests we use diverse demographics (such as education and race) as pre-treatment variables and time spent in health-related activities as a placebo outcome. This approach may miss other changes that happened in the past, for example a one-off incentive to retire given precisely one year ago, which affected only individuals aged 64 or older. In this new example, the RD estimates of discounts on attendance at cultural events are likely to be contaminated by the one-off incentive to retire given in the past.

Intermediate contamination poses a threat to RD designs that use age as a running variable at a later life stage, for example those evaluating the effect of the legal minimum drinking age. Age depends on the date of birth, which affects the timing of entry into education (Blanden, Del Bono, Hansen, & Rabe, 2017; Crawford, Dearden, & Meghir, 2010; Elder, 2010; Fredriksson & Ockert, 2005). Del Bono and Galindo-Rueda (2007) explain intuitively and mathematically how intermediate contamination may affect length of school estimates. Despite some researchers' awareness of this threat, the current RD falsification test guidelines may overlook it. This highlights the importance of strengthening RD designs that rely on running variables determined long before treatment by providing evidence that observations at one side of the threshold are not affected by intermediate contamination.

The three main conceptual contributions of this paper are as follows: i) administrative sorting and intermediate contamination are overlooked threats to internal validity in RD designs, ii) lack of individuals' manipulation does not translate automatically into variation in treatment assignment being as good as random, and iii) distinguishing among pre-treatment variables (between baseline characteristics and intermediate variables) in falsification tests is beneficial to assess whether intermediate contamination is affecting the internal validity of the RD design.

An additional contribution of this paper is that it raises issues that applied RD practitioners can encounter. Administrative sorting and intermediate contamination may invalidate RD designs in contexts where indexes such as proxy means tests (PMTs) are used as running variables.[43] In Colombia, Barrientos and Villa (2015) use a running variable and threshold previously used by a school fee reduction programme (Barrera-Osorio, Linden, & Urquiola, 2007) to assess the impact of the CCT *Familias en Acción*. In Ecuador, Ponce and Bedi (2010) do not consider whether administrative sorting explains why more individuals are found above the threshold of a CCT. The authors only claim that this difference could not be attributed to manipulation. Centro de Microdatos (2012) faces a related situation in Chile. At least 20 developing countries use PMTs for targeting (Australian Aid, 2011; Brown et

[43] A proxy means test refers to a system or situation where information correlated with income is used in a formula to proxy income. The formula (parameters and weights) is obtained through statistical analysis and tends to use data that is easily observable by public officials (Coady et al., 2004; Grosh & Baker, 1995).

al., 2016; Coady et al., 2004). Hence, this paper could also prove useful for designing new social policies using PMTs that are expected to be evaluated with RD designs.

My findings highlight the importance of fully understanding the data generation process of a running variable. If administrative rules, not individual manipulation, explain the shape of an index density near the threshold then useful variation for identification may still exist. Potential solutions can arise then from an appropriate diagnostic. For example, the causes behind administrative sorting can potentially be fixed for future versions of the index and facilitate the implementation of an RD design. In other cases, when treatment probabilities are not affected, administrative sorting may not invalidate a retrospective RD design entirely. If administrative sorting is not recognised correctly, it can be mistaken as manipulation, which could potentially lead to viable present or future research RD designs being discarded.

The structure of the paper is as follows. The second and third sections present the cases studied. Both sections have a similar aim, which is to introduce one example each of administrative sorting and intermediate contamination when implementing an RD design. The second section focuses on a CCT called BARE, which used the IVSE index for targeting. The third section builds on two CCTs (SUF and AS) that used the SPF index, although with different thresholds, to allocate the transfers. Both sections have a similar structure. The last section of the paper integrates the discussion of the previous three sections and concludes.

4.2 Case 1: *Beca de Apoyo a la Retención Escolar*

This section presents the case of *Beca de Apoyo a la Retención Escolar* (BARE). The first subsection describes the rollout of the CCT and the IVSE index used for targeting. The second part presents the data sources and provides summary statistics. The third subsection develops a case of administrative sorting. In this case, administrative sorting invalidates the RD design as individuals near the IVSE threshold used by BARE are not similar. The fourth subsection uses a factor of the IVSE index, day of birth, in an RD design.[44] This subsection elaborates on a threat to identification associated with intermediate contamination.

[44] I use the concepts date of birth and day of birth in this chapter. The date of birth refers to the exact date in which an individual is born, including the year. The day of birth or birthday only considers the day and month.

4.2.1 *Programme and Targeting Description*[45]

BARE Overview and Rollout

BARE is the smallest CCT in Chile; it reached only 18,000 students in 2014. This cash transfer is focused exclusively on secondary school students. The goal of BARE is to encourage these students to stay in school. In 2014, it consisted of an annual total monetary contribution of $178,000 CLP (approximately $280 USD on June 30th 2015).

The target population of BARE is secondary school students at risk of dropping out. This concept is operationalised through the IVSE index. Each year those enrolled secondary school students with the highest IVSE scores become eligible to be BARE new entrants. Eligible students are identified at the national level at the beginning of the academic year. Soon after the year starts, the people in charge of BARE in the field have to actively locate these students in their schools and encourage them to apply. There is no perfect match between eligible students and BARE recipients, as not all eligible students are encouraged to apply and a small fraction of non-eligible students end up accessing BARE.

Due to BARE's limited budget, new entrants only come from "targeted schools". These have been derived from a list of schools with high school dropout rates. Only public and private subsidised schools in the traditional education system are targeted. No student from the adult education system or private schools can be a new BARE entrant. Additionally, by design, students in their last grade of secondary education cannot be a BARE new entrant.

The IVSE Index

In 2014 and 2015 BARE's targeting index, the IVSE, was derived from six variables that are highly correlated with observing a future drop out. I present these variables in Table 4.1.

[45] The information in this subsection comes from an official report (Opazo et al., 2015).

Table 4.1: IVSE Variables

Variables	Description
Attendance	Attendance rate (%) in the previous academic year
$\lfloor \Delta age \rfloor$[46]	Difference between age and expected age for the grade (months)
Welfare Recipient	Student's household in *Chile Solidario* or *Ingreso Ético Familiar*
Paternity	Student is a father or a mother
Pregnancy	Student is pregnant
Mother Schooling	Mother's years of schooling

Source: author's compilation based on official documents, Ministry of Education

Each of the variables of the index was transformed into a score using a function $f_k(\cdot)$. After applying $f_k(\cdot)$, IVSE scores for each student *i* were calculated using the following formula:

$$\begin{aligned} IVSE_i = {} & 0.165 f_1(Attendance_i) + 0.135 f_2(\lfloor \Delta age_i \rfloor) + 0.21 f_3(WelfareRecipient_i) \\ & + 0.21 f_3(Paternity_i) + 0.21 f_3(Pregnancy_i) \\ & + 0.07 f_4(MotherSchooling_i) \end{aligned}$$

The maximum possible IVSE score was 182.650 and the minimum score was 30.775. The threshold used to determine eligibility changed between 2014 and 2015. In 2014 those with a score equal to or above 84.625 were eligible to be a new entrant for BARE, while in 2015 only those students with a score no lower than 84.150 became eligible for BARE.

4.2.2 *Data, Sample and Descriptive Statistics*

Data and Sample

My analysis focuses on the subgroup of potential BARE new entrants in 2014 and 2015. The sample contains only students who were: i) not BARE renewals, ii) enrolled in one of the first three grades of secondary education, and iii) enrolled in a "targeted school".

My analysis relies on administrative data provided by *Junta Nacional de Auxilio Escolar y Becas* (JUNAEB). At my request JUNAEB created two datasets (2014 and 2015 cohorts), which contain for each student from the ninth to eleventh grades: i) personal information (such as gender, date of birth, and previous academic performance), ii) each variable of the IVSE index, iii) their IVSE score, iv) whether the student became a BARE recipient during the year, and, v) future academic performance and enrolment in secondary education.

[46] $\lfloor X \rfloor$ represents the floor function. This function returns the integer part of a number. For example, $\lfloor 2.9 \rfloor$=2.

Descriptive Statistics

Table 4.2 provides descriptive statistics by cohort and by IVSE scores. Panel A shows that adolescents scoring above the IVSE score used to determine eligibility were more likely to receive BARE afterwards. In 2014, 49.2% of adolescents scoring no less than 84.625 in the IVSE were recipients of the CCT later in the year. Conversely, less than 1% of adolescents scoring below this threshold ended up being beneficiaries. In 2015 (with a threshold of 84.150) these percentages are 40.1% and 1.8%, respectively.

Panel B shows that students scoring above the thresholds had lower levels of attendance, were more likely to be lagging in their pathway to graduation, were more likely to belong to a household whose members were welfare recipients, were more likely to be parents or pregnant, and had mothers with fewer years of schooling. These facts are unsurprising as these six variables are part of the IVSE formula.

Panels C and D show that adolescents above the IVSE threshold were (compared to those below this threshold): slightly less likely to be male and more likely to be older, more likely to be enrolled in a public school and ninth grade. The group with higher IVSE scores also had a lower average grade. Additionally, a smaller fraction had progressed to the following grade in the previous year. Finally, minor differences across groups exist in the proportion who were enrolled in technical-professional schools and resided in the metropolitan region.

Table 4.2: Descriptive Statistics (Mean Values) by BARE Cohorts and IVSE scores

	BARE Cohorts								
Variables	2014			2015			2014 & 2015		
	< 84.625	>=84.625	Total	< 84.150	>=84.150	Total	< Cutoff	>=Cutoff	Total
Panel A: IVSE Score and BARE Recipient									
IVSE	48.92	94.29	51.60	50.55	96.02	53.55	49.72	95.19	52.55
BARE	0.005	0.492	0.034	0.018	0.401	0.043	0.011	0.445	0.038
Panel B: IVSE Variables									
Attendance (%)	89.5	68.4	88.3	90.6	70.1	89.2	90.0	69.3	88.7
$\lfloor \Delta age \rfloor$ (Months)	0.39	15.31	1.27	-0.13	14.87	0.85	0.14	15.08	1.07
Welfare Recipient	0.236	0.935	0.278	0.301	0.927	0.342	0.268	0.931	0.309
Parent	0.001	0.063	0.005	0.001	0.132	0.010	0.001	0.099	0.007
Pregnancy	0.001	0.058	0.005	0.000	0.064	0.004	0.001	0.061	0.004
Mother Schooling (Years)[47]	9.54	7.92	9.45	10.07	8.28	9.95	9.81	8.11	9.70
Panel C: Demographic Information									
Male	0.525	0.494	0.523	0.513	0.498	0.512	0.519	0.496	0.517
Age (Years)	16.06	17.19	16.13	16.00	17.17	16.08	16.03	17.18	16.10
Metropolitan Region	0.287	0.278	0.287	0.261	0.263	0.261	0.275	0.270	0.274
Panel D: Academic Information									
Technical-Professional School	0.513	0.506	0.512	0.519	0.513	0.518	0.516	0.510	0.515
Public School	0.694	0.788	0.700	0.708	0.775	0.712	0.701	0.781	0.706
Ninth Grade	0.335	0.407	0.339	0.350	0.402	0.353	0.342	0.404	0.346
Tenth Grade	0.347	0.316	0.345	0.336	0.314	0.334	0.341	0.315	0.340
Eleventh Grade	0.319	0.277	0.316	0.314	0.284	0.312	0.317	0.280	0.314
Previous Average Grade	5.20	4.28	5.15	5.31	4.73	5.27	5.25	4.51	5.21
Progressed Previous Year	0.905	0.585	0.886	0.924	0.624	0.904	0.914	0.605	0.895
Number of Observations	245,307	15,372	260,679	233,442	16,445	249,887	478,749	31,817	510,566

Source: own calculations using administrative datasets, Chilean Ministry of Education

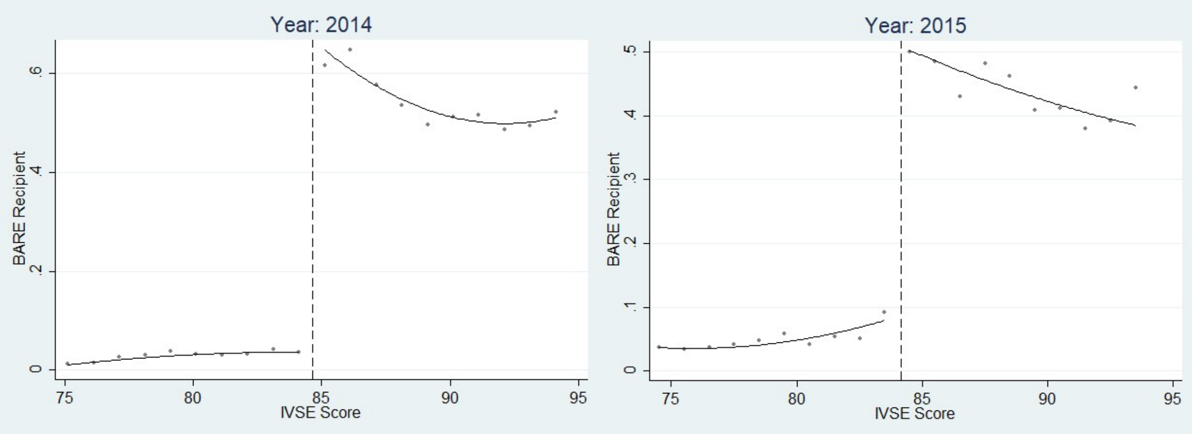
[47] This variable has a 6% to 7% of observations that are not available. The percentage of data that is not available does not vary depending on IVSE scores (above and below the cutoff).

4.2.3 *Administrative Sorting Invalidates the RD Design*

Can IVSE be used as a Running Variable?

A researcher interested in obtaining the causal relationship between being a BARE recipient and future enrolment in secondary education needs a suitable identification strategy. An intuitive approach is to implement a fuzzy RD design using IVSE scores as a running variable. Figure 4.1 shows that in principle this method is appealing. IVSE scores are a strong predictor of being a recipient of BARE in 2014 and 2015. In particular, BARE recipient status changes abruptly after crossing the respective IVSE threshold used each year.

Figure 4.1: BARE Recipient by IVSE Scores (by Different Years)[48]



Source: own calculations using administrative datasets, Chilean Ministry of Education

However, having a strong relationship between the running variable and participation in BARE is not sufficient to recover unbiased estimates in a fuzzy RD design. It is also required that, at least close to the threshold, the variation in eligibility for BARE is as good as random. If RD falsification tests are passed there is an increased likelihood that the continuity assumption will hold. If I use a local randomisation RD framework, then I should show balance in pre-treatment variables as in experimental designs. Not observing balance among these variables would cast doubt on the comparability of units close to the cutoff point.

Table 4.3 presents comparisons of means for pre-treatment variables for adolescents just

[48] I derive the continuous lines from a quadratic regression. The parameters of the regressions vary at each side of the threshold. The dots correspond to the average of BARE recipient for each non-overlapping bin of one point of the IVSE score. The vertical dashed lines are the thresholds used by BARE to determine eligibility.

above and just below the IVSE thresholds. These two groups of adolescents differ substantially. For example, in 2014 adolescents scoring between 84.625 and 85 in the IVSE (and therefore who were eligible for BARE) had on average (relative to adolescents scoring between 84.150 but lower than 84.625): lower levels of attendance in the previous year (0.068 percent points), a lower $\lfloor \Delta age \rfloor$ (10.13 months) and mothers with more years of schooling (3.78 years). All these differences are statistically significant at a 99% level of confidence.

A similar phenomenon can be observed in 2015. Adolescents scoring between 84.150 but lower than 84.625 (who were thus eligible for BARE) had different pre-treatment variables relative to those scoring above 83.500 and below 84.150. The former group had on average (relative to the latter group): more levels of attendance in the previous year (0.077 percentage points), a higher $\lfloor \Delta age \rfloor$ (13.17 months) and mothers with fewer years of schooling (4.68 years).

Statistically significant differences between groups can be observed not only in variables that are part of the IVSE index but also in other demographic and academic characteristics. For example, in 2014, adolescents just above the threshold (compared to those just below): were more likely to live in the metropolitan region, were less likely to be in eleventh grade, and had on average lower grades and rates of progress to the following grade in 2013. In 2015, those adolescents just above the threshold (relative to those just below): were less likely to live in the metropolitan region, were more likely to be enrolled in a technical-professional school, were more likely to be in the eleventh grade and had a higher academic performance in 2014. All these differences are significant at a 99% level of confidence.

To provide unbiased causal estimates, the local randomisation framework requires independence between potential outcomes and the running variable within a neighbourhood near the threshold (Cattaneo et al., 2018b). A potential explanation for the lack of balance in average pre-treatment variables is that the neighbourhood I use is too large. In other words, as the IVSE neighbourhood increases it becomes less likely to observe variation in treatment assignment being as good as random and more likely to observe differences in pre-treatment variables below and above the threshold. To avoid this problem, I compare these variables only at the threshold where it should be guaranteed that variation is as good as random.

The continuity-based framework in RD designs is suitable for this purpose. In this approach, the distribution of each pre-treatment variable along the running variable should not show a discontinuity exactly at the threshold. If there are no such discontinuities, then it is more likely that variation in eligibility is as good as random at the IVSE threshold.

Using local regressions, I test for discontinuities for multiple pre-treatment variables following Cattaneo et al. (2018a). The regression model I use in this part is as follows:

$$X_{1i} = \alpha + \beta I_{1i} + \gamma Z_{1i} + \theta Z_{1i} I_{1i} + \omega {Z_{1i}}^2 + \delta {Z_{1i}}^2 I_{1i} + \varepsilon_{1i} ,$$

where X_{1i} is a pre-treatment variable for adolescent i, Z_{1i} is the difference between the IVSE score for adolescent i and the IVSE threshold. I_{1i} is an indicator function, which takes the value of one if $Z_{1i} \geq 0$ and zero otherwise. ε_{1i} represents the error term of the regression.

Table 4.4 presents the results of the RD falsification tests for the continuity-based framework. The first two columns show the results for the 2014 BARE cohort. The next two columns focus in the 2015 group while in the last two columns I pool both BARE cohorts. Each pair of estimates vary by the size of bandwidth I use in the regression. Almost all the β coefficients are statistically significant at a 99% level of confidence. Therefore, I cannot assume that a continuous conditional distribution exists for each pre-treatment variable at the IVSE threshold used by BARE. Hence, the continuity assumption is unlikely to hold in this case and the treatment estimates are likely to be biased.

Table 4.3 and Table 4.4 suggest that the variation in BARE eligibility in 2014 and 2015 cannot be considered as good as random near each IVSE threshold. There are systematic differences in adolescents across the thresholds each year. For example, in 2015 adolescents just above the threshold (compared to those just below) had a higher $\lfloor \Delta age \rfloor$ and attendance. These differences will affect impact estimates in ways that are not straightforward to anticipate as these variables are negatively and positively correlated with future school enrolment, respectively.

To summarise, this part of the subsection has shown that the IVSE scores are not useful as a running variable in an RD design. The following part explains the reason for these findings.

Table 4.3: Mean Values by IVSE Scores and RD Estimates for Pre-Treatment Variables in Local Randomisation Framework

Pre-Treatment Variables	BARE Cohorts							
	2014				2015			
	< 84.625 & >=84.15	>=84.625 & <85	Difference in Means	p-value <	< 84.15 & >83.5	>=84.15 & < 84.625	Difference in Means	p-value <
Panel A: IVSE Variables								
Attendance (%)	96.1	89.3	-6.8	0.000	88.5	96.2	7.7	0.000
$\lfloor \Delta age \rfloor$ (Months)	27.63	17.50	-10.13	0.000	15.08	28.25	13.17	0.000
Welfare Recipient	0.986	0.979	-0.007	0.408	0.984	0.986	0.001	0.875
Parent	0.011	0.011	0.000	0.998	0.016	0.014	-0.001	0.875
Mother Schooling (Years)	5.63	9.42	3.78	0.000	10.86	6.18	-4.68	0.000
Panel B: Demographic Information								
Male	0.609	0.609	0.000	1.000	0.610	0.650	0.040	0.224
Age (Years)	18.35	17.24	-1.102	0.000	17.09	18.29	1.203	0.000
Metropolitan Region	0.065	0.197	0.132	0.000	0.252	0.117	-0.135	0.000
Panel C: Academic and School Information								
Technical Professional School	0.552	0.506	-0.047	0.144	0.485	0.593	0.107	0.001
Public School	0.683	0.724	0.041	0.157	0.739	0.689	-0.050	0.102
Ninth Grade	0.411	0.393	-0.017	0.581	0.352	0.439	0.086	0.009
Tenth Grade	0.218	0.419	0.201	0.000	0.418	0.219	-0.199	0.000
Eleventh Grade	0.371	0.188	-0.183	0.000	0.230	0.342	0.112	0.000
Previous Average Grade	5.07	4.83	-0.238	0.000	4.93	5.08	0.157	0.000
Progressed Previous Year	0.827	0.734	-0.093	0.001	0.765	0.801	0.035	0.207
Number of Observations	353	793	1,146		579	351	930	

Source: own calculations using administrative datasets, Chilean Ministry of Education

Table 4.4: RD Estimates for Pre-Treatment Variables in Continuity-Based Framework (Observed IVSE as a Running Variable)

Pre-Treatment Variables	BARE Cohorts (Columns Additionally Vary by Bandwidth Size)					
	2014		2015		2014 & 2015	
Panel A: IVSE Variables						
Attendance (%)	-1.574***	4.724***	7.566***	12.072***	3.575***	8.163***
	(0.368)	(0.274)	(0.294)	(0.227)	(0.231)	(0.174)
$\lfloor \Delta age \rfloor$ (Months)	-5.116***	6.403***	13.337***	15.871***	5.579***	11.104***
	(0.511)	(0.324)	(0.391)	(0.282)	(0.315)	(0.211)
Mother Schooling (Years)	0.565***	-0.587***	-0.210	-2.967***	0.155	-2.028***
	(0.181)	(0.111)	(0.211)	(0.130)	(0.140)	(0.087)
Panel B: Non IVSE Variables						
Male	0.029	0.064***	0.112***	0.139***	0.072***	0.100***
	(0.028)	(0.018)	(0.025)	(0.017)	(0.018)	(0.012)
Metropolitan Region	0.039	-0.144***	-0.104***	-0.174***	-0.053***	-0.160***
	(0.025)	(0.015)	(0.021)	(0.014)	(0.016)	(0.010)
Ninth Grade	-0.011	0.134***	0.100***	0.227***	0.037**	0.176***
	(0.026)	(0.017)	(0.024)	(0.016)	(0.017)	(0.012)
Previous Average Grade	-0.101***	-0.149***	0.027	-0.034*	-0.053***	-0.106***
	(0.032)	(0.022)	(0.027)	(0.019)	(0.020)	(0.014)
Progressed Previous Year	-0.018	-0.107***	-0.046**	-0.110***	-0.040***	-0.109***
	(0.021)	(0.011)	(0.018)	(0.010)	(0.013)	(0.007)
Bandwidth Size	5	10	5	10	5	10
Number of Observations	15,540	41,183	18,306	51,598	33,846	92,781

Standard errors in parentheses

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

Source: own calculations using administrative datasets, Chilean Ministry of Education

Understanding the IVSE´s Lack of Suitability as a Running Variable

A potential reason for my previous findings is that adolescents sorted themselves above the threshold, making themselves eligible for BARE. However, IVSE scores are hard for adolescents to manipulate as the variables from which the index is derived are difficult to manipulate. For example, mother´s schooling, parenthood and pregnancy decisions are unlikely to be affected by BARE. Δage is a variable that partly depends on the date of birth. Being a welfare recipient depends on the characteristics of the adolescent´s household and is determined after complex and long assessments.

The variable that adolescents could exercise more control over is attendance. However, manipulation of attendance does not imply necessarily precise manipulation of IVSE scores. Even if some adolescents could manipulate their attendance, they would need to have precise knowledge of the structure of the IVSE formula, the weights assigned to each variable (which are not public), and the threshold to be used in the future by BARE (which changed between 2014 and 2015) to precisely manipulate their eligibility.

For all the given reasons, the lack of suitability of the IVSE as a running variable for an RD design seems surprising. Features of the design of the IVSE cause adolescents just above and just below the threshold to be dissimilar. Table 4.5 illustrates this point.

Table 4.5: Three Types of Adolescents With IVSE Scores Close to the BARE Threshold

Variables	Adolescent Type		
	Adolescent A	Adolescent B	Adolescent C
Panel A: IVSE Score and Eligibility			
IVSE Score	83.925	84.150	84.625
Eligible in 2014	No	No	Yes
Eligible in 2015	No	Yes	Yes
Panel B: IVSE Variables (All types are welfare recipients, but not parents nor pregnant)			
Attendance	Between 85% and 94%	95% or over	Between 85% and 94%
$\lfloor \Delta age \rfloor$ (Months)	Between 13 and 24	25 or over	Between 13 and 24
Mother Schooling	12 Years	1 to 7 Years	9 to 11 Years

Adolescents who are close in terms of IVSE scores near the threshold are, by design, different in regard to fundamental characteristics. In 2014, type C adolescents ended up being eligible for BARE while type B adolescents (one of the closest in terms of IVSE) ended up not being eligible. Both types of adolescents differ in their values of attendance,

$\lfloor \Delta age \rfloor$ and their mothers' schooling. A similar situation happened in 2015, when type B adolescents were eligible, but type A adolescents were not. As in the previous example these two types of adolescents are entirely different in regard to the three mentioned variables.

In practice, the IVSE formula generates clusters of adolescents. Adolescents who have the same score in the IVSE also share the same category of values for attendance, $\lfloor \Delta age \rfloor$ and mother schooling. This logic explains the results of Table 4.3. In 2014, the comparison of adolescents just above and below the threshold is primarily between type C adolescents and type B. Likewise, in 2015 the comparison is mainly between type A and type B adolescents.

These points become clear from Table 4.6, which describes the functions $f_1(\cdot)$, $f_2(\cdot)$, $f_3(\cdot)$, and $f_4(\cdot)$ that transform all IVSE variables into the scores that are used in the IVSE formula (shown in subsection 4.2.1). Table 4.6 shows that attendance is divided into five categories. On the one hand, the minimum score given by $f_1(\cdot)$ is 5, which corresponds to students whose attendance is between 95% and 100%. On the other hand, the maximum score returned by $f_1(\cdot)$ is 165. There are four categories for $\lfloor \Delta age \rfloor$ and a similar number of scores associated with $f_2(\cdot)$. In this case, the minimum and maximum scores reach 30 and 135, respectively.

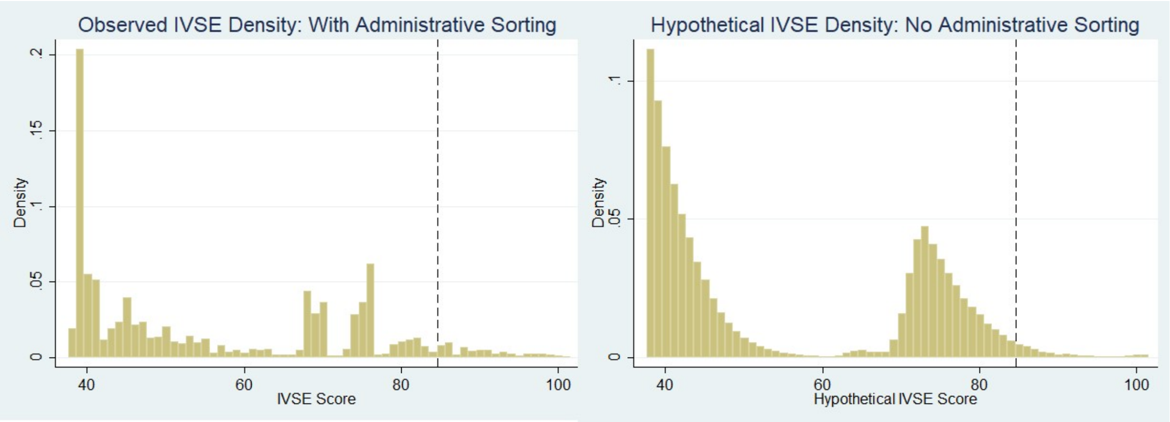
Table 4.6: Functions $f_k(\cdot)$ Transforming IVSE Variables into Scores

$f_k(\cdot)$	Variables	Formula	Score
$f_1(\cdot)$	Attendance	if Attendance is between 0% and 50%	165
		if Attendance is between 51% and 74%	125
		if Attendance is between 75% and 84%	85
		if Attendance is between 85% and 94%	45
		if Attendance is between 95% and 100%	5
$f_2(\cdot)$	$\lfloor \Delta age \rfloor$	if $\lfloor \Delta age \rfloor$ is lower or equal than 6 months	30
		if $\lfloor \Delta age \rfloor$ is between 7 and 12 months	65
		if $\lfloor \Delta age \rfloor$ is between 13 and 24 months	100
		if $\lfloor \Delta age \rfloor$ is equal to or higher than 25 months	135
$f_3(\cdot)$	Welfare Recipient	if No	40
	Paternity	if Yes	210
	Pregnancy		
$f_4(\cdot)$	Mother Schooling	if Mother Schooling is 0	70
	(Years)	if Mother Schooling is between 1 and 7	60
		if Mother Schooling is 8	50
		if Mother Schooling is between 9 and 11	40
		if Mother Schooling is 12	30
		if Mother Schooling is between 13 and 17	20
		if Mother Schooling is equal to or higher than 18	10

Source: author's compilation based on official documents, Ministry of Education

The design of the IVSE leads to an irregular distribution of scores - see the left panel of Figure 4.2. The IVSE distribution could have been smoother. The latter argument can be understood by looking at the right-hand panel of Figure 4.2 in which I show a hypothetical distribution of IVSE scores. The only difference between this hypothetical distribution and the original is that I replace $f_1(\cdot)$ and $f_2(\cdot)$ by two new functions $f_1'(\cdot)$ and $f_2'(\cdot)$, respectively.

Figure 4.2: Observed IVSE Density and Hypothetical IVSE Density (Year 2014)



Source: own calculations using administrative datasets, Chilean Ministry of Education

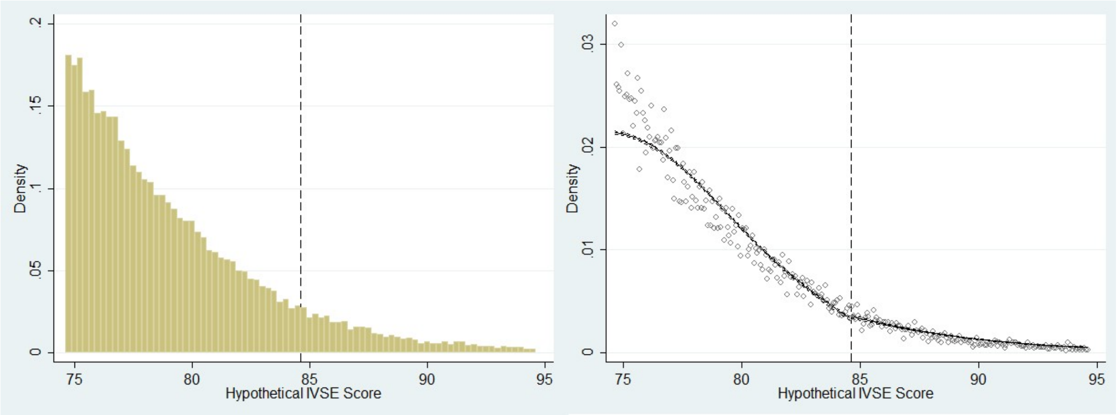
The irregular distribution of the IVSE scores is mainly explained by the design of $f_1(\cdot)$ and $f_2(\cdot)$, the functions that transform attendance and $\lfloor \Delta age \rfloor$ into scores respectively. Table 4.7 presents $f_1'(\cdot)$ and $f_2'(\cdot)$, the functions that I use to estimate hypothetical IVSE scores. The logic I utilise to design $f_1'(\cdot)$ and $f_2'(\cdot)$ follows two criterions. Firstly, to preserve the original scale of IVSE, $f_1'(\cdot)$ and $f_2'(\cdot)$ return the same minimum and maximum scores as $f_1(\cdot)$ and $f_2(\cdot)$, respectively. Secondly, $f_1'(\cdot)$ and $f_2'(\cdot)$ intend to break with the categorisation of attendance and $\lfloor \Delta age \rfloor$ of $f_1(\cdot)$ and $f_2(\cdot)$. Instead, in these new functions, I use almost the entire range of values available for each variable to return a unique score for each value.

Table 4.7: Functions $f_k'(\cdot)$ Transforming IVSE Variables into Scores

$f_k'(\cdot)$	Variables	Formula	Min. Score	Max. Score
$f_1'(\cdot)$	Attendance	5+1.6(100–Attendance)	5	165
$f_2'(\cdot)$	$\lfloor \Delta age \rfloor$	30 if $\lfloor \Delta age \rfloor$ is lower than –13 months 135 if $\lfloor \Delta age \rfloor$ is higher than 37 months 30+2.1($\lfloor \Delta age \rfloor$+13) otherwise	30	135

The left panel of Figure 4.3 shows the hypothetical distribution of the IVSE (from the right-hand panel of Figure 4.2) but close to the BARE threshold. The right-hand panel of Figure 4.3 shows that a McCrary test does not reject this new distribution. No discontinuity can be observed in the density of the hypothetical IVSE index at the threshold used by BARE.

Figure 4.3: Hypothetical IVSE Density & McCrary Test (Year 2014)



Source: own calculations using administrative datasets, Chilean Ministry of Education

Table 4.8 presents the RD continuity-based estimates for the pre-treatment variables using the hypothetical IVSE index as a running variable instead of the observed IVSE scores. The organisation of Table 4.8 follows a similar pattern as Table 4.4. Almost no coefficient is statistically significant at the 95% level. Figure 4.4 shows that the conditional distribution of each pre-treatment variable varies smoothly along the hypothetical index, showing no statistically significant discontinuities at the BARE threshold.[49]

The IVSE index is an inappropriate running variable for an RD design due to the design of $f_1(\cdot)$ and $f_2(\cdot)$. These functions transform discrete variables into categorical ones to build the index. The design of the IVSE affects non-randomly the position of adolescents near the threshold. The continuity assumption is unlikely to hold when using this index. Without administrative sorting, a researcher could have found that variation close to the threshold was as good as random. The continuity assumption is likely to have held. For BARE administrative sorting, not manipulation, undermines the RD design. In summary, this finding demonstrates that lack of ability to manipulate a running variable does not automatically translate into randomised variation in treatment assignment near the threshold.

[49] I derive the continuous lines in each graph from a quadratic regression. The parameters of the regressions vary at each side of the threshold (the vertical dashed line). The dots correspond to the average of the pre-treatment variable for each non-overlapping bin of one point of the hypothetical IVSE score. The dashed lines surrounding the continuous lines in each graph represent the 95% confidence interval of each polynomial fit.

Table 4.8: RD Estimates for Pre-Treatment Variables in Continuity-Based Framework (Hypothetical IVSE as a Running Variable)

Pre-Treatment Variables	BARE Cohorts (Columns Additionally Vary by Bandwidth Size)					
	2014		2015		2014 & 2015	
Panel A: IVSE Variables						
Attendance (%)	-0.368	0.075	0.432	-0.114	0.064	-0.019
	(0.610)	(0.344)	(0.538)	(0.303)	(0.404)	(0.228)
$\lfloor \Delta age \rfloor$ (Months)	-0.512	-0.071	0.165	-0.088	-0.165	-0.098
	(0.578)	(0.321)	(0.490)	(0.278)	(0.376)	(0.211)
Mother Schooling (Years)	0.061	-0.100	-0.363*	-0.168	-0.168	-0.134
	(0.194)	(0.140)	(0.220)	(0.162)	(0.149)	(0.109)
Panel B: Non IVSE Variables						
Male	-0.008	-0.017	0.005	0.007	-0.001	-0.004
	(0.030)	(0.022)	(0.028)	(0.020)	(0.020)	(0.015)
Metropolitan Region	-0.006	-0.006	-0.049**	-0.012	-0.029*	-0.009
	(0.026)	(0.018)	(0.024)	(0.017)	(0.018)	(0.012)
Ninth Grade	-0.028	0.005	0.019	-0.006	-0.002	-0.001
	(0.029)	(0.021)	(0.027)	(0.019)	(0.020)	(0.014)
Previous Average Grade	0.036	0.007	0.051	-0.001	0.045*	0.004
	(0.038)	(0.025)	(0.034)	(0.022)	(0.025)	(0.017)
Progressed Previous Year	0.001	0.016	-0.002	-0.000	-0.000	0.008
	(0.027)	(0.016)	(0.024)	(0.014)	(0.018)	(0.010)
Bandwidth Size	5	10	5	10	5	10
Number of Observations	12,630	37,827	14,292	45,803	26,922	83,630

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean Ministry of Education

Figure 4.4: Pre-Treatment Variables by Hypothetical IVSE Scores

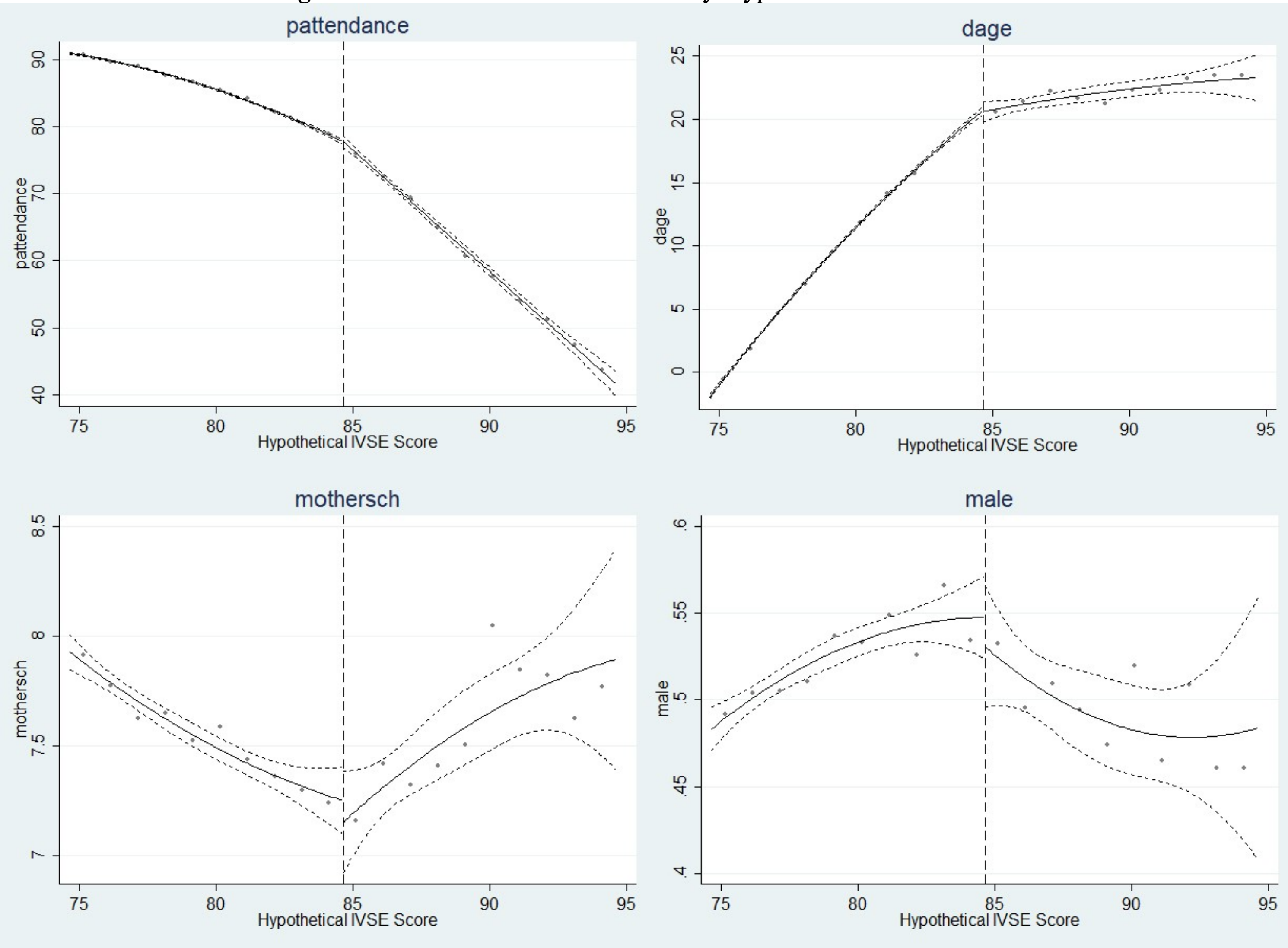
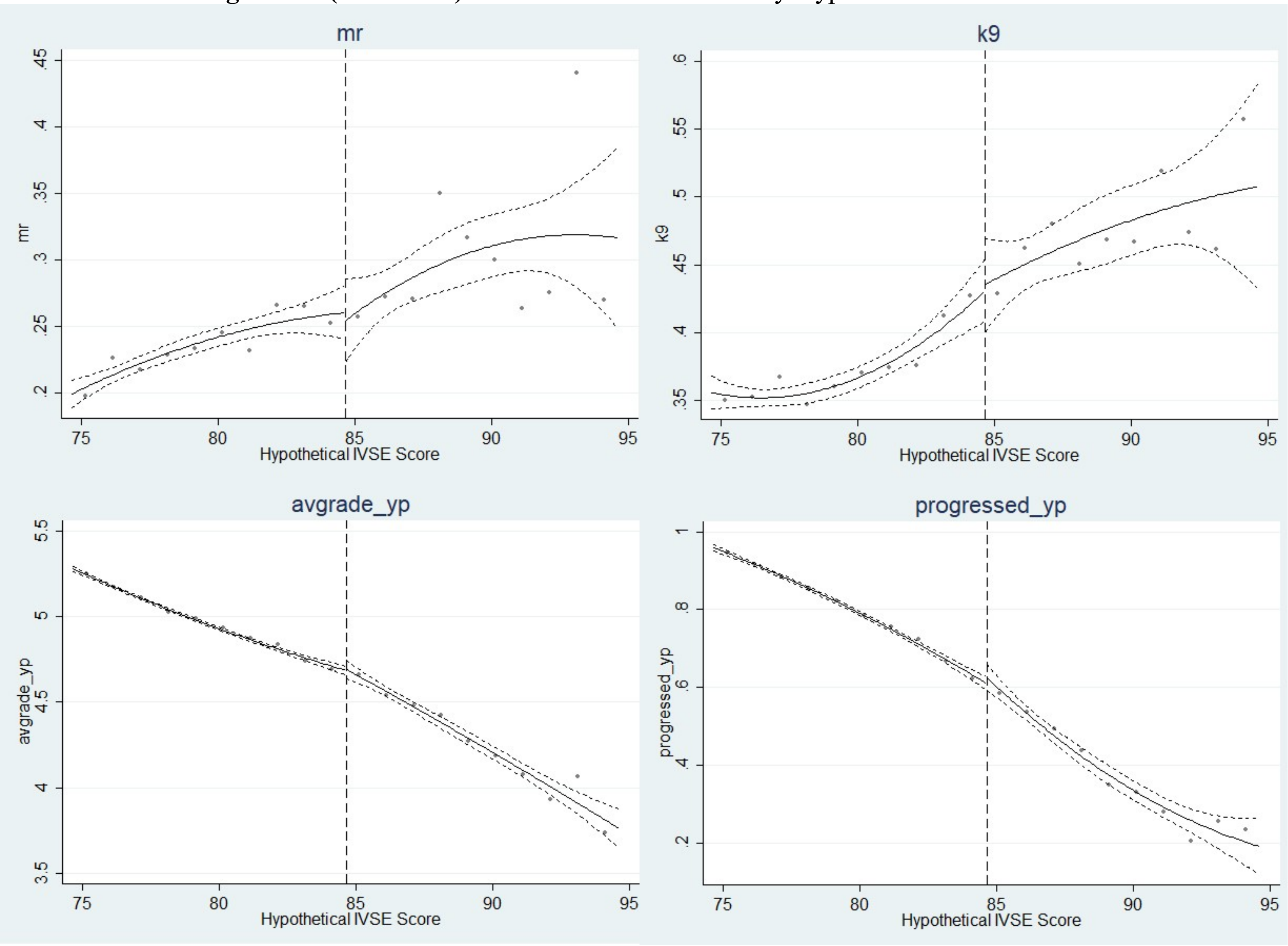


Figure 4.4 (continued): Pre-Treatment Variables by Hypothetical IVSE Scores



Source: own calculations using administrative datasets, Chilean Ministry of Education

4.2.4 *Intermediate Contamination Threatens the RD Design*

Can Day of Birth be Used as a Running Variable rather than the IVSE?

The existence of administrative sorting for the IVSE does not rule out the use of an RD design to recover the causal effect of BARE on school enrolment. A careful analysis of the sources of variation in BARE eligibility opens up new options, as I explain in this subsection.

A promising feature for an RD design corresponds to the rules associated with $\lfloor \Delta age \rfloor$. The first rule is the creation of categorical boundaries when using $f_2(\cdot)$. Table 4.6 shows that the boundaries for $\lfloor \Delta age \rfloor$ are 7, 13 and 25 months. The second rule is that the IVSE formula is based on $\lfloor \Delta age \rfloor$, not Δage. The reference day used to calculate $\lfloor \Delta age \rfloor$ is June 30th. There is no difference in $\lfloor \Delta age \rfloor$ for two students who are in the same grade, one of whom has a birth date on the 1st and the other on the 30th of the same month and year. However, there is a difference of one month in $\lfloor \Delta age \rfloor$ between two students in the same grade, where one's birth date is the 30th of a given month while the other's is the 31st of the same month.

Because of the floor function and the categorical boundaries created by $f_2(\cdot)$, a difference of one day in the day of birth can impact the score derived from $f_2(\lfloor \Delta age \rfloor)$, and along these lines, eligibility for BARE and access to the CCT. For example, a student with Δage of 7 months has a higher probability of being a BARE recipient than a student with Δage slightly less than 7 months. Individuals born on November 30th, who are also exactly seven months behind with respect to their expected age for the grade, end up receiving a score of 65 from $f_2(\lfloor \Delta age \rfloor)$. In contrast, those born on December 1st, who are also almost seven months behind, receive a score of 30. All other characteristics being equal, the first group have an IVSE score 4.725 points higher than the second group and are more likely to receive BARE.

A similar situation occurs at the two categorical boundaries of 13 months and 25 months. In these cases, the key day of birth corresponds to May 30th. Adolescents born on this day receive 35 more points from $f_2(\lfloor \Delta age \rfloor)$ than those adolescents whose birthday is on May 31st or June 1st (among the group 13 or 25 months behind their expected age for the grade).

Δage is affected by two factors. The first is the number of years that a student is behind concerning his or her expected age for the grade. For example, students who are not behind have values of $\lfloor \Delta age \rfloor$ of between –12 and –1 months. Students who are one year behind have values of $\lfloor \Delta age \rfloor$ of between 0 and 11 months. Where in these ranges they are located depends on their birthday. For example, an adolescent who is one year behind and born between November 1st and 30th has a value of 7 in $\lfloor \Delta age \rfloor$, whereas an adolescent who is one year behind and born between December 1st and 30th has a value of 6 in $\lfloor \Delta age \rfloor$.

The BARE datasets contain the exact birth date of the students. Hence, I can use variation in day of birth to identify the causal effect of BARE on school enrolment. In this case the running variable corresponds to $\Delta days$ (the difference in days between the birthday and the day associated with the categorical boundary). Local variation in BARE caused by variation in $\Delta days$ can be thought of as exogenous variation, at least close to the point where $\Delta days$ equals zero. The running variable $\Delta days$ can be defined in two ways as explained below:

Table 4.9: Definition of $\Delta days$ by Categorical Boundaries of $\lfloor \Delta age \rfloor$ and Day of Birth

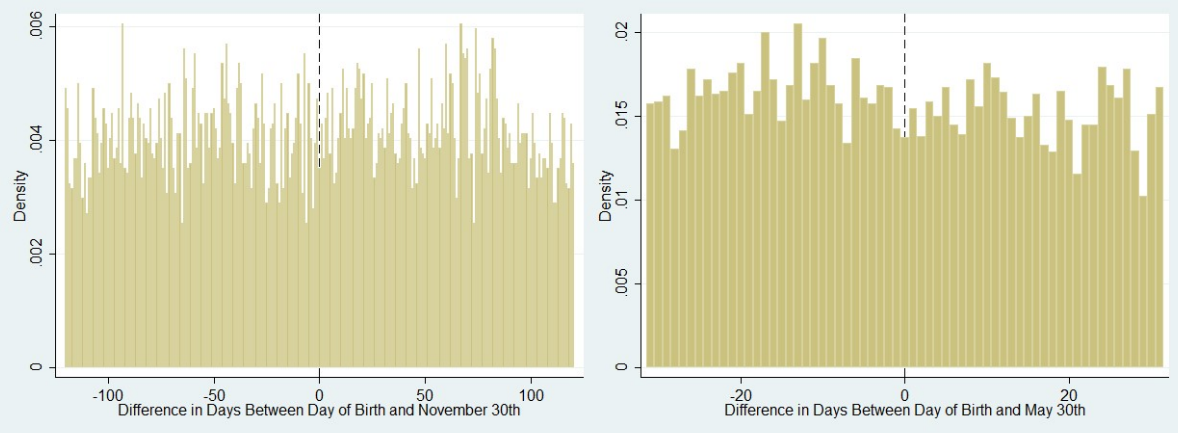
$\Delta days$	Categorical Boundaries of $\lfloor \Delta age \rfloor$	
	7 Months	13 & 25 Months
$\Delta days$<0	From March 31st ($\Delta days$ = –121 or –122) until December 1st ($\Delta days$ = –1)	From June 30th ($\Delta days$ = –31) until May 31st ($\Delta days$ = –1)
$\Delta days$=0	November 30th	May 30th
$\Delta days$>0	From November 29th ($\Delta days$ = 1) until July 1st ($\Delta days$ = 152)	From May 29th ($\Delta days$ = 1) until April 1st ($\Delta days$ = 59)

For the first boundary ($\lfloor \Delta age \rfloor$=7) the reference day of birth is November 30th ($\Delta days$=0). Adolescents who were born between July 1st and November 29th have positive values for $\Delta days$. Conversely, those born between December 1st and March 31st have negative values for $\Delta days$. For the second and third boundaries, the reference day of birth is May 30th. Those born between April 1st and May 29th have positive values for $\Delta days$ while their peers whose birthday is between May 31st and June 30th have negative values for $\Delta days$. Adolescents with a positive value for $\Delta days$ (or zero) had a higher probability of receiving BARE.

My selection of the overall range of dates for the first boundary (from July 1st until March 31st) and the second and third boundaries (from April 1st until June 30th) is consistent with the admission rules in the Chilean Education system, as I explained in section 1.4.

Variation in $\Delta days$ should be a good source from which to identify the treatment effect because it is improbable that it could have been manipulated precisely. Birth dates are hard to manipulate and are related to a period when BARE did not exist. To highlight this, Figure 4.5 presents $\Delta days$ density. The left and right-hand panels use November and May 30th as reference birthdays, respectively. Neither panel shows a higher density of observations at one side of the threshold (the dashed line), suggesting that there is no sorting around it.

Figure 4.5: Distribution of $\Delta days$ (by Reference Day of Birth)



Source: own calculations using administrative datasets, Chilean Ministry of Education

In some cases, using variation in the day of birth to identify a causal effect is not free of risks. For example, Gans and Leigh (2008) show that births are not evenly distributed over the week in Australia. The authors claim that nearly one-third of births that would have occurred on a weekend were moved to a weekday and that this trend is explained by the rise of caesarean sections and inductions. Even if weekends do not play a role in explaining the day of birth, these could affect the availability of professional support. Fitzsimons and Vera-Hernández (2016) show that in the United Kingdom babies born on weekends were less likely to be breastfed as some mothers take advantage of hospital staff to start this process.

If any of these situations are prevalent in Chile, then students born on weekdays could result in being different from those born on weekends. The continuity assumption may not hold, and RD designs will provide biased estimates. For BARE, "weekend-birth effects" are less likely to be a matter for concern. In the sample, I pool together adolescents born in different years (I use two BARE cohorts where each cohort has students from ninth, tenth and eleventh grade). In other words, in the sample adolescents whose birthday is on May 30th or May 31st were not born on a specific day of the week (for example Sunday and Monday, respectively).

Age Requirements to Entry Primary School Contaminates the New Running Variable

Following the standard approach in RD falsification tests, if variation in $\Delta days$ is as good as random in a neighbourhood around the threshold, we should observe balance or continuity for the pre-treatment variables. For $\Delta days$ there is no need to extrapolate to assess continuity for these variables at the threshold. Rather, the RD local randomisation framework is more suitable than the continuity-based framework.[50] In the former framework, balance is tested through differences in the means for observations just above and below the threshold.

Table 4.10 presents the results for the second reference day of birth (when $\Delta days$=0 on May 30th) for those adolescents whose scores are close to the IVSE threshold. I show the calculations for different $\Delta days$ windows. Each row in the table represents a different pre-treatment variable, while each column contains a different window size. Thus, each cell is equivalent to the difference in the means of a pre-treatment variable between observations with positive (including zero) and negative values of $\Delta days$ for a given range of birthdays.

There are almost no statistically significant coefficients in Table 4.10. There are no differences larger than 0.2% in previous attendance, 0.04 in years of mothers' schooling and two percentage points in gender, course grade, type of school and the rate of progress in the previous year. Consequently, at first glance $\Delta days$ is a suitable running variable. The new running variable does not seem to be prone to manipulation and there is balance among the different pre-treatment variables for those adolescents whose birthday is close to May 30th.

However, a more detailed analysis reveals some problems. Table 4.11 shows that adolescents who were born before or on May 30th are on average 0.097 to 0.176 years younger than those whose birthday is after May 30th (with results varying depending on the range of days used for the estimation). All these differences are statistically significant at a 95% level. Also, the F-statistic of joint significance for all the pre-treatment variables is statistically significant.

[50] The local randomisation framework assumes independence between potential outcomes and the running variable within a neighbourhood around the threshold. In this context, this means assuming that differences in the day of birth (or relative age) are uncorrelated with the potential outcomes in the neighbourhood.

The most likely explanation for these findings is the discontinuous relationship between date of birth and age to enrol in primary school in Chile. May 30th or May 31st are not fixed days determining the age requirements to access primary education in the country. However, for children born between April 1st and June 30th, the age at which they can enter school depends on the discretion of academic directors. In practice, Chilean children born on June 1st, or some days later, are more likely to enrol in the first grade of primary education at an older age than children whose birthday is in the final days of May (McEwan & Shapiro, 2007). Thus, Chilean children born in May are more likely to be the youngest in their cohorts.

Using $\Delta days$ as a running variable in an RD design for BARE offers initial promise. However, intermediate contamination partly rules out the approach. The new running variable largely depends on the date of birth. The running variable, which was determined long ago, and the threshold affected other determinants before BARE was implemented. As a result, observations on either side of the threshold become incomparable. This fact does not rule out the RD design. The main problem appears if these determinants are not neutral regarding their relationship with the outcomes of interest. In this case, age is likely to be an explanatory factor of secondary school enrolment. In this scenario the continuity assumption at the new threshold is unlikely to hold and RD estimates are most likely to be biased.

Table 4.10: RD Local Randomisation Estimates for Pre-Treatment Variables

Pre-Treatment Variables	Window Size				
	10	20	31	45	59
Panel A: IVSE Variables					
Attendance (%)	-0.171	-0.081	-0.067	-0.086	-0.064
	(0.107)	(0.077)	(0.062)	(0.057)	(0.054)
Mother Schooling (Years)	-0.032	-0.040	-0.021	-0.033	-0.006
	(0.102)	(0.073)	(0.060)	(0.054)	(0.051)
Panel B: Non IVSE Variables					
Male	-0.005	0.016	0.003	-0.004	-0.002
	(0.018)	(0.013)	(0.010)	(0.010)	(0.009)
Metropolitan Region	0.003	0.016	0.009	0.006	0.004
	(0.014)	(0.010)	(0.008)	(0.007)	(0.007)
Technical Professional School	0.021	-0.003	0.004	-0.002	0.001
	(0.018)	(0.013)	(0.010)	(0.009)	(0.009)
Public School	0.025	0.010	-0.001	0.003	0.001
	(0.016)	(0.011)	(0.009)	(0.008)	(0.008)
Ninth Grade	0.009	0.009	0.007	0.008	0.004
	(0.017)	(0.012)	(0.010)	(0.009)	(0.009)
Tenth Grade	-0.012	-0.007	-0.006	-0.007	-0.009
	(0.017)	(0.012)	(0.010)	(0.009)	(0.009)
Previous Average Grade	-0.027	-0.014	-0.020*	-0.024**	-0.027***
	(0.019)	(0.014)	(0.011)	(0.010)	(0.010)
Progressed Last Year	-0.003	-0.007	-0.009	-0.008	-0.008
	(0.009)	(0.007)	(0.006)	(0.005)	(0.005)
Number of Observations	3,090	6,092	9,249	11,292	13,206

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean Ministry of Education

Table 4.11: RD Local Randomisation Estimates for Age

Pre-Treatment Variable	Window Size				
	10	20	31	45	59
Age (Years)	-0.097**	-0.143***	-0.153***	-0.172***	-0.176***
	(0.042)	(0.030)	(0.025)	(0.023)	(0.022)
Number of Observations	3,090	6,092	9,249	11,292	13,206

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean Ministry of Education

4.3 Case 2: *Subsidio Familiar & Asignación Social*

This section analyses the cases of *Subsidio Familiar* (SUF) and *Asignación Social* (AS). The first part provides an overview of the two CCTs and their targeting strategy. Both CCTs use the Social Protection File (SPF) index to assess eligibility. The second subsection presents the data sources and summary statistics. The third part introduces a case of administrative sorting near the SPF threshold used by SUF. This subsection shows that administrative sorting induces abrupt changes in the SPF index density, threatening the RD design. The fourth part elaborates on a case of intermediate contamination near the SPF threshold used by AS.

4.3.1 *Programmes and Targeting Description*[51]

SUF Overview and Rollout

SUF is a monthly cash benefit that operates as a CCT in practice. SUF has two goals. First, the CCT seeks to increase households' present income and, secondly, it intends to promote human capital accumulation among children and adolescents no more than 18 years old.

The mother is the recipient. The monthly SUF-related income she receives varies by the number of entitlements within the household. In 2013, the average number of entitlements per household that received SUF was 2.60. 64.9% of the entitlements corresponded to children until 18 years old while mothers accounted for 34.7% of the share of entitlements. From July 2014 until June 2015 the monthly cash transfer per entitlement was $9,242 CLP ($14.6 USD at June 30th 2015). The government updates this amount annually.

SUF is currently the largest CCT in Chile, with more than 760,000 households and 2 million entitlements since 2010. In 2013, approximately 11.7% of Chileans were entitled to SUF. SUF's target population is households belonging to the poorest 40% of the national population. Low-income status was measured using the SPF index. Since September 2007, when the SPF was implemented, selection into SUF has been primarily decided by two

[51] The information in this section comes from various official reports (Comité de Expertos Ficha de Protección Social, 2010; Focus. Consultorías y Estudios, 2016; Universidad del Desarrollo, 2014).

factors: i) having a score equal to or lower than 11,734 points in the SPF index, and ii) self-selection - a household member must apply for SUF at the local government (municipality) office.

These rules hold if a person is claiming SUF for the first time or if they are renewing it after three years. To receive SUF the person must bring, at the time of application (or renewal), proof of school enrolment for those children aged between 7 and 18 years old within the household.

AS Overview and Rollout

AS was a pilot programme provided in 2011 and 2012 before the implementation of *Ingreso Ético Familiar* (IEF), currently the largest programme in Chile designed to reduce extreme poverty. AS had different goals. First, the scheme intended to increase households' income. Additionally, it intended to promote human capital accumulation among children. AS was launched in April 2011, providing cash transfers to 123,757 households. The programme then grew to 148,775 households in January 2012. In 2013, AS was replaced by IEF.

AS provided different types of cash transfers, which can be classified into unconditional and conditional cash transfers. The conditional cash transfers were provided as follows, per child: i) $5,000 CLP ($10.4 USD on March 31st 2011) per month for those aged no more than 6 years old attending regular health check-ups, and ii) $25,000 CLP per year ($51.9 USD) for those aged between 7 and 18 years old enrolled in primary or secondary education, plus $35,000 CLP ($72.6 USD) per year if the annual attendance rate reached 85%.

The target population of AS was households in extreme poverty. Every household that met the eligibility criteria on March 31st 2011 was classified as a "stock" household. This was operationalised to mean: i) having a score equal to or lower than 4,213 points in the SPF, and ii) being an active member of *Chile Solidario* (previously the largest programme in Chile designed to reduce extreme poverty), where this implied receiving the unconditional cash transfers of this programme (provided for five continuous years) or its psychosocial support component (which was provided only in the first two years following take-up).

AS took advantage of the infrastructure developed for *Chile Solidario* to deliver its cash transfers in addition. Hence, enrolment in AS was automatic and 98% of the beneficiaries received payments through the same channels as *Chile Solidario*. Compliance with the conditions of school attendance and enrolment was verified by the Chilean Ministry of Social Development (MSD) using administrative records.

The Social Protection File (SPF)

Until December 2015, the SPF was the largest targeting instrument in Chile. The instrument was developed by the MSD as a tool to assign social programmes among the population.

Any household could ask for an SPF assessment in their respective municipality. However, in practice, the wealthiest tended not to ask for it. The data collection was carried out through an interview conducted by municipality personnel that had a standardised procedure. The process started in 2006 and the first scores were provided to households in late 2007. As of January 2010, the dataset had 10,782,270 individuals, approximately 63.5% of Chile's population.

SPF scores estimate household income. Specifically, the formula estimates household members' income using variables correlated to their income. Therefore, the instrument is a form of a proxy means test. The score of the household derives in part from the sum of the predicted incomes of each member. Most of the variables that go into the formula are collected during the household interview. The general formula for the SPF score is as follows:

$$SPF\_Score_h = G\left(\frac{\sum_i^n\{[CGI_i*0.9+YD_i*0.1]+P_i*YP_i\}}{IN_h}\right),$$

where CGI_i is the proxy means test prediction of the (potential) labour income of individual i in household h.[52] YD_i is the self-reported income (mostly from labour) of individual i in household h. YP_i is the permanent income (public pensions being the most relevant) for

[52] Six different equations are used to estimate the CGI depending on the gender of individual i and whether he or she is: i) a paid employee, ii) self-employed or, iii) unemployed or economically inactive. This categorisation acknowledges the differences in (potential) income across groups within the country. Education plays a vital role in all equations. For example, more years of schooling and concluding levels (such as secondary or higher education) are positively correlated with CGI predictions. Additionally, other variables such as the type of education received, the age, and the district where the individual lives affect these predictions.

individual i in household h. IN_h is a needs index (or a form of equivalence scale) for household h. Finally, $G(\cdot)$ is a monotonic function that transforms household h income prediction into an SPF score.

Given its formula, in theory, the SPF index ranks households from the poorest to the richest in a similar way to a ranking of households by income per capita.[53] The SPF scale ranges from 2,072 points (the poorest households) to infinity (the richest) in theory. In practice, by 2014, only two households had an SPF score higher than 16,000 points.

4.3.2 *Data, Sample and Descriptive Statistics*

Data and Sample

Every administrative dataset I use in this section was provided by the Chilean Ministry of Social Development at my request. I link individuals in each dataset using the ID number provided by the Chilean State. For privacy purposes the ID numbers were changed by the MSD using an algorithm that is unknown to me but that enabled me to link the datasets.

The Social Protection File Dataset provide information on: i) SPF scores, which are crucial for the identification of both SUF and AS, ii) household structure, and iii) variables such as years of schooling, employment status, and self-reported income. The Social Protection File is available on a monthly basis from September 2007 until December 2015.

The Ministry of Education annual Performance Dataset contains information on enrolment, attendance, performance and end of year academic classification for all Chilean primary and secondary school students (except those in adult flexible and differential education). This dataset is available annually from the MSD. Additionally, I use the SUF and AS datasets to identify treated and non-treated adolescents for each CCT. These two datasets are available on a monthly basis. The *Chile Solidario* historic dataset is useful to identify eligible adolescents (active members of *Chile Solidario*) for AS.

[53] SPF scores do not account for all sources of income (for example many subsidies are not included). Additionally, the SPF formula uses a needs index not the number of household members n in the denominator. Despite these differences, for simplicity, I refer to SPF scores as a prediction of household income per capita.

My analysis focuses on adolescents aged between 13 and 17 years old (using December 31st as the reference day to calculate their age). I restrict the sample by age range because of two factors. First, in Chile, dropout rates are most problematic in secondary studies and so including children younger than 13 years old does not seem pertinent in the context of evaluating the impact of Chilean CCTs on school enrolment. Second, adolescents older than 17 years old were not eligible for SUF and AS.

Descriptive Statistics

The first three columns of Table 4.12 provide descriptive statistics relevant to SUF for two cohorts, corresponding to adolescents in 2013 and 2014. The last three columns of Table 4.10 present descriptive statistics for AS. The sample I use in this case is adolescents who were living in an active *Chile Solidario* household in March 2011.

Panel A shows that adolescents scoring no more than 11,734 in the SPF were more likely to be entitled to SUF. Overall, 41.2% of eligible adolescents received SUF. Conversely, less than 1% of adolescents scoring above this threshold received SUF. Panel A additionally shows AS high rates of compliance. 97.6% of adolescents who were active in *Chile Solidario* and had no more than 4,213 in the SPF of March 2011 received AS. In contrast, practically no *Chile Solidario* adolescents scoring above this threshold received AS.

Panel B presents variables that are used in the SPF formula. Adolescents scoring below 11,734 are more vulnerable than those scoring above this threshold. On average, the former group has a head of household with fewer years of schooling and a lower chance of working formally (defined as working and contributing to social security). Their households also have a lower income and are larger. *Chile Solidario* adolescents with SPF scores no higher than 4,213 live in households with a lower income that is more likely to be headed by a female compared to *Chile Solidario* adolescents with SPF scores higher than 4,213.

Concerning their demographic and academic information, Panels C and D show that adolescents scoring below the SUF threshold are also more disadvantaged in terms of academic features compared to adolescents with SPF scores higher than 11,734. Contrarily, *Chile Solidario* adolescents with an SPF score no higher than 4,213 are similar to their counterparts whose SPF score is higher than 4,213.

Table 4.12: Descriptive Statistics (Mean Values) by Conditional Cash Transfer With Its Respective SPF Threshold

Variables	Conditional Cash Transfer and Sample Utilised					
	SUF: All Adolescents			AS: Active *Chile Solidario* Adolescents		
	<= 11,734	>11,734	Total	<= 4,213	>4,213	Total
Panel A: SPF Score and CCT Recipient						
SPF Score	5,907.2	13,222.8	6,976.0	2,939.2	7,013.0	4,261
SUF Recipient	0.412	0.008	0.353	-	-	-
AS Recipient	-	-	-	0.976	0.000	0.659
Panel B: SPF Relevant Variables (HH: Head of Household)						
HH Years of Schooling	9.17	12.10	9.60	8.08	7.68	7.95
HH Working	0.741	0.838	0.755	0.682	0.670	0.678
HH Working Formally	0.351	0.730	0.406	0.118	0.261	0.164
Household Monthly Income ($CLP)	154,970	423,502	194,201	62,181	124,632	82,444
Household Size	4.30	4.00	4.26	4.66	4.78	4.70
Female HH	0.510	0.291	0.478	0.660	0.430	0.586
Panel C: Demographic Information						
Male	0.513	0.515	0.513	0.506	0.519	0.510
Age (Years)	14.93	14.95	14.93	14.91	14.96	14.93
Metropolitan Region	0.335	0.429	0.349	0.222	0.231	0.225
Panel D: Academic and School Information						
Enrolment Previous Year	0.913	0.954	0.919	0.899	0.877	0.892
Seventh or Eighth Grade	0.377	0.354	0.373	0.386	0.380	0.384
Ninth or Tenth Grade	0.424	0.446	0.427	0.409	0.411	0.409
Eleventh or Twelfth Grade	0.144	0.178	0.150	0.129	0.130	0.129
Attendance Previous Year (%)	90.20	91.74	90.43	91.00	91.24	91.08
Average Grade Previous Year	5.32	5.48	5.35	5.25	5.24	5.24
Progressed Previous Year	0.914	0.939	0.918	0.893	0.895	0.894
Number of Observations	1,627,331	278,419	1,905,750	60,934	29,266	90,200

Note: grades, attendance, average grade and progression measured only among those who were enrolled in the previous year

Source: own calculations using administrative datasets, Chilean Ministry of Social Development and Ministry of Education

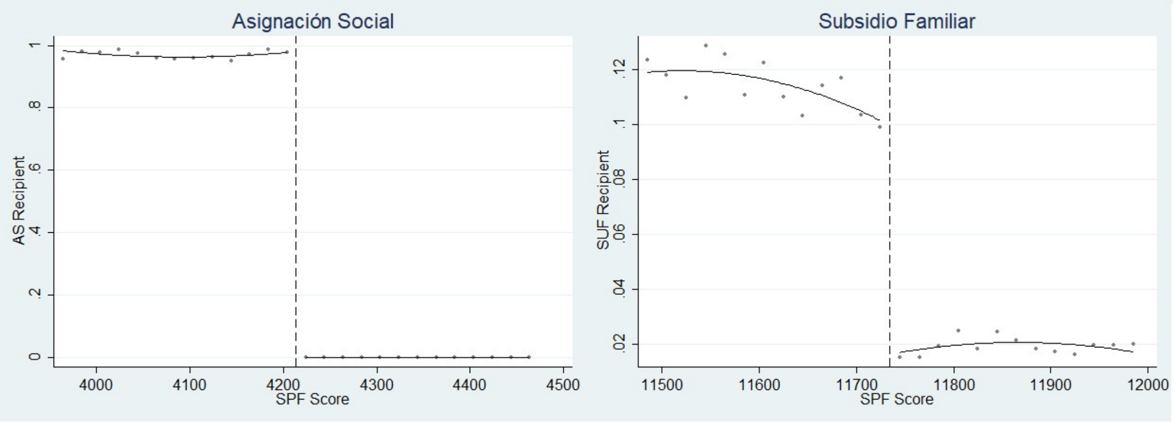
4.3.3 *Administrative Sorting Threatens the RD Design*

Can SPF Scores be used as a Running Variable?

A regression discontinuity design seems to be a suitable method to estimate the causal effect of SUF and AS on school enrolment of adolescents. The high degree of compliance in AS take-up suggests that a sharp RD design could be implemented for this CCT. Conversely, SUF needs to be assessed through a fuzzy RD design, as many factors, beyond having an SPF score of 11,734, play a role in determining which adolescents access the cash transfer.

Figure 4.6 shows the relationship between the treatment and SPF scores for AS and SUF. Figure 4.6 supports the idea of using an RD design because for each CCT there exists an SPF threshold (the dashed line) that affects treatment propensity. For AS the change in treatment status is almost deterministic after crossing the 4,213 SPF threshold. For SUF only a change in the probability of treatment is observed for adolescents with an SPF score no higher than 11,734.

Figure 4.6: *Asignación Social* and *Subsidio Familiar* Recipient by SPF Scores[54]



Source: own calculations using administrative datasets, Ministry of Social Development

SPF scores act as a potential running variable. Observing a smooth distribution in the running variable in the neighbourhood of the threshold is essential to believe that treatment assignment is as good as random. Conversely, any discontinuity in the density of the running variable close to the threshold is generally interpreted as a sign of manipulation. Whenever

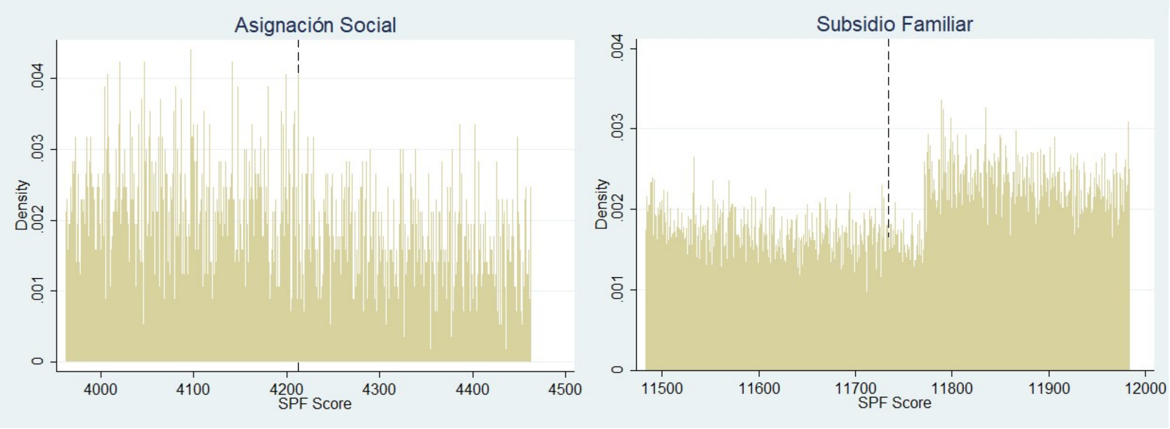
[54] The dots correspond to the proportion of recipients for each CCT by 20-point SPF bins. I derive the continuous lines from a quadratic regression. The parameters of the regressions vary at each side of the cutoffs.

there is manipulation the plausibility of the continuity assumption holding weakens. Therefore, it becomes essential to assess whether any discontinuities emerge in the distribution of SPF scores.

Figure 4.7 presents the density of SPF scores in a small neighbourhood close to the AS and SUF thresholds. The left panel of the figure presents the case of AS: this is the distribution of SPF scores for *Chile Solidario* active adolescents around the 4,213 SPF score (the dashed line). The density is uneven, but no clear discontinuity is observed when crossing this threshold.

The right-hand panel of Figure 4.7 presents the case of SUF: this is the distribution of SPF scores for all adolescents around the 11,734 SPF threshold (the dashed line). Unlike the case of AS, a clear discontinuity can be observed in the density of SPF scores to the right side of the 11,734 SPF threshold, though this is not observed immediately after crossing the threshold.

Figure 4.7: SPF Scores Distribution for *Asignación Social* and *Subsidio Familiar*

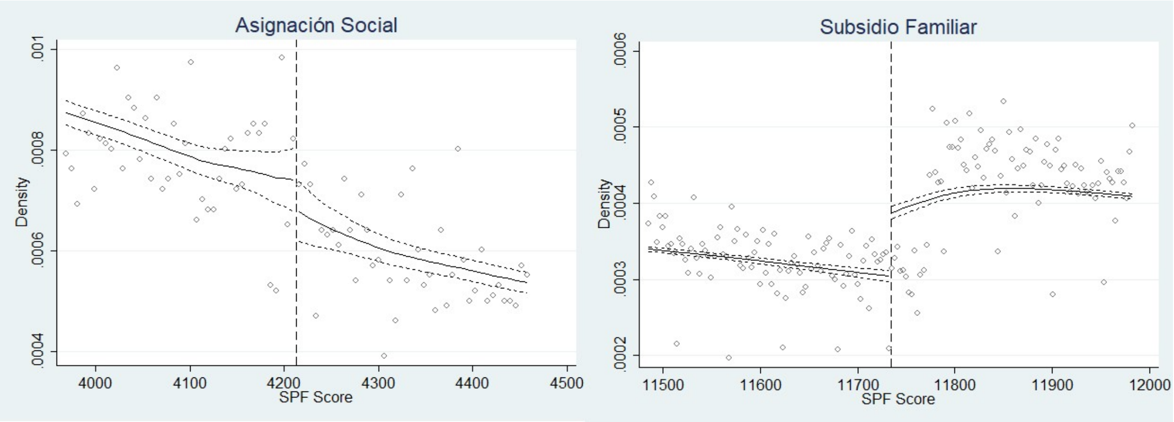


Source: own calculations using administrative datasets, Ministry of Social Development

A McCrary test (2008) formally assesses whether there is a discontinuity in the density of a running variable at a threshold. The test produces an estimation of the distribution of the running variable at each side of the threshold. If the estimate of the discontinuity in the density is not statistically significant there is no evidence to claim that such a discontinuity exists. Graphically, this is mostly the case when the confidence intervals of the estimates overlap at the threshold.

Figure 4.8 presents the results of the McCrary test for SUF and AS. The figure shows the estimated density of SPF scores (in the wider central line) with its 95% confidence interval (the dashed side lines) at each side of the threshold. In the right-hand panel, no overlap between the estimated density of SPF scores can be observed at the 11,734 SPF threshold. In contrast, the left panel of Figure 4.8 shows some degree of overlap in the confidence intervals of the estimated density of SPF scores at the 4,213 SPF threshold.

Figure 4.8: McCrary Test for *Asignación Social* and *Subsidio Familiar*



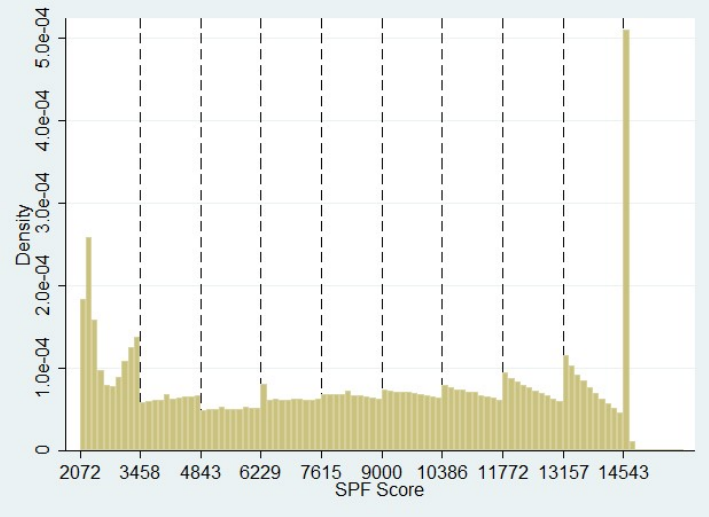
Source: own calculations using administrative datasets, Ministry of Social Development

The McCrary test result raises the question of the suitability of using SPF scores as a running variable to evaluate SUF. Conversely, the test does not reject the use of SPF scores as a running variable to evaluate AS. At first glance, the discontinuity in the density of SPF scores around 11,734 suggests that the variation in SUF eligibility is not as good as random close to this threshold. In this case, the continuity assumption is less likely to hold and RD estimates for SUF are more likely to be biased. The next subsection explains the underlying cause of these results.

Understanding SPF Scores Distribution

The discontinuity in the distribution of SPF scores shown previously is not an isolated case. The SPF scores distribution shows clear discontinuities in at least four parts of its range. Moreover, I distinguish a sequence in these discontinuities. This can be seen in Figure 4.9.

Figure 4.9: Social Protection File Scores Distribution



Source: own calculations using administrative datasets, Ministry of Social Development

The dashed lines divide the SPF scores distribution into ten sections. Each section has a size of 1,385 or 1,386 SPF points. The first discontinuity in the distribution is observed at the SPF score of 3,458 (where the first dashed line is), 1,386 points away from the lowest SPF score of 2,072. On the right-hand side of the figure, larger discontinuities can be observed at the SPF scores of 11,772, 13,157 and 14,543 (the seventh, eighth and ninth lines, respectively).

Figure 4.9 suggests the existence of a pattern in the SPF scores. Every 1,385 or 1,386 SPF points there is a discontinuity in the distribution. Public policies in Chile used none of the points where these discontinuities are observed. Hence, these discontinuities are unlikely to be explained by individuals' manipulation of SPF scores. That argument seems more suitable for explaining the accumulation of observations at the very left side of the distribution.

A better explanation is that the discontinuities are administratively produced. The function $G(\cdot)$, which transforms households' income per capita predictions into SPF scores is responsible for these discontinuities. This hypothesis is supported by the feasibility of obtaining from the Chilean income per capita distribution a density with a similar shape to the SPF in a short number of steps. This sequence of steps, or function $g(\cdot)$, is described as follows:

1. Estimate the independent income per capita for each household (using CASEN).[55] [56]
2. Obtain the nine percentiles of income that divide the sample into ten decile groups.
    - The ten decile groups have different sizes in terms of the income range. For example, income per capita for the first group ranges from $0 CLP to $43,960 CLP, while for the ninth group it ranges from $329,507 CLP to $567,581 CLP.
3. Divide each decile group into ten additional subgroups with an equal income range.
    - Each subgroup has the same income band within each decile group (for example each subgroup in the first decile has an income band of $4,396 CLP).
    - Income ranges differ for two subgroups from different income decile groups. Thus, the income band for the tenth subgroup is different from the eleventh.
4. Assign each household in CASEN to 1 of the 100 independent income per capita subgroups by observing their independent income per capita.

I present the resulting distribution in the left panel of Figure 4.10.[57] The dashed lines in this panel separate income decile groups. This is compared with the right-hand panel of Figure 4.10, which shows the density of SPF scores when 100 bins of 138.6 points are used to construct the histogram. The dashed lines in this panel separate SPF scores by 1,386 points.

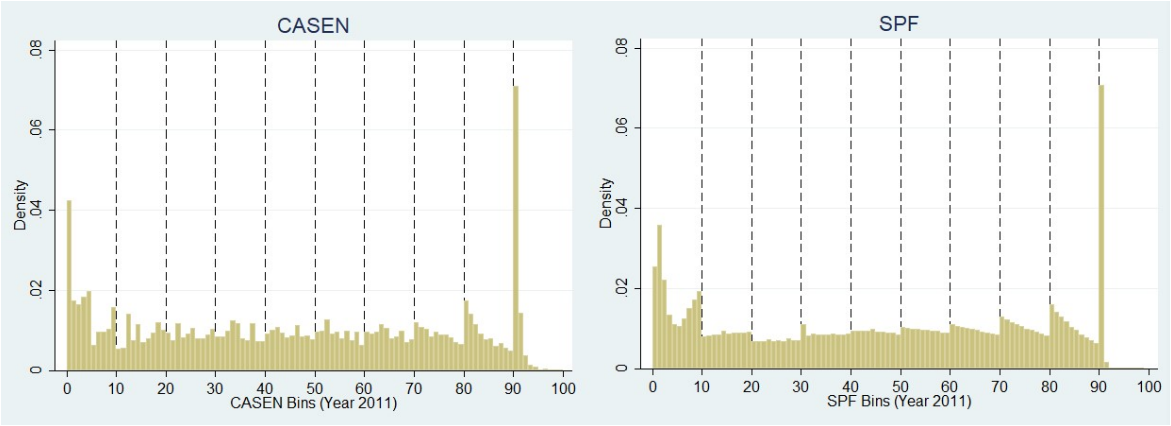
There are many similarities between these two distributions, including: i) an increasing density from the sixth to the tenth bin, ii) a sharp decline at the eleventh bin, iii) a flat density thereafter (albeit with more variance in CASEN), iv) a moderate discontinuity at the 71st bin, v) a steady decline in the density until the 80th bin, vi) a repetition of the pattern noticed in points iv) and v) between the 81st to the 90th bin, and finally, vii) a very sharp increase in the 91st bin that leads to the highest concentration of observations in both distributions.

[55] The *Encuesta de Caracterización Socioeconómica de Hogares* (CASEN) is the cornerstone of Chilean welfare measurement. In Chile, official statistics about poverty and inequality are based on income, not consumption. Therefore, CASEN collects detailed information about every source of income within the household. Since 1990, CASEN has been held every two or three years. The survey is representative at a national and a regional level. In CASEN 2011, 86,836 households (approximately 1.75% of the population) were interviewed. CASEN datasets are available from http://observatorio.ministeriodesarrollosocial.gob.cl/

[56] Independent income refers to all monthly payments received related to labour and property of assets.

[57] After using $g(\cdot)$ I randomly assign 7% of observations from the second to the tenth decile groups to one of the first five subgroups to account for the higher density in the left side of the distribution of SPF scores.

Figure 4.10: One Hundred Bins of Income per Capita in CASEN and the SPF[58]



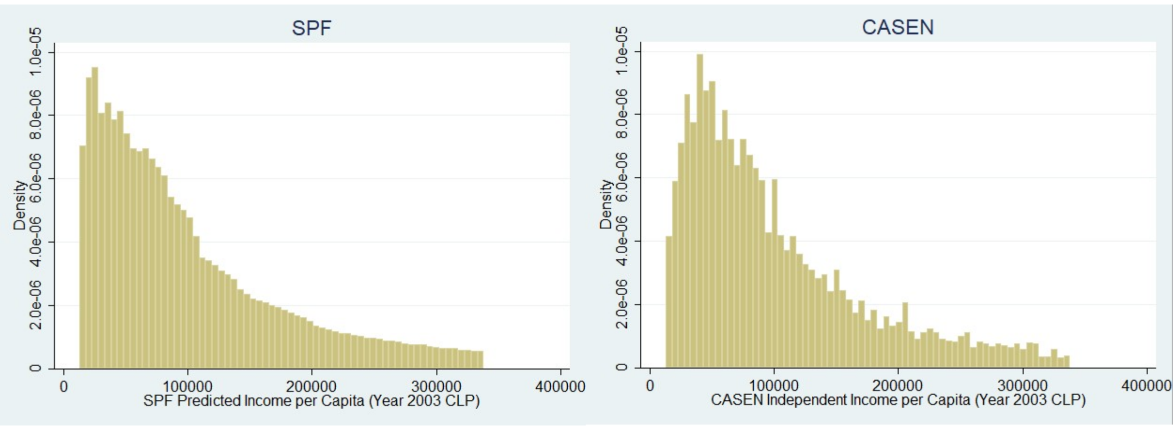
Source: public and administrative datasets, Ministry of Social Development

The monotonic function $G(\cdot)$ that produces SPF scores from income predictions seems to affect non-randomly the position of units near some SPF thresholds. Thus, administrative sorting in the running variable is the primary explanation for the McCrary test result in the case of SUF. The 11,734 SPF threshold used by SUF for targeting is too close to 11,772, where one of the discontinuities in the density of SPF scores explained by $G(\cdot)$ emerges. The AS threshold of 4,213 SPF points is not close enough to 3,458 and 4,843 (scores where discontinuities caused by $G(\cdot)$ are observed) to lead to a similar McCrary test outcome.

As SPF scores are derived from a prediction of income per capita and a monotonic function $G(\cdot)$, using the inverse function of $G(\cdot)$ on SPF scores will provide the original income prediction. Accordingly, if the national household income per capita distribution is continuous then the SPF predicted income per capita distribution is likely to be continuous. I present the result of inverting the SPF scores, using similar ideas as in function $g(\cdot)$, in the left panel of Figure 4.11. The density of SPF predicted income per capita does not show the discontinuities of the SPF distribution. This is the case in all parts shown in this distribution except on its tails, which are not shown. Moreover, the density is similar to CASEN's (which is shown in the right-hand panel of Figure 4.11).

[58] SPF scores do not precisely predict independent income per capita, but a different measure of income (which includes some public pensions and subsidies and excludes income related to the property of assets) divided by a needs index. Many of these factors cannot be easily accounted for in CASEN. For this reason, I select independent income per capita as the standard for comparison.

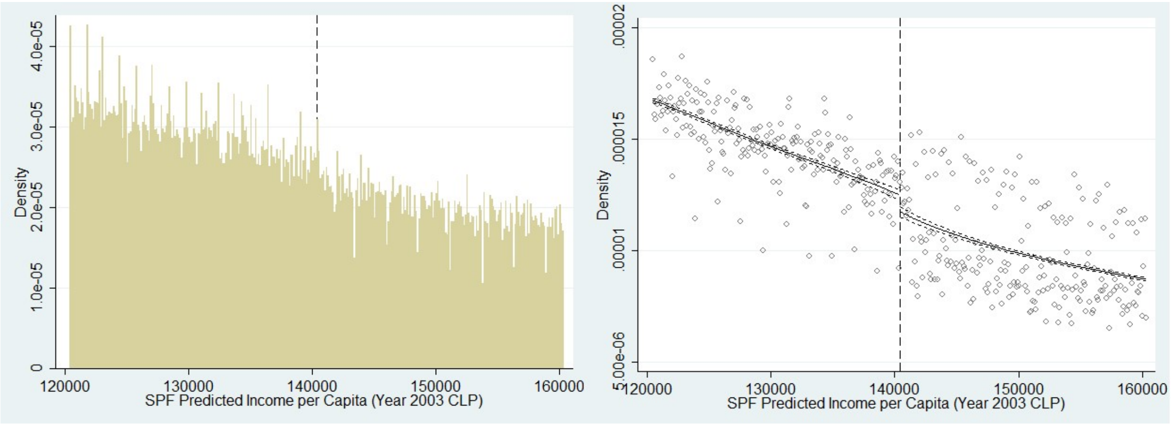
Figure 4.11: Income per Capita Distribution (SPF Prediction and CASEN)



Source: public and administrative datasets, Ministry of Social Development

SPF predicted income per capita is a more suitable running variable for an RD design for SUF than SPF scores because this running variable is smoother around the threshold used by SUF. This can be better appreciated from Figure 4.12. The left panel of the figure shows the distribution of SPF predicted income per capita around the equivalent (to 11,734 SPF points) threshold. The density is smooth in the relevant neighbourhood. The discontinuity in the estimated density is narrower relative to the one I presented in Figure 4.8. However, the new running variable still does not pass the McCrary test (in the right-hand panel of Figure 4.12).

Figure 4.12: SPF Predicted Income per Capita Distribution and McCrary Test



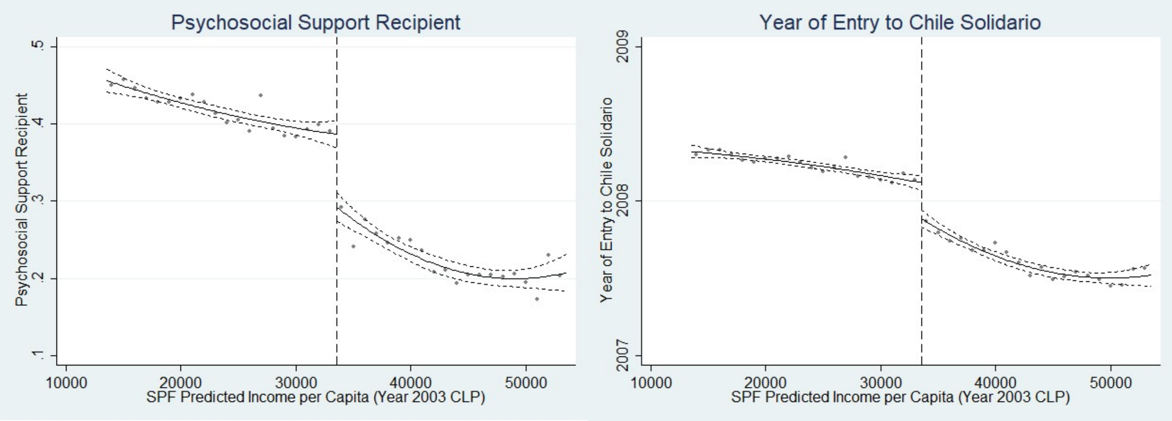
Source: own calculations using administrative datasets, Ministry of Social Development

All these findings confirm that administrative sorting, not manipulation, is the primary driver of the discontinuities in the distribution of SPF scores. This situation highlights the relevance of fully understanding the running variable generating process when applying an RD design.

4.3.4 *Intermediate Contamination Invalidates the RD Design*

In Figures 4.7 and 4.8 I show that the neighbourhood of scores close to the 4,213 SPF threshold is free of discontinuities. For this reason, using an RD design for AS seems more promising relative to the case of SUF, due to SUF's threshold being affected by administrative sorting. However, an RD design for AS presents additional complications caused by intermediate contamination. Active *Chile Solidario* adolescents just below and above the 4,213 SPF threshold were not comparable when AS was launched. This can be observed in Figure 4.13.

Figure 4.13: *Chile Solidario* Variables by SPF Predicted Income per Capita ($CLP)[59]



Source: own calculations using administrative datasets, Ministry of Social Development

Figure 4.13 shows the SPF predicted income per capita on the horizontal axis. I use this variable instead of SPF scores due to its better properties as a running variable. In the vertical axis of Figure 4.13 I present two pre-treatment variables related to the *Chile Solidario* programme. Active *Chile Solidario* adolescents just below the SPF predicted income per capita threshold, equivalent to 4,213 SPF points (the vertical lines), were more likely to be receiving the psychosocial component of *Chile Solidario* and had a later date of entry into this programme. A discontinuity in the distribution of these two intermediate variables is visible at the threshold.

I show estimates of the discontinuities in the distribution of both variables in Table 4.13. Given that income and *Chile Solidario* participation are expected to be correlated, I implement a

[59] The dots correspond to the proportion of psychosocial support recipients and year of entry to *Chile Solidario* by $1000 Chilean Pesos non-overlapping bins of SPF predicted income per capita. I derive the continuous lines from a quadratic regression. The parameters of the regressions vary at each side of the threshold. The dashed lines surrounding each of the continuous lines represent the 95% confidence interval of each polynomial fit.

continuity-based RD framework for the falsification test. I use the following local regression:

$$X_{2i} = \alpha + \beta I_{2i} + \gamma Z_{2i} + \theta Z_{2i} I_{2i} + \omega {Z_{2i}}^2 + \delta {Z_{2i}}^2 I_{2i} + \varepsilon_{2i}$$

where X_{2i} is a pre-treatment variable for adolescent i, Z_{2i} is the difference between the SPF predicted income per capita equivalent to 4,213 points and the SPF predicted income per capita for adolescent *i*. I_{2i} is an indicator function, which takes the value of one if $Z_{2i} \geq 0$ and zero otherwise. ε_{2i} represents the difference between the observed value and the model prediction.

The continuity-based RD estimates (β) are statistically significant and robust to bandwidth selection. Adolescents just below the threshold were approximately ten percentage points more likely to have been receiving the *Chile Solidario* (CS) psychosocial support when AS was launched. Additionally, those adolescents who barely qualified for AS were around a quarter of a year younger in *Chile Solidario* relative to their peers who just missed out on AS.

Table 4.13: Continuity-Based RD Estimates for *Chile Solidario* Variables

Pre-Treatment Variables	Bandwidth ($CLP)			
	5,000	10,000	15,000	20,000
Psychosocial Support	0.094***	0.109***	0.101***	0.094***
Recipient (March 2011)	(0.027)	(0.019)	(0.016)	(0.014)
Year of Entry	0.252***	0.271***	0.239***	0.232***
to *Chile Solidario*	(0.075)	(0.053)	(0.043)	(0.038)
Number of Observations	11,318	23,830	37,834	50,677

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Ministry of Social Development

Because one requirement for receiving AS was being an active member of CS, the difference across groups in terms of the CS-related variables could come as a surprise. However, the fact that from late 2007 treatment into CS was mainly decided by having a score equal to or lower than 4,213 in the SPF (same criterion as AS) helps to explain these results. Older CS cohorts, which entered this five-year programme before late 2007, could have had, in March 2011, any SPF score without their CS condition being affected. Conversely, the cohorts that entered the programme after late 2007 could only have entered CS if they had less than 4,213 in the SPF. Thus, in March 2011 those CS recipients scoring just above 4,213 in the SPF tended to be on average older beneficiaries of CS relative to those scoring just below this threshold.

The school enrolment decisions of adolescents are likely to be affected by differential exposure to CS components. Therefore, the differences I observe for these intermediate variables at the threshold weaken the plausibility of the continuity assumption holding in an RD design. Thus, AS impact estimates can be biased. This is explained by intermediate contamination. The running variable, which was determined long before AS was implemented, differentially affected *Chile Solidario* active adolescents near the 4,213 SPF threshold.

4.4 Conclusion

This paper focuses on two threats to internal validity in RD designs that have received little attention in the methodological RD literature. I analyse administrative sorting and intermediate contamination in the context of the evaluation of three CCTs using RD designs. Table 4.14 outlines the CCTs, running variables and thresholds I use and the threats to validity I observe.

Table 4.14: Summary of CCTs, Running Variables, Thresholds and Threats to Validity

CCT	Running Variable	Threshold	Threat to Internal Validity
BARE	IVSE	84.625 or 84.150	Administrative Sorting
	Day of Birth	May 30th	Intermediate Contamination
SUF	SPF	11,734	Administrative Sorting
AS		4,213	Intermediate Contamination

Administrative sorting differs from the threat of manipulation, which is characterised by individuals' deliberate action for their benefit. Administrative sorting is the result of administrative procedures that are beyond the control and knowledge of individuals. These procedures affect non-randomly the position of individuals in the running variable near the threshold. As a result, the continuity assumption becomes less plausible.

I present two cases of administrative sorting. The first is caused by the categorisation of discrete variables in the index used to target BARE. The index formula generates clusters of students rather than a smooth distribution of scores, so students just above and below the threshold differ in terms of key features. I show that minor adjustments in the formula would have prevented these problems and made the index suitable to be used as a running variable in an RD design.

In the second case, administrative sorting is caused by the monotonic function used to transform the predicted values of a proxy means test into ordinal scores. The function generates noticeable discontinuities in the distribution of index' scores. For one of the CCTs evaluated, SUF, a discontinuity is close to the threshold used for targeting. As a result, the density of the index fails to pass the McCrary test. The paper shows that the predicted income of the targeting instrument is a more suitable running variable for an RD design.

How administrative sorting unfolds differs in each of my examples. In the first case, BARE, one type of adolescents ends up at one side of the threshold while a different type of adolescents ends up at the other side of the threshold. A better design of the index would have caused both types of adolescents to have equal probabilities of being at one side or the other of the threshold. In the second case, SUF, adolescents at one side of the threshold have been artificially compressed inducing a discontinuity in the density of the running variable. In this case, what is affected is the distance of individuals from the threshold rather than the side where they finish.

In these two cases, lack of manipulation of the running variable does not translate automatically into variation in treatment assignment being as good as random near the threshold. This finding contradicts one of the most highlighted points in Lee and Lemieux's (2010) influential paper on RD designs. Here administrative sorting undermines the RD design, not manipulation. Accordingly, lack of manipulation is a necessary condition but not sufficient for valid RD designs. My paper shows that a McCrary test can fail for reasons other than manipulation of the running variable. This finding broadens our understanding of the interpretation of the test.

If administrative sorting is not properly recognised as driving a McCrary test result it could be mistaken as manipulation in contexts where indexes such as proxy means tests are used as running variables. In Chile, a lack of acknowledgement of this kind may lead readers to discard viable research designs when using the SPF (Centro de Microdatos, 2012). In Ecuador, administrative sorting has not been considered as a potential cause that could explain why more individuals are found above the threshold used by the CCT *Bono de Desarrollo Humano* (Ponce & Bedi, 2010). The authors only explain that this difference could not be attributed to manipulation of the index.

My findings highlight the importance of fully understanding the data generation process. If administrative rules, not individual manipulation, explain the shape of the running variable density near the threshold then useful variation for identification may still exist. For example, future research on the effect of BARE on school enrolment could exploit the categorisation of attendance in the IVSE formula if adolescents were unaware of the effect of this variable for selection into BARE. Adolescents who differ by a few decimal places in their attendance have a different likelihood of being eligible for BARE. In the case of SUF, future research could focus on fully reconstructing the SPF predicted income. If this variable passes the McCrary test, it may be useful as a running variable to identify the impact of SUF on school enrolment.

I discuss two examples of intermediate contamination in the paper. In each case, the running variable and threshold affected other factors, potentially associated with the outcomes of the treatment, before the running variable and the threshold were used to assign the treatment of interest. Timing matters here. Units near the threshold were no longer similar when the treatment of interest occurred, hence affecting the plausibility of the continuity assumption.

The first case of intermediate contamination relates to the BARE assessment that uses day of birth. The paper shows that adolescents born before and after May 30th have similar pre-treatment characteristics. For example, it is not possible to observe statistically significant differences in multiple variables such as gender, region and rate of progression. However, these adolescents are statistically significantly different in their age, a result that is likely to be explained by the discontinuous effect of date of birth on age of entry to primary school in Chile.

Settings in which a substantive time lag exists between the realisation of the running variable and its use for the treatment of interest are more prone to intermediate contamination. Thus, RD designs that use age at a later stage in life as a running variable deserve careful analysis. Age depends on the date of birth, which affects among other factors, the timing of entry into school and, as a consequence, future outcomes in multiple social dimensions (Crawford et al., 2010; Elder, 2010; Fredriksson & Ockert, 2005). In the first years of life, date of birth could also affect exposure to vaccination campaigns (Helleringer, Asuming, & Abdelwahab, 2016) and access to pre-school education and care (Blanden et al., 2017). Multiple RD designs that use age at a later stage in life can be found in the causal inference literature, such as evaluations of the effect of the minimum legal drinking age (Carpenter & Dobkin, 2009; Yörük & Yörük, 2011) or of the impact of labour market policies on young adults (Cockx & Dejemeppe, 2012; Dickens, Riley, & Wilkinson, 2014; Lemieux & Milligan, 2008).

The second case of intermediate contamination is associated with the use of a specific threshold of a proxy means test by another intervention in the past. The threshold used for eligibility for one CCT had previously affected the timing of entry into another social programme. Consequently, individuals were no longer comparable when the treatment of interest, AS, took place. RD practitioners should take additional precautions in settings where their running variables are expected, or have been used, to serve multiple purposes. For example, Barrientos and Villa (2015) use a running variable and threshold previously used by another policy in Colombia (Barrera-Osorio et al., 2007) to assess the impact of the CCT *Familias en Acción.*

Intermediate contamination can be classified within the concern, "other changes at the same threshold of the running variable" (Imbens & Lemieux, 2008) (p.631). However, intermediate contamination is more complex and subtler than other changes that happen at the same time at the threshold, such as discounts and retirement for individuals who have just turned 65 years old. If intermediate contamination occurred, then balance is likely to be observed on baseline characteristics but not on intermediate variables. RD designs that rely on running variables determined long before treatment can strengthen their validity by providing tailored evidence against intermediate contamination. In this context, my finding highlights the relevance of scrutinising beyond baseline characteristics and placebo outcomes in RD falsification tests.

A complete set of robustness checks in RD designs should include: i) a McCrary test for manipulation and administrative sorting, and ii) intent to treat estimates for baseline characteristics, intermediate variables and placebo outcomes to assess the quality of randomisation and any signs of intermediate contamination and simultaneous interventions.

Many of the threats to validity in RD designs discussed in this paper are closely related to the use of proxy means tests as running variables. At least 20 developing countries in Latin America, Africa and Asia such as Peru, Nicaragua, Bangladesh, Cambodia, Cameroon and Rwanda have used this mechanism for targeting (Australian Aid, 2011; Brown et al., 2016; Coady et al., 2004). Hence, this paper could also prove useful for public officers who are designing new social policies using PMTs that are expected to be evaluated with RD designs.

The recent breakthrough of RD designs in causal inference has been a major contribution. This is particularly true where experimental approaches are less feasible. In this spirit, this paper contributes clarifications regarding RD applications. It does so by highlighting administrative sorting and intermediate contamination, which are relatively overlooked threats to validity.

Chapter 5 Conclusion

The development of quality administrative records has expanded the possibilities for conducting cost-effective research outside developed countries. Utilising large and rich administrative datasets from Chile, in this thesis I answer three questions located at the intersection of conditional cash transfers (CCTs), targeting mechanisms of CCTs, especially proxy means tests (PMTs), and educational outcomes in primary and secondary school.

Each empirical chapter or paper addresses a different type of inquiry. Chapter 2 provides a targeting evaluation of a CCT. In other words, this chapter investigates the extent to which the target groups of CCTs (would) receive them in practice. Specifically, the second chapter analyses the effectiveness of a PMT and a predictive model of school dropout to find two target groups, poor students and future school dropouts. Chapter 3 focuses on measuring the causal effects of a cash transfer on subsequent attendance and academic performance. I use a regression discontinuity (RD) design for this purpose given that the cash for grades programme was targeted using different indexes. Hence, this chapter offers an impact assessment among subgroups of students who received a cash transfer. Chapter 4 elaborates on threats to internal validity in RD designs that have been overlooked in the context of three evaluations of CCTs, which use PMTs or similar indexes, on school enrolment.

Together the papers advance current knowledge both for social policy and methodology. I summarise the main findings and contributions of the papers in the first section of this chapter. Its second part discusses the thesis' most important implications for the improved design and evaluation of CCTs, and beyond CCTs. The third section of this chapter elaborates on further projections of my research and some limitations. The final section concludes the thesis.

5.1 Main Findings and Contributions

Chapter 2 addresses a noticeable gap in the targeting literature of CCTs and provides novel contributions to the social policy targeting field. I argue that the poor are not the exclusive target group of CCTs. Students at risk of leaving primary or secondary education are an additional target group of CCTs. However, targeting assessments of CCTs have primarily focused on the income dimension, with the notable exception of Azevedo and Robles (2013).

I compare the targeting effectiveness of a PMT to find poor students or future school dropouts with other simulated mechanisms that use the outputs of a predictive model of school dropout.

I build the predictive model of school dropout using machine learning algorithms, providing one of the first applications of this tool to predict school dropout outside a developed country. I compare the algorithms using receiver operating characteristic curves. My best algorithm produces results that are better than the ones obtained in Guatemala and Honduras (Adelman et al., 2017), in the same region or above as 107 out of the 110 dropout flags analysed by Bowers et al. (2013) and in line with the best models tested in the United States (Knowles, 2015; Sorensen, 2018).

The targeting assessment shows that a trade-off exists between using a PMT relative to the predictive model of school dropout. Using the PMT for targeting, instead of the predictive model, is more income progressive as poor undercoverage and non-poor leakage are reduced. However, future dropout undercoverage and non-dropout leakage increase. The paper shows that using the outputs of the predictive model in conjunction with the PMT increases targeting effectiveness by identifying more students who are either poor or future dropouts. This joint targeting approach increases effectiveness in different scenarios except when the social valuation of the two target groups differs to a large extent. In these cases, the most likely optimal approach is to use only the mechanism designed to find the target group that is valued the most.

While Chapter 2 aims to assess whether two target groups of a CCT were (would be) reached, Chapter 3 analyses the impact of a cash transfer. The paper does not focus on the most common design of CCTs in developing countries, in which there are conditions on school enrolment and attendance. Instead it analyses whether it is convenient to reward students according to their academic performance. Specifically, Chapter 3 estimates the impact of a cash for grades programme, *Bono por Logro Escolar* (BLE) in 2013, on future attendance and average grade. The conditional cash transfer was targeted using two indexes from 2012, a PMT and academic performance. I implement a sharp RD design along each running variable. I show that students marginally at each side of the two thresholds used for targeting only differ in terms of receiving the BLE in 2013.

The main causal estimates for the outcomes are not statistically significantly different from zero. Additionally, the analysis by subgroups does not consistently show estimates that are statistically significantly different from zero. If anything, the BLE local average effects in 2013 are modest and smaller than for previous interventions of this kind in developing countries, where effects of near 0.2 of a standard deviation on test scores have been found (Behrman et al., 2015; Kremer et al., 2009).

These findings are not necessarily surprising given that the evidence of the effectiveness of cash for grades schemes is mixed. In developed countries the literature has shown negligible effects (Fryer, 2011; Riccio et al., 2013) and statistically significant impacts only among specific topics or subgroups (Angrist & Lavy, 2009; Bettinger, 2012).

Like Chapter 3, Chapter 4 is written in the context of impact assessments of CCTs using RD designs supported by proxy means tests or similar indexes. However, the purpose of this paper is not to identify the causal effects of CCTs *per se*. Rather, Chapter 4 highlights atypical threats to validity in RD designs. The paper contributes to the causal inference literature by strengthening our theoretical understanding of this popular and rapidly developing method.

The main conceptual contributions of Chapter 4 are threefold: i) administrative sorting and intermediate contamination represent overlooked threats to internal validity in RD designs, ii) lack of individuals' manipulation (of the index) does not translate automatically into variation in treatment assignment being as good as random, and iii) distinguishing among pre-treatment variables (between baseline and intermediate) in RD falsification tests is beneficial for assessing whether intermediate contamination is affecting the internal validity of an RD design.

The second finding of Chapter 4, related to the manipulation of the index near the threshold used, questions a conclusion from Lee and Lemieux's (2010) influential paper on RD designs. The authors claim that, if individuals are unable to manipulate the index precisely, variation in treatment assignment near the threshold is as good as random. However, Chapter 4 shows twice that variation in treatment assignment is not as good as random even in the absence of individuals' manipulation. Additionally, Chapter 3 offers an additional equivalent case (partly introduced in Appendix H). Three times in my research, administrative procedures outside of individuals' control or knowledge, affect non-randomly the position of units near the threshold. As a result, units at each side of the respective thresholds are not comparable. Therefore, administrative sorting decreases the plausibility of the continuity assumption of RD designs.

The third finding of Chapter 4 is associated with the concept of time. Intermediate contamination is more likely to originate if a substantive time lag exists between the realisation of the index and the actual use of the index to assign the treatment of interest. Intermediate contamination relates to the impact on key variables of using the index and threshold within the referred time frame, making units near the threshold no longer comparable. Unlike the case of administrative sorting, which represents the most novel contribution of the paper, the concern about other changes happening at the same threshold of the index has been pointed out in the methodological RD literature (Imbens & Lemieux, 2008). However, in the context of a running variable created well in advance and that is used to assign the treatment of interest, the guidelines in RD falsification tests can be too broad and may not identify cases of intermediate contamination specifically.

5.2 Policy and Methodological Implications

The findings of the thesis have several implications for advancements in the design of CCTs. For example, in Chapter 2, I show that the targeting of CCTs can be improved when other dimensions beyond income, related to human capital accumulation, are incorporated in the targeting design. Given that CCTs have multiple target groups, a problem of misidentification exists if only one of them is considered for targeting. Households with students who are likely to drop out of school but who are not poor would most likely not receive a CCT. Public officials that value equally providing a CCT to a student who is poor or who is likely to drop out of school may find opportunities for increased targeting effectiveness by modifying the allocation rules of CCTs from using only a PMT to using a predictive model of school dropout with a PMT.

This conclusion does not necessarily hold if public officials mostly prioritise finding the poor over future dropouts. In this case maintaining the status quo, which is targeting CCTs on the basis of income, is appropriate. Alternatively, policy designers should evaluate the cost-effectiveness of adopting new targeting mechanisms for CCTs. A first step in this direction would be to: i) estimate the costs of developing and implementing a new targeting mechanism, ii) estimate the gains in targeting effectiveness, and iii) compare these with the default scenario.

Section 1.3 explains that CCTs might have gone too far in Latin America. The number of CCT beneficiaries overtook the number of poor on the continent in 2006 (Stampini & Tornarolli, 2012). A country in this position has increased difficulties in finding new CCT recipients who are poor. However, that CCTs have gone too far in the income dimension does not necessarily imply that CCTs have gone equally far in every important dimension. Instead of reducing the scope of CCTs, a potential policy alternative would be to redirect part of these cash transfers from non-poor households to households that underinvest in the human capital of their children. The findings of Chapter 2 invite the inclusion of students that are most likely to drop out of school in the targeting design of CCTs.

Taking CCTs away from multiple households, even if these are above the poverty line, may prove too complicated politically. Instead, another alternative to incorporate students who are likely to drop out of school in the targeting design of CCTs would be to modify only the amount of the cash transfer for current CCT recipients. For example, giving an increased amount for each child who has a high risk of dropping out of school, as informed by the predictive model.

If CCTs have gone too far in Latin America another policy alternative would be to reduce their coverage and invest the resources that are freed up in employment subsidies or cash for grades programmes. This path should improve the targeting accuracy of current CCTs in respect to income measures if the households abandoning the CCTs are mostly non-poor. Additionally, in theory, the new programmes should promote increased levels of employment among the unemployed and of students' effort and academic performance, respectively. These outcomes are not generally considered in the design of CCTs typically observed in developing countries.

However, Chapter 3 shows no statistically significant results in the assessment of the BLE. This highlights the need to understand why this cash for grades evaluation does not show a major impact on attendance and academic performance. For example, if children's unawareness of the cash transfer is a relevant factor then actions to increase this awareness should be pursued. Conversely, if students are aware of the benefits of the cash transfer but for them it is too difficult to improve their average grade then subject-specific goals could be considered.

Another potential explanation for not finding statistically significant results in Chapter 3 is that the size of the transfer is too small to affect behaviour. If this is true for all the population receiving the BLE then increasing this amount should be considered. However, it is also possible that the size of the cash transfer is appropriate for the poorest students and not enough for the non-poor. The RD design only allows for capturing the impact of the BLE for students near the thresholds used in targeting. These thresholds are the 30% with the lowest income in the population (measured by a PMT) and the 30% with the highest academic achievement in their cohorts. Therefore, for the BLE it is not possible to estimate the impact for a student at the lower end of the income distribution who is also at the median of academic performance within the cohort.

Given the results found in Chapter 3 and the multiple goals of conditional cash transfers an appealing alternative would be to modify the thresholds used by the BLE. A potential budget neutral policy would be to lower the income threshold (for example from the 30% to the 20% poorest as measured by the PMT) and raise the academic performance threshold (for example from the 30% to the 40% of highest academic achievement). This change would increase access to the cash transfer for the first quintile group of PMT scores as students between the 30% and 40% of academic performance and within the 20% poorest would be able to access the BLE. Even if no impacts were observed on educational outcomes for students near these new thresholds the modification should at least have a positive impact on income measures for the 20% poorest.

Beyond CCTs, the thesis shows that appropriate predictive models of school dropout are at hand for public officials. The models I implement in Chapter 2 rely only on data currently available in the Chilean government. Using these models for targeting, once large and rich administrative datasets exist within a country, can be cost-effective as no surveys are required. Where countries are improving their administrative records, this idea deserves consideration. Predictive models of this kind can prove useful not only for CCTs but also for further policies, such as Early Warning Systems, whose purpose is to prevent school dropouts. After the model has identified every student who is likely to drop out of school an initial diagnostic of the student could be implemented to understand the main drivers behind the likelihood of dropping out. A targeted intervention could follow to encourage them to keep enrolled in school.

The thesis also contributes in methodological grounds. For example, an essential and implicit lesson from the thesis is that effective targeting depends on consistency. There is a need to avoid misalignment between the policy goals, the target groups identified and the selection of the targeting mechanisms. Ideally, targeting design must follow the goals of the policy and its consequential definition of target groups. If a policy has multiple purposes and target groups, then using only one dimension for targeting is most likely not to be the most effective approach. For example, CCT targeting mechanisms, using primarily income measures, have not been fully aligned with CCT objectives and their target population (Azevedo & Robles, 2013).

In a similar vein, CCT targeting assessments should ideally cover the totality of the target groups involved. These lessons invite to broaden the outreach on CCT targeting evaluations, as these evaluations have provided valuable information though mostly regarding income measures (Maluccio, 2009; Robles et al., 2015; Skoufias et al., 2001; Stampini & Tornarolli, 2012). An additional methodological contribution of my thesis is that it offers two indicators to assess CCT targeting that combine information about poor students and future school dropouts.

Regarding causal inference methodological contributions, my work shows that administrative sorting and intermediate contamination may play a role in threatening or invalidating RD designs in contexts where indexes such as PMTs are used as a running variable. Given the central role of PMTs in the targeting of CCTs, these findings could be especially useful within this context.

In terms of administrative sorting, my thesis highlights the importance of fully understanding the data generation process behind any index. Chapter 3 shows that useful variation could still be used for identification as administrative procedures explain that students are not comparable near the 30% relative ranking (of academic performance) threshold used by the BLE. Chapter 4 demonstrates that administrative sorting explains the shape of two running variables' density near their respective thresholds and that minor adjustments would have prevented administrative sorting. Additionally, I show that a McCrary test can fail due to reasons other than manipulation of the running variable. This finding broadens our understanding of the interpretation of the test.

All these findings are important because if administrative sorting is not recognised correctly it could be mistaken for manipulation, potentially leading to viable research designs being discarded. The causes behind administrative sorting can potentially be fixed for future versions of the index and facilitate the implementation of an RD design. In other cases, for example when treatment probabilities are not affected, administrative sorting may not invalidate an RD design.

The thesis also emphasises that RD practitioners should take increased precautions in settings where their running variables are expected, or have been used, to serve multiple purposes. In contexts where the running variable is in place well in advance of the treatment of interest, RD designs are more prone to suffer from intermediate contamination. For example, in Colombia, to assess the impact of a CCT, Barrientos and Villa (2015) use a running variable and threshold previously used by a school fee reduction policy (Barrera-Osorio et al., 2007).

Considering all the RD method findings a complete routine for robustness checks in RD designs should include: a McCrary or Frandsen test to assess for manipulation or administrative sorting, and intent to treat estimates for baseline characteristics, intermediate variables and placebo outcomes to assess the plausibility of the continuity assumption on which RD designs rely.

5.3 Further Research

Many CCTs promote human capital accumulation beyond educational dimensions. For example, a common condition used by CCTs is the regular attendance of children no older than six years old at health check-ups. None of my empirical chapters consider health outcomes. For example, in Chapter 2, children who have not been taken by their parents to preventive health check-ups could be included as a target group for CCTs. Unfortunately, unlike educational variables, quality administrative records for CCT health-related outcomes are not available in Chile. Future research could be implemented, related to either CCT targeting or impact, after children's health-related administrative datasets improve in the country.

Further research could explore the feasibility of measuring the impact of Chilean CCTs on school enrolment, for example, assessing the impact of BARE on preventing school dropout and evaluating whether eligibility to receive SUF helps to bring back to school those who are outside. The findings of Chapter 4 could be helpful to the latter goals in an RD design. Administrative sorting explains the shape of the BARE and SUF targeting indexes near their respective thresholds. Fully understanding the generation process of the scores for each index could help in making amendments to implement RD designs in the future or to identify any useful variation left for identification in the present. Prospective research on BARE or SUF could rely on revised indexes that are not affected by administrative sorting. Retrospective research could exploit the categorisation of one variable in the BARE index and fully reconstruct the income prediction of the SUF index.

RD estimates are informative for the population around the thresholds but not necessarily away from them. Given this caveat, it is not possible to generalise the results of Chapter 3 for the entire population who received the BLE in 2013. As RD provides average effects for the population near the thresholds, we cannot learn about the effect of the BLE for all students at the lower end of the income distribution. An RD design estimating the impact of SUF on school enrolment would have a similar problem. This subgroup of students may be the one that is more susceptible to benefitting from conditional cash transfers. Future research could overcome these limitations by introducing some degree of randomisation, allowing for obtaining estimates for this entire subgroup of interest.

Acknowledging the limitations of the findings of Chapter 3 it is essential to understand more deeply why the BLE is producing little effect on educational outcomes. Interviewing parents and students could help us to comprehend the causal mechanisms (or lack of such mechanisms) between providing this cash transfer and subsequent gains in attendance and academic performance. Additionally, these interviews could prove useful to find out how aware families are of the BLE implementation. Given the nature of these inquiries, this research probably ought to be qualitative. Experiments could follow that analyse different modalities of implementing the BLE. For example, it could be tested whether giving a diploma to students increases their awareness of the BLE and subsequent outcomes.

Distinguishing between the poorest of the population and those who are close to a certain threshold (for example those who are marginally poor or who barely access a CCT given their proxy means test score) can be important in targeting assessments too. A limitation of Chapter 2 is that I implicitly assume that there is equal social value in finding any poor student. Given the measures I use, I make no distinction between students at the bottom of the income distribution and those who are slightly below the poverty line. Building upon my research, future projects could account for these differences to strengthen the targeting assessment.

Regarding the predictive model of school dropout, future research has room for improvement. The efficacy of the predictions could be further enhanced if the Chilean government were to add new administrative variables such as motherhood (which can be identified through the Civil Register administrative datasets). Major modifications to the structure of the predictive model also deserve further exploration. For example, the model could be run on a monthly instead of an annual basis. The Chilean Ministry of Education possesses monthly attendance records at the individual level that could be useful for capturing additional variation in school dropout. Another alternative would be to implement longitudinal growth models instead of machine learning algorithms. Also, the logic and variables I use in my model could be applied to predict other educational outcomes such as performance, attendance and enrolment in higher education.

Future research to uncover further improvements in CCT targeting, estimate the impact of Chilean CCTs on school enrolment and health-related outcomes, identify the effects for individuals in the lower end of the income distribution and provide a better understanding of the channels that explain the relationship, or lack of such a relationship, between CCTs and educational outcomes for Chilean students will inform policymakers allowing them to enhance the design of CCTs in the country. This research can be informative not just only for Chile but all nations using or considering CCTs in their toolkit.

5.4 Concluding Remarks

During the last 30 years the world has observed the rise of different types of social policies. Latin American countries were the first to implement income support schemes targeted towards poor households, conditional on recipients carrying out actions to foster human capital accumulation among their children and adolescents. These conditional cash transfers later expanded across the globe reaching Asia, the Middle East, Africa and even the United States. The surge in the prevalence of CCTs has been correlated with a rise in the use of income targeting through proxy means tests. CCTs continue to be an important social policy model across the globe. Their goals of poverty alleviation and human capital accumulation remain valid in multiple regions.

Have CCTs worked? In multiple countries CCTs have been successful in alleviating poverty, increasing health care utilisation, and raising school enrolment and attendance rates. Have they gone too far? In Latin America the number of beneficiaries of CCTs surpassed the number of the poor in 2006. Within the latter policy framework, my research has answered questions located at the intersection of CCTs, their targeting (especially through PMTs) and educational outcomes using large and rich Chilean administrative datasets. Although the thesis has focused its analysis on one country, it has provided findings and implications that will be useful in multiple contexts.

My thesis advances knowledge in the policy domains of the targeting and impact of CCTs. My findings show that CCT targeting can be improved if students at risk of dropping out of school are included as a target group and if a predictive model of school dropout is used in conjunction with a PMT for targeting. If a CCT has gone too far in terms of reaching non-poor households these findings offer an appealing policy alternative to increase their targeting effectiveness. Another appealing policy alternative is to partly replace CCTs with cash for grades initiatives. However, my research shows that a Chilean cash for grades scheme did not have a major effect on educational outcomes of students around the thresholds used for targeting. This finding suggests that some modifications to this programme are needed for increased impact.

My thesis offers substantial contributions in regard to methodological grounds too. Among these are one of the first machine learning applications to predict school dropout outside a developed nation, implementing a CCT targeting assessment using more dimensions than income, and highlighting administrative sorting as a threat to the internal validity in RD designs and potential solutions to overcome it. All these findings not only contribute to strengthening the future targeting and impact evaluations of CCTs but are also helpful beyond the context of CCT assessments.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2014). *Finite Population Standard Errors*. http://www.nber.org/papers/w20325.

Adelman, M., Haimovich, F., Ham, A., & Vazquez, E. (2017). *Predicting School Dropout With Administrative Data*. http://documents.worldbank.org/curated/en/273541499700395624/pdf/WPS8142.pdf.

Alatas, V., Banerjee, A., Hanna, R., Olken, B., Purnamasari, R., & Wai-Poi, M. (2016). Self-Targeting: Evidence from a Field Experiment in Indonesia. *Journal of Political Economy, 124*(2), 371-427.

Alatas, V., Banerjee, A., Hanna, R., Olken, B. A., & Tobias, J. (2012). Targeting the Poor: Evidence from a Field Experiment in Indonesia. *American Economic Review, 102*(4), 1206-1240.

Almeida, R., Orr, L., & Robalino, D. (2014). Wage Subsidies in Developing Countries as a Tool to Build Human Capital: Design and Implementation Issues. *IZA Journal of Labor Policy, 3*(1), 1-24.

Alzúa, M., Cruces, G., & Ripani, L. (2013). Welfare Programs and Labor Supply in Developing Countries: Experimental Evidence from Latin America. *Journal of Population Economics, 26*(4), 1255-1284.

Angrist, J., & Lavy, V. (2009). The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial. *American Economic Review, 99*(4), 1384-1414.

Angrist, J., & Pischke, S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton and Oxford: Princeton University Press.

Athey, S., & Imbens, G. W. (2015a). *Machine Learning Methods for Estimating Heterogeneous Causal Effects*. https://www.gsb.stanford.edu/gsb-cmis/gsb-cmis-download-auth/406621.

Athey, S., & Imbens, G. W. (2015b). A Measure of Robustness to Misspecification. *American Economic Review, 105*(5), 476-480.

Attanasio, O., Battistio, E., Fitzsimons, E., Mesnard, A., & Vera-Hernández, M. (2005). *How Effective Are Conditional Cash Transfers? Evidence from Colombia*. https://www.ifs.org.uk/bns/bn54.pdf.

Attanasio, O., Fitzsimons, E., Gomez, A., Gutiérrez, M. I., Meghir, C., & Mesnard, A. (2010). Children's Schooling and Work in the Presence of a Conditional Cash Transfer Program in Rural Colombia. *Economic Development and Cultural Change, 58*(2), 181-210.

Australian Aid. (2011). *Targeting the Poorest: An Assessment of the Proxy Means Test Methodology*. https://www.unicef.org/socialpolicy/files/targeting-poorest.pdf.

Azevedo, V., & Robles, M. (2013). Multidimensional Targeting: Identifying Beneficiaries of Conditional Cash Transfer Programs. *Social Indicators Research, 112(2)*, 447-475.

Baird, S., Ferreira, F., Özler, B., & Woolcock, M. (2014). Conditional, Unconditional and Everything in Between: A Systematic Review of the Effects of Cash Transfer Programmes on Schooling Outcomes. *Journal of Development Effectiveness, 6*(1), 1-43.

Barrera-Osorio, F., Bertrand, M., Linden, L. L., & Perez-Calle, F. (2011). Improving the Design of Conditional Transfer Programs: Evidence from a Randomized Education Experiment in Colombia. *American Economic Journal: Applied Economics, 3*(2), 167-195.

Barrera-Osorio, F., Linden, L. L., & Urquiola, M. (2007). *The Effects of User Fee Reductions on Enrollment. Evidence from a Quasi-Experiment*. http://siteresources.worldbank.org/EDUCATION/Resources/278200-1121703274255/1439264-1171379341729/SessionIII_FelipeBarrera3.pdf.

Barrientos, A., & Villa, J. M. (2015). Antipoverty Transfers and Labour Market Outcomes: Regression Discontinuity Design Findings. *The Journal of Development Studies, 51*(9), 1224-1240.

Bassi, M., Busso, M., & Muñoz, J. S. (2015). Enrollment, Graduation and Dropout Rates in Latin America: Is the Glass Half Empty or Half Full? *Economía, 16(1)*, 113-156.

Baulch, B. (2002). *Poverty Monitoring and Targeting Using ROC Curves: Examples from Vietnam*. http://www.ids.ac.uk/publication/poverty-monitoring-and-targeting-using-roc-curves-examples-from-vietnam.

Behrman, J., Parker, S., Todd, P., & Wolpin, K. (2015). Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools. *Journal of Political Economy, 123*(2), 325-364.

Besley, T. (1990). Means Testing versus Universal Provision in Poverty Alleviation Programmes. *Economica, 57*(225), 119-129.

Besley, T., & Kanbur, R. (1990). *The Principles of Targeting*. http://documents.worldbank.org/curated/en/212811468739258336/pdf/multi0page.pdf

Bettinger, E. (2012). Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores. *Review of Economics and Statistics, 94*(3), 686-698.

Biblioteca del Congreso Nacional de Chile. (1980). *Decreto Ley 3464. Aprueba Nueva Constitución Política y la Somete a Ratificación del Plebiscito*.

Biblioteca del Congreso Nacional de Chile. (1997). *Decreto 511 Exento. Aprueba Reglamento de Evaluación y Promoción Escolar de Niñas y Niños de Enseñanza Básica*.

Biblioteca del Congreso Nacional de Chile. (1999). *Decreto 112 Exento. Establece Disposiciones para que Establecimientos Educacionales Elaboren Reglamento de Evaluación y Reglamenta Promoción de Alumnos de Primer y Segundo Año de Enseñanza Media, Ambas Modalidades*.

Biblioteca del Congreso Nacional de Chile. (2003). *Ley 19876. Reforma Constitucional que Establece la Obligatoriedad y Gratuidad de la Educación Media*.

Biblioteca del Congreso Nacional de Chile. (2007). *Decreto 2169 Exento. Aprueba Reglamento de Evaluación y Promoción Escolar para Educación Básica y Media de Adultos*.

Biblioteca del Congreso Nacional de Chile. (2009). *Ley 20370. Establece la Ley General de Educación*.

Biblioteca del Congreso Nacional de Chile. (2011). *Decreto 1718 Exento. Determina las Fechas en que se Deberán Cumplir los Requisitos de Edad de Ingreso a la Educación Básica y Media Regular y la Fecha que se Considerará para el Ingreso al Primer y Segundo Nivel de Transición de la Educación Parvularia*.

Biblioteca del Congreso Nacional de Chile. (2012). *Decreto 332. Determina Edades Mínimas para el Ingreso a la Educación Especial o Diferencial, Modalidad de Adultos y de Adecuaciones de Aceleración Curricular*.

Biblioteca del Congreso Nacional de Chile. (2013). *Decreto 24. Aprueba Reglamento que Regula el Bono por Esfuerzo*.

Biblioteca del Congreso Nacional de Chile. (2015). *Ley 20845. De Inclusión Escolar que Regula la Admisión de los y las Estudiantes, Elimina el Financiamiento Compartido y Prohíbe el Lucro en Establecimientos Educacionales que Reciben Aportes del Estado*

Blanden, J., Del Bono, E., Hansen, K., & Rabe, B. (2017). *The Impact of Free Early Childhood Education and Care on Educational Achievement: A Discontinuity Approach Investigating Both Quantity and Quality of Provision.* https://www.surrey.ac.uk/sites/default/files/DP06-17_0.pdf.

Bowers, A., Sprott, R., & Taff, S. (2013). Do We Know Who Will Drop Out? A Review of the Predictors of Dropping Out of High School: Precision, Sensitivity and Specificity. *The High School Journal, 96(2)*, 77-100.

Brewer, M., & Browne, J. (2006). *The Effect of the Working Families' Tax Credit on Labour Market Participation*. http://www.ifs.org.uk/bns/bn69.pdf.

Brown, C., Ravallion, M., & Van de Walle, D. (2016). *A Poor Means Test? Econometric Targeting in Africa*. https://www.nber.org/papers/w22919.pdf.

Bushaw, W. J., & Lopez, S. J. (2010). *Highlights of the 2010 Phi Delta Kappa/Gallup Poll. What Americans Said About the Public Schools*. https://larrycuban.files.wordpress.com/2010/11/2010_poll_report1.pdf.

Calefati, J. (2008). Giving Students Cash for Grades. *US News*. Retrieved from https://www.usnews.com/education/articles/2008/11/28/giving-students-cash-for-grades.

Camacho, A., & Conover, E. (2011). Manipulation of Social Program Eligibility. *American Economic Journal: Economic Policy, 3*(2), 41-65.

Cameron, J. (2001). Negative Effects of Reward on Intrinsic Motivation—A Limited Phenomenon: Comment on Deci, Koestner, and Ryan (2001). *Review of Educational Research, 71*(1), 29–42.

Cameron, J., Banko, K. M., & Pierce, W. D. (2001). Pervasive Negative Effects of Rewards on Intrinsic Motivation: The Myth Continues. *The Behavior Analyst, 24*(1), 1-44.

Cameron, J., & Pierce, W. D. (1994). Reinforcement, Reward and Intrinsic Motivation: A Meta-Analysis. *Review of Educational Research, 64*(3), 363– 423.

Carneiro, P., Galasso, E., & Ginja, R. (forthcoming). Tackling Social Exclusion: Evidence from Chile. *The Economic Journal*.

Carpenter, C., & Dobkin, C. (2009). The Effect of Alcohol Consumption on Mortality: Regression Discontinuity Evidence from the Minimum Drinking Age. *American Economic Journal: Applied Economics, 1*(1), 164-182.

Cattaneo, M., & Escanciano, J. C. (2017). Introduction: Regression Discontinuity Designs. In M. Cattaneo & J. C. Escanciano (Eds.), *Advances in Econometrics* (Vol. 38, pp. i-xxv). Bingley, UK: Emerald Publishing Limited.

Cattaneo, M., Idrobo, N., & Titiunik, R. (2018a). A Practical Introduction to Regression Discontinuity Designs: Part I. In *Cambridge Elements: Quantitative and Computational Methods for Social Science*. Cambridge, UK: Cambridge University Press.

Cattaneo, M., Idrobo, N., & Titiunik, R. (2018b). A Practical Introduction to Regression Discontinuity Designs: Part II. In *Cambridge Elements: Quantitative and Computational Methods for Social Science*. Cambridge, UK: Cambridge University Press.

Cattaneo, M., Jansson, M., & Ma, X. (2017). *Simple Local Polynomial Density Estimators*. https://eml.berkeley.edu/~mjansson/Papers/CattaneoJanssonMa_LocPolDensity.pdf.

Cattaneo, M., Keele, L., Titiunik, R., & Vazquez-Bare, G. (2016a). Interpreting Regression Discontinuity Designs With Multiple Cutoffs. *Journal of Politics, 78*(4), 1229-1248.

Cattaneo, M., Titiunik, R., & Vazquez-Bare, G. (2017b). Comparing Inference Approaches for RD Designs: A Reexamination of the Effect of Head Start on Child Mortality. *Journal of Policy Analysis and Management, 36*(3), 643-681.

Centro de Microdatos. (2012). *Evaluación de Impacto del Programa Subsidio al Empleo Joven*. http://www.dipres.gob.cl/594/articles-119350_doc_pdf.pdf.

Coady, D. (2006). The Welfare Returns to Finer Targeting: The Case of the Progresa Program in Mexico. *International Tax and Public Finance, 13*(2-3), 217-239.

Coady, D., Grosh, M., & Hoddinott, J. (2004). *Targeting of Transfers in Developing Countries: Review of Lessons and Experience*. http://siteresources.worldbank.org/SAFETYNETSANDTRANSFERS/Resources/281945-1138140795625/Targeting_En.pdf.

Coady, D., & Skoufias, E. (2004). On the Targeting and Redistributive Efficiencies of Alternative Targeting Instruments. *Review of Income and Wealth, 50(1)*, 11-27.

Cockx, B., & Dejemeppe, M. (2012). Monitoring Job Search Effort: An Evaluation Based on a Regression Discontinuity Design. *Labour Economics, 19*(5), 729-737.

Comité de Expertos Ficha de Protección Social. (2010). *Informe Final Comité de Expertos Ficha de Protección Social*. http://www.ministeriodesarrollosocial.gob.cl/btca/txtcompleto/mideplan/c.e-fps-infinal.pdf.

Crawford, C., Dearden, L., & Meghir, C. (2010). *When You Are Born Matters: The Impact of Date of Birth on Educational Outcomes in England*. https://www.ifs.org.uk/wps/wp1006.pdf.

De la Mata, D. (2012). The Effect of Medicaid Eligibility on Coverage, Utilization, and Children's Health. *Health Economics, 21*(9), 1061-1079.

De Wachter, S., & Galiani, S. (2006). Optimal Income Support Targeting. *International Tax and Public Finance, 13(6)*, 661-684.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation. *Psychological Bulletin, 125*(6), 627– 668.

Deci, E. L., Koestner, R., & Ryan, R. M. (2001). Extrinsic Rewards and Intrinsic Motivation in Education: Reconsidered Once Again. *Review of Educational Research, 71*(1), 1–27.

Del Bono, E., & Galindo-Rueda, F. (2007). *The Long Term Impacts of Compulsory Schooling: Evidence from a Natural Experiment in School Leaving Dates*. https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2006-44.pdf.

Dickens, R., Riley, R., & Wilkinson, D. (2014). The UK Minimum Wage at 22 Years of Age: A Regression Discontinuity Approach. *Journal of the Royal Statistical Society, 177*(1), 95-114.

Dickert-Conlin, S., & Holtz-Eakin, D. (2000). Employee-Based versus Employer-Based Subsidies to Low-Wage Workers: A Public Finance Perspective. In D. Card & R. Blank (Eds.), *Finding Jobs: Work and Welfare Reform* (pp. 262-296). New York: Russell Sage Foundation.

Dong, Y. (2015). Regression Discontinuity Applications With Rounding Errors in the Running Variable. *Journal of Applied Econometrics, 30*(3), 422–446.

Eissa, N., & Hoynes, H. (2004). Taxes and the Labor Market Participation of Married Couples: The Earned Income Tax Credit. *Journal of Public Economics, 88*(9), 1931-1958.

Eissa, N., & Liebman, J. B. (1996). Labor Supply Response to the Earned Income Tax Credit. *The Quarterly Journal of Economics, 111*(2), 605-637.

Elder, T. (2010). The Importance of Relative Standards in ADHD Diagnoses: Evidence Based on Exact Birth Dates. *Journal of Health Economics, 29*(5), 641-656.

Filmer, D., & Schady, N. (2011). Does More Cash in Conditional Cash Transfer Programs Always Lead to Larger Impacts on School Attendance? *Journal of Development Economics, 96*(1), 150-157.

Fiszbein, A., & Schady, N. (2009). *Conditional Cash Transfers: Reducing Present and Future Poverty*. https://siteresources.worldbank.org/INTCCT/Resources/5757608-1234228266004/PRR-CCT_web_noembargo.pdf.

Fitzsimons, E., & Vera-Hernández, M. (2016). Breastfeeding and the Weekend Effect: An Observational Study. *BMJ Open*; 6:7.

Focus. Consultorías y Estudios. (2016). *Evaluación de Impacto Subsidio Familiar y Asignación Familiar*. http://www.dipres.gob.cl/595/articles-146449_informe_final.pdf.

Frandsen, B. (2017). Party Bias in Union Representation Elections: Testing for Manipulation in the Regression Discontinuity Design When the Running Variable Is Discrete. In M. D. Cattaneo & J. C. Escanciano (Eds.), *Advances in Econometrics* (Vol. 38, pp. 281-315). Bingley, UK: Emerald Publishing Limited.

Fredriksson, P., & Ockert, B. (2005). *Is Early Learning Really More Productive? The Effect of School Starting Age on School and Labor Market Performance*. https://pdfs.semanticscholar.org/ac91/6c36bed5d153cc9695938c0d7451eb988e2d.pdf

Fryer, R. (2011). Financial Incentives and Student Achievement: Evidence from Randomized Trials. *Quarterly Journal of Economics, 126*(5), 1755-1798.

Galiani, S., & McEwan, P. J. (2013). The Heterogeneous Impact of Conditional Cash Transfers. *Journal of Public Economics, 103*, 85-96.

Gans, J., & Leigh, A. (2008). *What Explains the Fall in Weekend Births?* https://www.academia.edu/2808187/What_Explains_the_Fall_in_Weekend_Births.

García, S., & Saavedra, J. (2017). Educational Impacts and Cost-Effectiveness of Conditional Cash Transfer Programs in Developing Countries: A Meta-Analysis. *Review of Educational Research, 87*(5), 921-965.

Glewwe, P. (1992). Targeting Assistance to the Poor: Efficient Allocation of Transfers When Household Income Is not Observed. *Journal of Development Economics, 38(2)*, 297-321.

Glewwe, P., & Olinto, P. (2004). *Evaluating the Impact of Conditional Cash Transfers on Schooling: An Experimental Analysis of Honduras' PRAF Program*. http://web.worldbank.org/archive/website01404/WEB/IMAGES/GLEWWEOL.PDF.

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and Why Incentives (Don't) Work to Modify Behavior. *Journal of Economic Perspectives, 25*(4), 191-210.

Grosh, M., & Baker, J. (1995). *Proxy Means Tests for Targeting of Social Programs: Simulations and Speculation*. http://documents.worldbank.org/curated/en/750401468776352539/pdf/multi-page.pdf.

Guttenplan, D. D. (2011). Motivating Students With Cash-for-Grades Incentive. *The New York Times*. Retrieved from https://www.nytimes.com/2011/11/21/world/middleeast/21iht-educLede21.html.

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and Estimation of Treatment Effects With a Regression-Discontinuity Design. *Econometrica, 69*(1), 201-209.

Handa, S., & Davis, B. (2006). The Experience of Conditional Cash Transfers in Latin America and the Caribbean. *Development Policy Review, 24*(5), 513-536.

Hanna, R., & Olken, B. (2018). Universal Basic Incomes versus Targeted Transfers: Anti-Poverty Programs in Developing Countries. *Journal of Economic Perspectives, 32*(4), 201-226.

Helleringer, S., Asuming, P. O., & Abdelwahab, J. (2016). The Effect of Mass Vaccination Campaigns Against Polio on the Utilization of Routine Immunization Services: A Regression Discontinuity Design. *Vaccine, 34*(33), 817–3822.

Higgins, L. (2015). Think and Grow Rich? Michigan School Offers Cash for Grades. *USA Today*. Retrieved from https://www.usatoday.com/story/news/nation-now/2015/11/01/michigan-high-school-cash-for-grades/75003438/.

Hirshleifer, S. (2017). *Incentives for Effort or Outputs? A Field Experiment to Improve Student Performance*. https://economics.ucr.edu/repec/ucr/wpaper/201701.pdf.

Honorati, M., Gentilini, U., & Yemtsov, R. G. (2015). *The State of Social Safety Nets*. http://documents.worldbank.org/curated/en/415491467994645020/pdf/97882-PUB-REVISED-Box393232B-PUBLIC-DOCDATE-6-29-2015-DOI-10-1596978-1-4648-0543-1-EPI-1464805431.pdf.

Hunt, F. (2008). *Dropping Out from School: A Cross Country Review of Literature*. http://www.create-rpc.org/pdf_documents/PTA16.pdf.

Ibarrarán, P., Medellín, N., Regalia, F., & Stampini, M. (2017). *How Conditional Cash Transfers Work: Good Practices after 20 Years of Implementation*. https://publications.iadb.org/bitstream/handle/11319/8159/How-conditional-cash-transfers-work.PDF?sequence=9.

Imbens, G. W., & Lemieux, T. (2008). Regression Discontinuity Designs: A Guide to Practice. *Journal of Econometrics, 142*(2), 615-635.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. New York: Springer.

Kabeer, N., Piza, C., & Taylor, L. (2012). *What Are the Economic Impacts of Conditional Cash Transfer Programmes? A Systematic Review of the Evidence*. https://assets.publishing.service.gov.uk/media/57a08a6840f0b649740005a4/CCTprogrammes2012Kabeer.pdf.

Katz, L. (1998). Wage Subsidies for the Disadvantaged. In R. Freeman & Gottschalk (Eds.), *Generating Jobs* (pp. 21-53). New York: Russell Sage Foundation.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction Policy Problems. *The American Economic Review, 105(5)*, 491-495.

Knowles, J. (2015). Of Needles and Haystacks: Building an Accurate Statewide Early Warning System in Wisconsin. *Journal of Educational Data Mining, 7(3)*, 18-67.

Kohn, A. (1999). *Punished by Rewards: The Trouble With Gold Stars, Incentive Plans, A's, Praise, and Other Bribes*. Boston: Houghton Mifflin.

Kolesár, M., & Rothe, C. (2018). Inference in Regression Discontinuity Designs With a Discrete Running Variable. *American Economic Review, 108*(8), 2277-2304.

Kremer, M., Miguel, E., & Thornton, R. (2009). Incentives to Learn. *Review of Economics and Statistics, 91*(3), 437-456.

Kuhn, M. (2008). Building Predictive Models in R Using the Caret Package. *Journal of Statistical Software, 28(5)*, 1-26.

Lagarde, M., Haines, A., & Palmer, N. (2007). Conditional Cash Transfers for Improving Uptake of Health Interventions in Low and Middle-Income Countries. *JAMA, 298*(16), 1900-1910.

Lamote, C., Van Damme, J., Van den Noortgate, W., Speyboreck, S., Boonen, T., & Bilde, J. (2013). Dropout in Secondary Education. An Application of a Multilevel Discrete-Time Hazard Model Accounting for School Changes. *Quality & Quantity, 47(5)*, 2425-2446.

Lee, D. (2008). Randomized Experiments from Non-Random Selection in U.S. House Elections. *Journal of Econometrics, 142*(2), 675-697.

Lee, D., & Card, D. (2008). Regression Discontinuity Inference With Specification Error. *Journal of Econometrics, 142*(2), 655–674.

Lee, D., & Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature, 48*(2), 281-355.

Lemieux, T., & Milligan, K. (2008). Incentive Effects of Social Assistance: A Regression Discontinuity Approach. *Journal of Econometrics, 142*(2), 807-828.

Levitt, S. D., List, J. A., Neckermann, S., & Sadoff, S. (2012). *The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance*. https://www.nber.org/papers/w18165.pdf.

Lindo, J. M., Sanders, N. J., & Oreopoulos, P. (2010). Ability, Gender, and Performance Standards: Evidence from Academic Probation. *American Economic Journal: Applied Economics, 2*(2), 95-117.

Maluccio, J. (2009). Household Targeting in Practice: The Nicaraguan Red de Protección Social. *Journal of International Development, 21(1)*, 1-23.

Maluccio, J., & Flores, R. (2005). *Impact Evaluation of a Conditional Cash Transfer Program: The Nicaraguan Red de Protección Social*. http://www.ifpri.org/publication/impact-evaluation-conditional-cash-transfer-program-2.

McBride, L., & Nichols, A. (2016). Retooling Poverty Targeting Using Out-of-Sample Validation and Machine Learning. *The World Bank Economic Review, 32*(3), 531-550.

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

McEwan, P., & Shapiro, J. (2007). The Benefits of Delayed Primary School Enrollment. Discontinuity Estimates Using Exact Birth Dates. *The Journal of Human Resources, 43(1)*, 1-29.

Meyer, B. D., & Rosenbaum, D. T. (2001). Welfare, the Earned Income Tax Credit, and the Labor Supply of Single Mothers. *The Quarterly Journal of Economics, 116*(3), 1063-1114.

Ministerio de Desarrollo Social. (2015). *Casen 2013. Una Medición de la Pobreza Moderna y Transparente para Chile*. http://observatorio.ministeriodesarrollosocial.gob.cl/documentos/Presentacion_Resultados_Encuesta_Casen_2013.pdf.

Ministerio de Educación. (2013). *Serie Evidencias: Medición de la Deserción Escolar en Chile*. https://centroestudios.mineduc.cl/wp-content/uploads/sites/100/2017/06/A2N15_Desercion_escolar.pdf.

Molina-Millan, T., Barham, T., Macours, K., Maluccio, J. A., & Stampini, M. (2016). *Long-Term Impacts of Conditional Cash Transfers in Latin America: Review of the Evidence.* https://publications.iadb.org/handle/11319/7891.

Morris, S., Flores, R., Olinto, P., & Medina, J. M. (2004). Monetary Incentives in Primary Health Care and Effects on Use and Coverage of Preventive Health Care Interventions in Rural Honduras: Cluster Randomised Trial. *The Lancet, 364*(9450), 2030-2037.

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives, 31(2)*, 87-106.

Neumark, D. (2013). Spurring Job Creation in Response to Severe Recessions: Reconsidering Hiring Credits. *Journal of Policy Analysis and Management, 32*(1), 142-171.

Nichols, A. (2018). *Implementing Machine Learning Methods in Stata*. https://www.stata.com/meeting/uk18/slides/uk18_Nichols.pdf.

OECD. (2015). *National Accounts at a Glance 2015*. http://www.oecd-ilibrary.org.gate2.library.lse.ac.uk/economics/national-accounts-at-a-glance-2015_na_glance-2015-en.

Opazo, V., Ormazabal, C., & Crespo, C. (2015). *Informe Final de Evaluación. Beca de Apoyo a la Retención Escolar*. http://www.dipres.gob.cl/597/articles-141242_informe_final.pdf.

Owusu-Addo, E., & Cross, R. (2014). The Impact of Conditional Cash Transfers on Child Health in Low and Middle-Income Countries: A Systematic Review. *International Journal of Public Health, 59*(4), 609-618.

Peyre Dutrey, A. (2007). *Successful Targeting? Reporting Efficiency and Costs in Targeted Poverty Alleviation Programmes*. http://www.unrisd.org/80256B3C005BCCF9/(httpAuxPages)/0B87C67449C938EDC12573D10049830B/$file/Peyrepap.pdf.

Ponce, J., & Bedi, A. (2010). The Impact of a Cash Transfer Program on Cognitive Achievement: The Bono de Desarrollo Humano of Ecuador. *Economics of Education Review, 29*(1), 116-125.

Pop-Eleches, C., & Urquiola, M. (2013). Going to a Better School: Effects and Behavioral Responses. *American Economic Review, 103*(4), 1289–1324.

Ravallion, M., & Chao, K. (1989). Targeted Policies for Poverty Alleviation Under Imperfect Information. *Journal of Policy Modelling, 11(2)*, 213-224.

Reardon, S. F., & Robinson, J. P. (2012). Regression Discontinuity Designs With Multiple Rating-Score Variables. *Journal of Research on Educational Effectiveness, 5*(1), 83-104.

Riccio, J., Dechausay, N., Miller, C., Nunez, S., Verma, N., & Yang, E. (2013). *Conditional Cash Transfers in New York City: The Continuing Story of the Opportunity NYC-Family Rewards Demonstration*. https://www.mdrc.org/sites/default/files/Conditional_Cash_Transfers_FR%202-18-16.pdf.

Ripley, A. (2010). Should Kids Be Bribed to Do Well in School? *Time Magazine*.

Roberts, D., Becker, C., & Ibanga, I. (2008). Chicago Offers Students Cash for Good Grades. *ABC News*. Retrieved from https://abcnews.go.com/GMA/Parenting/story?id=6371073&page=1.

Robles, M., Rubio, M., & Stampini, M. (2015). *Have Cash Transfers Succeeded in Reaching the Poor in Latin America and the Caribbean*. https://publications.iadb.org/handle/11319/7223.

Rumberger, R. W., & Lim, S. A. (2008). *Why Students Drop Out of School. A Review of 25 Years of Research*. https://www.issuelab.org/resources/11658/11658.pdf.

Sara, N. B., Halland, R., Igel, C., & Alstrup, S. (2015). High-School Dropout Prediction Using Machine Learning: A Danish Large-Scale Study. In M. Verleysen (Ed.), *Proceedings. ESANN 2015: 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* (pp. 319-324). Louvain-la-Neuve, Belgium: Ciaco.

Schady, N. R., & Araujo, M. C. (2008). Cash Transfers, Conditions, and School Enrollment in Ecuador. *Economía, 8*(2), 43-70.

Schultz, T. P. (2004). School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program. *Journal of Development Economics, 74*(1), 199-250.

Sekhon, J., & Titiunik, R. (2017). On Interpreting the Regression Discontinuity Design as a Local Experiment. In M. D. Cattaneo & J. C. Escanciano (Eds.), *Advances in Econometrics* (Vol. 38, pp. 1-28). Bingley, UK: Emerald Publishing Limited.

Sidorkin, A. M. (2007). Is Schooling a Consumer Good? A Case Against School Choice, But Not the One You Had in Mind. *Philosophy of Education*, 75–83.

Sidorkin, A. M. (2009). *Labor of Learning: Market and the Next Generation of Educational Reform*. Rotterdam: Sense.

Skoufias, E. (2005). *PROGRESA and Its Impacts on the Welfare of Rural Households in Mexico*. https://ageconsearch.umn.edu/bitstream/37891/2/rr139.pdf.

Skoufias, E., Davis, B., & De la Vega, S. (2001). Targeting the Poor in Mexico: An Evaluation of the Selection of Households into PROGRESA. *World Development, 29(10)*, 1769-1784.

Skovron, C., & Titiunik, R. (2015). *A Practical Guide to Regression Discontinuity Designs in Political Science*. https://pdfs.semanticscholar.org/5461/c817976f51a4fb0073b772c03cd670be8def.pdf.

Soares, F. V., Ribas, R. P., & Osório, R. G. (2010). Evaluating the Impact of Brazil's Bolsa Família: Cash Transfer Programs in Comparative Perspective. *Latin American Research Review, 45*(2), 173-190.

Sorensen, L. (2018). "Big Data" in Educational Administration: An Application for Predicting School Dropout Risk. *Education Administration Quarterly*, 1-43.

Stampini, M., Martinez-Cordova, S., Insfran, S., & Harris, D. (2018). Do Conditional Cash Transfers Lead to Better Secondary Schools? Evidence from Jamaica's PATH. *World Development, 101*(C), 104-118.

Stampini, M., & Tornarolli, L. (2012). *The Growth of Conditional Cash Transfers in Latin America and the Caribbean: Did They Go Too Far?* http://ftp.iza.org/pp49.pdf.

Theodoridis, S., & Koutroumbas, K. (2009). *Pattern Recognition*. California: Elsevier.

Toppo, G. (2008). Good Grades Pay Off Literally. *USA Today*. Retrieved from https://usatoday30.usatoday.com/news/education/2008-01-27-grades_N.htm.

Universidad del Desarrollo. (2014). *Evaluación de Impacto de la Bonificación Ingreso Ético Familiar*. http://www.dipres.gob.cl/595/articles-141198_informe_final.pdf.

Vakis, R., Rigolini, J., & Lucchetti, L. (2016). *Left Behind. Chronic Poverty in Latin America and the Caribbean*. http://documents.worldbank.org/curated/en/334891469074274116/pdf/107159-PUB-Box396279B-PUBLIC-PUBDATE-7-19-16.pdf.

Van der Vaart, A. W., Dudoit, S., & Van der Laan, M. J. (2006). Oracle Inequalities for Multi-Fold Cross Validation. *Statistics and Decisions, 24*(3), 351-371.

Varian, H. (2014). Big Data. New Tricks for Econometrics. *Journal of Economic Perspectives, 28(2)*, 3-28.

Warnick, B. (2017). Paying Students to Learn: An Ethical Analysis of Cash for Grades Programmes. *Theory and Research in Education, 15*(1), 71–87.

Wodon, Q. (1997). Targeting the Poor Using ROC Curves. *World Development, 25(12)*, 2083-2092.

Wong, V. C., Steiner, P. M., & Cook, T. D. (2013). Analyzing Regression-Discontinuity Designs With Multiple Assignment Variables: A Comparative Study of Four Estimation Methods. *Journal of Educational and Behavioral Statistics, 38*(2), 107-141.

Yörük, B., & Yörük, C. (2011). The Impact of Minimum Legal Drinking Age Laws on Alcohol Consumption, Smoking, and Marijuana Use: Evidence from a Regression Discontinuity Design Using Exact Date of Birth. *Journal of Health Economics, 30*(4), 740-752.