From Secondary School to the Labor Market in Colombia:

Turning Barriers into Building Blocks

Dissertation in order to acquire the doctoral degree from the Faculty of Economic Sciences,
at the Georg-August-Universität Göttingen

Submitted by

Luis Omar Herrera Prada

Born in Bogotá, Colombia

doi: 10.53846/goediss-10372

Göttingen, 2024

First Supervisor: Prof. Inmaculada Martínez-Zarzoso, Ph.D

Second Supervisor: Prof. Dario Maldonado, PhD

Third Supervisor: Prof. Dr Krisztina Kis-Katos, PhD

Date of oral examination February 2, 2024

Acknowledgments:

I am deeply indebted to Mary Lucy Herrera Contento and Inmaculada Martínez-Zarzoso. Mary Lucy's support allowed me to finish high school, enroll in college, and pursue my master's. Inmaculada (my advisor) not only opened the door for me to accomplish my dream of reaching the doctorate level, her patience, guides, and wise comments were key and drove me to deliver this document.

I am incredibly grateful to Juan Esteban Saavedra, who landed my ideas and motivated me during the process. Andrés Córdoba, who helped me with the code when I was lost. Sarah Belford encouraged me, discussed, and helped me with the English to improve this document.

I would like to extend my sincere thanks to Diana Durán, Paula Andrea Triana Romero, Elkin Garavito, Fabio Sánchez, Mónica Ospina, Carolina Guzmán, César López, Martha Susana Jaimes, and Jorge Franco. My discussions with them ignited the questions that triggered this document. I must also thank the SPADIES team 2005-2011 at Universidad de los Andes, Luz Emilse Rincón, Dolly Ovalle, and Activisual. They helped me with the data collection and process. I am also grateful to Fernando Rios-Ávila; my findings were incomplete without your code.

I would also like to extend my gratitude to Pelin Berkmen, Gustavo Canavire, Julián Cristia, Pablo Druck, Julia Faltermeier, Dmitry Gershenson, Luca Ricci, Belén Sbrancia, Miguel Urquiola, Hannah Winterberg, Román Andrés Zárate, and Román David Zárate. They showed interest in my research documents and provided valuable comments in many aspects.

Finally, I would like to express my most profound appreciation to my dad, mom, and brothers; my family was a keystone in this journey. Óscar Bernal is like my big brother. Clara Paola Pava significantly influenced my path, and through me, touched the lives of many others. God only knows what I'd be without you.

Comments received in the Development Economics Seminar - Universität Göttingen; WHD Seminar - International Monetary Fund; Public Policy Seminar - Universidad de los Andes (Bogotá-Colombia); Jornada NIP 2020-Universidad de la Republica (Montevideo-Uruguay); Midwest Economics Association (MEA-SOLE) 2021 and 2022 meeting; Association for Education Finance and Policy (AEFP) 2021; Colombian Central Bank Research Network Meeting 2022; Latin America and Caribbean Economics Association 2022 meeting, and International Economic Association World Congress 2023 were invaluable. All views and errors are only mine.

All IMF disclaimers apply, the views expressed in this thesis are mine and do not necessarily represent the views of the IMF, its Executive Board, or its management.

Keep interested in your own career, however humble; it is a real possession in the changing fortunes of time...

life is full of heroism...

Contents

List of Tables

List of Figures

List of Acronyms

ANM	National Agency of Mining
ATT	Average Treatment effect on Treated
bps	Basic Points
Co-As	Coal and Asbestos
COP	Colombian Peso
DANE	National Department of Statistics
DiD	Differences in Differences
DRI	Doulbly Robust Estimator
FE	Fixed Effects
GDP	Gross Domestic Product
Ha	Hectare
HE	Higher Education
HEI	Higher Education Institution
HS	Household Surveys
ICETEX	Colombian Bureau for Public Student Loans
ICFES	Colombian Institute for the Promotion of Higher Education
IMF	International Monetary Fund
IMP	Improved Doulbly Robust Estimator
IPW	Inverse Probability Weights
IV	Instrumental Variables

km	Kilometer
LAC	Latin America and the Caribbean
LATE	Local Average Treatment Effect
MEN	Ministry of Education
mmw	Minimum Monthly Wage
NL2SLS	Non-linear Two-stage Least Squares
OLE	Observatory for Educated Labor
OLS	Ordinary Least Squares
OM	Operating Mine
OR	OLS Regression
p.p	Percentual Points
PILA	Planilla Integrada de Liquidación de Aportes (Social Security Records)
PMETA	IMF Base Metals Price Index
PTA	Parallel Trend Assumption
RE	Random Effects
SACES	National System of Quality Certification in Higher Education
SENA	National Apprenticeship System
SNIES	National Information System for Higher Education
SPADIES	System for the Prevention and Analysis of Dropout in Higher Education Institutions
STEM	Science, Technology, Engineering, and Math
TWFE	Two-Way Fixed Effects
USD	United States Dollar

Chapter 1

Introduction

Education is one of the essential components of effective social policy. A nation's resources must be utilized judiciously to enhance well-being and promote social and economic development. Countries have long viewed education as a tool to facilitate their progress. A more educated populace is expected to be more productive (Becker, 1962) and experience greater social mobility, thereby promoting equity. However, individuals from impoverished backgrounds face numerous obstacles, including limited access to resources and a high opportunity cost associated with pursuing higher education. Consequently, many choose not to attend or drop out of higher education programs, perpetuating a vicious cycle of poverty. A dearth of human capital makes obtaining further knowledge and skills challenging. Moreover, the wage gap between workers with high and low levels of education has increased, and unequal access to higher education contributes to this wage disparity. Low enrollment rates result in a concentration of high-income individuals in the limited number of available spots, exacerbating the financial challenges of low-income earners. As the supply of uneducated people increases, their wages decrease, leading to a vicious cycle of poverty. This oversupply is also reflected in rising unemployment, intensifying financial vulnerability and augmenting the wage differential (Morley, 2002, Thurow, 1978; Weller, 2000).

The overall education level of a population serves as a barometer of a country's economic well-being. However, higher education levels that fail to align with labor market needs do not yield meaningful results. Therefore, to foster productivity, competitiveness, and prosperity, public policy must also ensure a level of quality that hinges on the ability of the higher education system to impart its knowledge to the productive sector. It is an onerous task, and it poses a significant challenge. It requires implementation and extension to the entire populace to ensure equitable opportunities and enhance the country's economy (CEPAL, 2004; Schultz, 1981).

Despite variations in their specific contexts, developing countries share common challenges in their pursuit of societal advancement. In the 2000s, Latin American countries confronted a universal challenge: a surge in demand for higher education, particularly among individuals with low income and limited qualifications, leading to diverse outcomes. This dissertation delves into the case of Colombia, a nation that has made remarkable strides in educational indicators during this century. Colombia's progress has positioned it as a regional role model, garnering attention in the literature on the economics of education. The World Bank has acknowledged the improvements in Colombia's educational system, highlighting the country's distinction in reporting the highest increase in enrollment rates and education expenditure as a percentage of GDP between 2000 and 2013. Furthermore, Colombia has reported the highest returns to education in the 2010s (Ferreyra et al., 2017).

However, significant efforts to increase attendance rates and achieve social improvements can falter if students drop out. A considerable portion of the success highlighted by the World Bank can be attributed to reductions in the dropout rate, as both attendance and returns on education hinge on students not leaving prematurely. The second chapter of this thesis investigates the effectiveness of the Colombian government's program, which utilized technology to facilitate information sharing among all market actors, in increasing enrollment and reducing dropout rates in higher education. The third chapter delves into the impact of college attendance on the lives of high school graduates, with explicit attention to comparing the returns on education for those who graduated, those who dropped out, and those who never attended college. Finally, the last chapter analyzes the influence of the exploitation of natural resources near schools on secondary and tertiary school outcomes. Given the diverse natural resources, environmental issues, and socio-economic contexts across regions, including rural and urban areas, those in rural settings often face challenges in accessing education and securing higher salaries. Additionally, considering the relevance of natural resource extraction in Colombia, this investigation explores how the demand for extraction might impact or improve educational outcomes for the population residing near these areas.

The Colombian education system's development has paralleled Colombian society's evolution. In the first half of the 20th century, the surge of the "boomer" generation increased the demand for education, leading to what was later termed the massification of higher education. To meet the demand, private institutions emerged (Lucio and Serrano, 1992). However, in the 1990s, another pressure on resources emerged as the children of the "boomers" entered high school. Intending to increase the enrollment rate, the government allowed for "automatic promotion," where students were promoted to the next academic year without meeting any academic requirements. As a result, educational quality was significantly reduced. During the early 2000s, many young Colombians sought to attend college, but the country's enrollment rate was relatively low, especially compared to other countries in the region. While the infrastructure was adequate to accommodate these new students, the Colombian economic situation and the coordination between the Secondary and Tertiary education systems were not optimal. Tertiary education was inaccessible to those with limited financial resources and/or less academic preparation.

To address the challenge of increasing access to higher education, the Colombian government implemented a strategy known as the "Revolución Educativa" (Education Revolution). The "Revolución Educativa" program positively affected the Colombian education system; as higher education enrollment rates increased, the dropout rate decreased, and the graduation rate stagnated. However, it also created new challenges, such as the need to improve education quality and provide adequate resources and support to the increasing number of students. The "Crowding Cohort"[1] phenomenon, resulting from the massive enrollment of students, highlights the importance of addressing the quality of education and ensuring that the educational system can cope with the demand. Overall, the case of Colombia shows that improving educational indicators is a complex process that requires a comprehensive and coordinated approach that considers the different aspects of the educational system and the needs of the population (Herrera-Prada and Kugler, 2017; Ministry of Education, 2010; Orozco, 2010).

This thesis makes a substantial contribution to knowledge by leveraging unique information generated by the Colombian education system. The level of detail presented in this research is unprecedented and, to the best of my knowledge, has not been explored in any previous studies with such granularity. The dissertation is organized into three essays that scrutinize the quality of the education system, the influence of externalities such as mining and a government program on diverse educational and labor market outcomes, and the returns of the higher education system in Colombia. Specifically, the thesis assesses the enrollment and performance of individuals in higher education and the formal labor market.

The advent of the new millennium presented a fresh challenge to the Colombian education system, with more students entering higher education who needed more preparation and had lower incomes than their predecessors. This resulted in a decline in the overall quality of the higher education system. In the first essay of this thesis, titled "How a Data-driven Tool Ended the Musical Chair Game in Higher Education in Colombia", I explore one of the tools employed by Colombia to address this dual challenge. The System for the Prevention and Analysis of Dropout in Higher Education Institutions (SPADIES) is a software dashboard that enables the collection, analysis, and visualization of student data, including modules that help institutions identify and track at-risk students. SPADIES also allows higher education institutions (HEIs) to compare their dropout rates to national or regional averages

[1] "Crowding cohort" is a phenomenon when there is increased demand for fixed or reducing resources. This term was used for the first time in Bound and Turner (2006).

and other HEIs. By facilitating the exchange of information between agents, SPADIES effectively improved higher education outcomes throughout the country. The positive impact of SPADIES on dropout rates and overall HEI quality was evident, regardless of the equilibrium model employed, including those that maximize income or quality from Epple et al. (2006) or the effect of peers from MacLeod and Urquiola (2015). Through enhancing access to and utilization of student and school performance data, SPADIES proved instrumental in significantly improving educational outcomes.

This thesis provides compelling evidence that SPADIES played a crucial role in enhancing higher education outcomes in Colombia. Applying a differences-in-differences approach within the framework of Callaway and Sant'Anna (2021), I demonstrate that SPADIES significantly reduced the probability of students becoming dropouts by an average of 70 basis points (bps). This effect was even more pronounced after 5 years of SPADIES implementation, reaching up to 210 bps (or 2.1 percent which is the 4.4% of the control mean) and saving around 14,000 students from dropping out—a remarkable achievement considering the average size of a higher education institution (HEI) in Colombia is 8,000 students. Furthermore, SPADIES increased the probability of students earning their degree on time by an average of 40 bps and earning their degree overall by 60 bps. After 5 years of SPADIES' installation, these figures rose to 1.4 and 1.9 percent, respectively, equivalent to 6.2% and 7.3% of the control mean. In total, SPADIES contributed to approximately 12,000 students earning their degrees, with 8,000 achieving this milestone on time. With a total cost of USD 4.5 million from 2005 to 2017 (in 2022 prices), the cost per saved-from-dropout student through SPADIES was approximately USD 320.5, and the cost per graduated student was around USD 373.9. These figures underscore SPADIES as a potentially cost-effective tool for advancing human capital formation in developing countries.

In the second essay, titled "Returns to Education in Colombia: New Empirical Evidence with a Comprehensive Dataset," I conducted a comprehensive analysis encompassing all secondary school graduates in Colombia from 2002 to 2012, totaling 5.4 million graduates.

Utilizing a modified Mincer equation within a differences-in-differences framework under Callaway and Sant'anna (2021)'s methodology, I estimated the premium of pursuing higher education by comparing the incomes of secondary school graduates who attended college with those of their fellow graduate classmates who did not pursue higher education. Additionally, I employed an instrumental variable (IV) approach to gauge the causal impact of attending college on wages. The analysis also delves into estimating the "Sheepskin Effect," a phenomenon where individuals with an academic degree earn a higher income than those with an equivalent schooling level but without the credential, first analyzed by Hungerford and Solon in 1987. The results indicate that attending college leads to a 38.4% increase in earnings compared to those who did not pursue higher education. However, for those who obtain a degree, the premium significantly rises to 50.6%, with specific premiums of 53.1% for bachelor's degree graduates and 92.83% for diploma holders (a graduate degree between bachelor and master's, detailed later in the document). Calculating the Sheepskin Effect involved computing the difference in Average Treatment Effect on the Treated (ATT) for the post-college period between graduates (50.6%) and individuals who completed over 90% of the coursework but did not receive a degree, resulting in a Sheepskin Effect of 68.1%.

In the essay titled "The Impact of Mining on Education and Labor Market Outputs in Colombia," I investigate the impact of mining resource exploitation on educational outcomes and labor market outputs. Developing countries, including Colombia, often possess abundant natural resources. I leverage the fact that mineral allocations were randomly assigned before population settlement, creating a quasi-experimental design. This design utilizes the operation of the nearest mine to each school in Colombia as a natural experiment, considering it as the treatment for the nation's secondary graduates from 2002 to 2012. I use a Difference in Difference approach under the Callaway and Sant'anna (2021) framework and an Instrumental Variables (IV) approach to establish a causal link between mining activities and various educational and labor market outcomes in Colombia. The study reveals significant heterogeneity among types of extracted products and their corresponding outcomes. Consistent with the findings of Angrist and Kugler (2008), the analysis indicates that the size of cohorts increases by approximately 6% to 7.2% when the nearest mine is in operation. The Saber 11 test score exhibits an improvement of 3.87 points (8.2% compared to the control mean) when the closest mine is operational, with sustained positive effects up to 9 years after the mine's opening. Moreover, positive outcomes are observed in the probability of enrolling in higher education, showing a substantial increase of 12.2% when the nearest mine is operational. Notably, this increase in cohort size does not compromise the quality of education, as evidenced by the consistent Saber 11 scores. However, the study also identifies a potential downside. The results indicate that if the mine closest to schools is in operation, the probability of securing formal employment decreases between 0.2% and 8.6%.

Finally, Individually, these three chapters provide context about and insight into three different factors that impact educational outcomes in Colombia. Collectively, they provide essential information for policymakers on articulating the system better, using the pre-established conditions of communities to their advantage, and transforming the country's human capital between secondary school and the labor market more efficiently.

Abstract Chapter 2

EN

The System for the Prevention and Analysis of School Dropouts in Higher Education (SPADIES) played a pivotal role in addressing the challenge of reducing drop-out rates in Colombian higher education, contributing to an improved enrollment rate from 20% in 2002 to 40% in 2010 and 53.9% in 2022. This software dashboard facilitated student data collection, analysis, and visualization, empowering higher education institutions to prevent dropouts and re-engage those who had already dropped out. Using a differences-in-differences approach within the framework proposed by Callaway and Sant'anna (2021), this study reveals that SPADIES reduced the probability of students becoming dropouts by 0.7 percentage points and increased the likelihood of graduating and graduating on-time by 0.6 and 0.4 percentage points, respectively. Although the impact may appear small, it reaches 2.1 percent after 5 years of the program's implementation (which compares to the control group mean of 47%), saving at least 14,000 students from dropping, nearly two times the size of the average higher education institution in Colombia. Furthermore, SPADIES contributed to increased future income by assisting students in obtaining their degrees, alleviating congestion in the higher education system, reducing the burden on institutions, and enhancing enrollment efficiency. These findings highlight the significant positive effects of a data-driven software dashboard on the quality and efficiency of higher education systems in developing countries.

DE

Das System zur Prävention und Analyse von Studienabbrüchen (SPADIES) spielte eine entscheidende Rolle bei der Verringerung der Abbrecherquote im kolumbianischen Hochschulwesen und trug dazu bei, die Einschreibungsquote von 20 % im Jahr 2002 auf 40 % im Jahr 2010 und 53,9 % im Jahr 2022 zu steigern. Das Software-Dashboard erleichterte die Erfassung, Analyse und Visualisierung von Studierendendaten und ermöglichte es den Hochschulen, Studienabbrüchen vorzubeugen und Studienabbrecher wieder zu integrieren.

Unter Verwendung eines Differenzen-in-Differenzen-Ansatzes innerhalb des von Callaway und Sant'anna (2021) vorgeschlagenen Rahmens zeigt diese Studie, dass SPADIES die Wahrscheinlichkeit, dass Studierende ihr Studium abbrechen, um 0,7 Prozentpunkte verringert und die Wahrscheinlichkeit, dass sie ihr Studium abschließen, um 0,6 bzw. 0,4 Prozentpunkte erhöht hat. Obwohl die Auswirkungen gering erscheinen mögen, erreichen sie nach fünf Jahren Programmdurchführung 2,1 Prozent (im Vergleich zum Durchschnitt der Kontrollgruppe von 47 Prozent), wodurch mindestens 14.000 Studierende vor dem Studienabbruch bewahrt wurden, was fast der doppelten Größe einer durchschnittlichen Hochschule in Kolumbien entspricht. Darüber hinaus trug SPADIES dazu bei, das zukünftige Einkommen der Studierenden zu erhöhen, indem es ihnen half, ihren Abschluss zu machen, die Überlastung des Hochschulsystems zu verringern, die Belastung der Institutionen zu reduzieren und die Effizienz der Einschreibungen zu erhöhen. Diese Ergebnisse verdeutlichen die signifikant positiven Auswirkungen eines datengestützten Software-Dashboards auf die Qualität und Effizienz von Hochschulsystemen in Entwicklungsländern.

Chapter 2

How a Data-driven Tool Ended the Musical Chair Game in Higher Education in Colombia

2.1 Introduction

The dropout phenomenon in higher education presents a substantial economic challenge, impacting the cost of training a qualified workforce and hindering productivity and efficiency in the labor force. To address this issue, countries globally are exploring innovative solutions to increase enrollment and retain students in school, particularly in higher education. Colombia is an interesting case since its enrollment rate increased from 20% in 2002 to 40% in 2010 (Ferreyra et al., 2017; Orozco Silva, 2010) .

Between 2002 and 2010, Colombia's higher education system faced challenges in meeting the increasing demand for higher education. These challenges stemmed from demographic growth and a policy that promoted secondary graduation without sufficiently considering academic merit. Consequently, the system encountered the "Crowding cohort"[1] phenomenon,

[1] "Crowding cohort" is a phenomenon when there is increased demand for fixed or reducing resources. This term was used for the first time in Bound and Turner (2007).

compounded by an economic depression, which put many students at risk of dropping out (Herrera-Prada, 2013; ICFES, 2002; Orozco Silva, 2010).

To address the challenge, the Ministry of Education (MEN) implemented a plan to improve higher education outcomes. This plan utilized existing infrastructure, provided targeted financial support, and established information systems to track students and evaluate quality. SPADIES (System for the Prevention and Analysis of School Dropout in Higher Education) was created as part of this effort. It provided real-time data on enrollment, academic performance, peer quality, dropout rates, and graduation rates to higher education institutions (HEIs), the MEN, and the public. HEIs used SPADIES to identify at-risk students and offer various forms of assistance, including academic support, financial aid, and mental health services, to reduce dropout rates (Ministerio Nacional de Educacion, 2008; Guzmán Ruiz et al., 2009; Ministerio de Educación Nacional, 2006).

This chapter evaluates the impact of SPADIES on education outcomes and system efficiency in Colombia. Using SPADIES data from 1998 to 2017 and employing a differences-in-differences approach under the framework proposed by Callaway and Sant'Anna (2021), the study examines the college paths of approximately 4 million students enrolled in HEIs with SPADIES.

The findings indicate that SPADIES reduced the dropout rate in average by 70 basis points -bps- (up to 2.1 percent (210 bps) after 5 years of SPADIES installed or 4.4% of the control mean) and increased the probability of earning a degree on time in average by 60 bps (up to 2 percent (200 bps after 5 years of SPADIES installed or 7.3% of the control mean). Although these effects may appear modest, they have significant implications, with approximately 14,000 students saved from dropout, nearly double the average size of a higher education institution in Colombia. Moreover, SPADIES was particularly effective in reducing the dropout rate among males, students from public institutions, and low-income students. Additionally, HEIs utilized SPADIES not only to prevent dropouts but also to re-engage students who had already dropped out, as evident from increased transitions from absent and dropout status. With a total cost of USD 4.5 million from 2005 to 2017 (in prices of 2022), the cost per saved-from-dropout student through SPADIES was approximately USD 320.5, or the cost per graduated student was around USD 373.9. These figures indicate that tools like SPADIES could be a cost-effective tool for improving human capital formation in developing countries. By investing in SPADIES, the education system was able to prevent dropouts and increase graduation rates at a reasonable cost, contributing to the overall development and advancement of the country. Furthermore, the results suggest that SPADIES enhanced the efficiency of the higher education system by managing the entry of a growing student population. SPADIES also facilitated the transition to a digital record-keeping system. These findings highlight the potential of data-driven tools to improve educational outcomes in developing countries.

The following section presents a review of the literature. Section 2.3 details the SPADIES program design and the context in which it was created. Section 2.4 describes the dataset and variables. Section 2.5 presents the model specification. Section 2.6 discusses the results. Finally, Section 2.7 presents the conclusions and policy implications.

2.2 Literature Review

This section presents a comprehensive literature review that underpins the foundation of this chapter. Initially, the focus will be on the literature concerning the demand and supply sides of higher education, their impact on quality, and the evolution of dropout analysis. Subsequently, the literature specific to the Colombian context will be examined. Additionally, the influence of technology on education and the analytical approaches identified in the literature will be discussed. Finally, the contribution of this study to the existing literature will be explored.

In 1999, the World Bank initiated the "Education for All" program, aiming to improve the monitoring of educational indicators such as access, enrollment, and quality. The objective was to apply the lessons learned from developed countries' advancements in higher education systems to developing regions like Latin America and encourage governments to track outcomes and indicators.

Canonical literature acknowledges the vital role of formal education, particularly higher education, in human capital formation and social mobility (Adams and VanderWaerdt, 1984; Bank et al., 1990; Becker, 1962; ONU, 2013; Trow, 1973). However, the dynamics of supply and demand in higher education give rise to an imbalanced market characterized by access barriers, low enrollment rates, excess demand or supply, and low quality (Epple et al., 2006). Factors such as unexpected increases in applicants leading to insufficient public resources contribute to this discrepancy (Bound et al., 2009). Cost-benefit analyses incorporate institutional factors influencing students' decisions to enroll, persist, or drop out, such as the quality of education, support services, and peer influences (Bank et al., 1990; MacLeod and Urquiola, 2015; Tinto, 1975, 1982). The theoretical discourse on attrition commonly falls into two categories: the student's integration or adaptation to the education system model (Tinto, 1975, 1982) and attrition as a set of conditions linked to individual socioeconomic factors, such as family circumstances or academic performance during school (Bean, 1980, 1985).

In the context of Colombia, initial studies on attrition in higher education focused on selected institutions. For example, the Universidad Nacional de Colombia (2007) examined lag, graduation rates, and dropout rates in Colombia's largest public university, finding that being a female, particularly 18 years old or younger, positively influenced the probability of obtaining a degree in any program. Financial aid or student loan programs were found to decrease the dropout rate. Affirmative Action programs, including unique admission mechanisms and alternative admission routes through pre-university courses, played a crucial role in improving students' retention in higher education and enhancing access conditions and social equity (Sánchez Torres, 2002). Castaño et al. (2006) conducted a study at the University of Antioquia, focusing on the School of Engineering and the School of Economics.

Their findings revealed that being male, single, and over 18 years old increased the risk of dropping out. On the other hand, living with parents, achieving better academic performance, not being employed, having parents with a high level of education, and being female were associated with a decreased risk of dropout. Public universities in Colombia utilized these studies to address government concerns about resource allocation in the education system, viewing the dropout phenomenon as a waste of economic resources, human capital, and infrastructure (Cárdenas, 1996; Córtes et al., 2011; Facundo Díaz, 2009). National-level research conducted by ICFES (2002) identified household financial conditions as the primary determinant of becoming a dropout student. However, Ministerio Nacional de Educacion (2008) revealed that low academic skills (measured by the secondary school exit exam score), mismatch between career choice and skills, poor academic performance, and gender were the main reasons for the dropout rate. Finally, Herrera-Prada (2013) demonstrated that the "Crowding cohort" phenomenon, accompanied by an increase in the average time required to graduate, resulted in an overall decrease in the graduation rate in Colombia, despite a reduction in the nationwide dropout rate.

The impact of technology on education has been widely studied in various contexts, primarily focused on improving learning, teaching, research, and administrative systems (Tongkaw, 2013). There have been numerous studies on the influence of technology on new models of online learning (López-Pérez et al., 2011), motivation, and learning strategies (Sailer et al., 2021; Valentín et al., 2013), as well as on the use of information systems to improve learning (Leong and Ibrahim, 2015; McGill and Klobas, 2009; Sari, 2014). Additionally, there is research on the expansion of the educational system (Rahman, 2020; Tongkaw, 2013), the adoption of technology in learning environments (Lacka and Wong, 2021), and improving school attendance to enhance academic performance (Gomis-Porqueras et al., 2011). However, direct evidence of the use of technology at the administrative level for an entire country's higher education institutions, which can affect the coverage and quality of the higher education system, is lacking. The literature about this is quite limited, with the international literature focusing on small programs in regions of Germany and Chile and using econometric models (uplift models) to predict dropout rates (Berens et al., 2021) or design tailored anti-dropout programs (Olaya et al., 2020). However, survival, tailored, and uplift models have been employed in HEIs in Colombia to reduce the dropout rate since 2006, with no academic references other than the MEN reports on the SPADIES experiences.

This chapter adds to the literature by revealing how SPADIES effectively achieved its policy goals. My analysis demonstrates how a data system that improved the flow of information among agents, coordinated policy and innovative policymakers developed tailored aid programs for pre-selected candidates, resulting in a new equilibrium with educational and social outcomes even better than the government had anticipated.

2.3 Context and Program Design

In this section, I present facts and statistics about the Colombian context and the history behind the creation of SPADIES.

2.3.1 Context

In the past five decades, Colombia has experienced significant demographic changes, characterized by extensive migration from rural to urban areas. This influx of young individuals, driven by the pursuit of social and economic mobility, encountered limited access to quality educational resources (Lucio and Serrano, 1992).

By the mid-1990s, the insufficient capacity of high schools became apparent, leading the government to promote students through primary and secondary education without academic restrictions. Although this approach increased secondary school coverage and completion rates, it resulted in a significant decline in educational quality and skills (Herrera-Prada, 2013; Orozco Silva, 2010; Orozco Silva et al., 2006).

The economic crisis of the late 1990s further exacerbated this situation, revealing that only a few Colombians had access to higher education, and those who did face a high risk of dropping out (Ministerio Nacional de Educacion, 2008; Orozco Silva, 2010; Orozco Silva et al., 2006). Studies conducted in the early 2000s indicated that household economic conditions were the primary barrier to higher education enrollment (ICFES, 2002). During this period, Colombia had a gross college enrollment rate of merely 20%, one of the lowest in the region (Ferreyra et al., 2017; Ministerio de Educación Nacional, 2006).

In response, the Ministry of Education (MEN) implemented the "Educational Revolution" plan to boost educational attainment levels in higher education, which had lagged behind regional counterparts. The plan increased the enrollment rates from 20% in 2002 to over 40% in 2010 and surpassing 45% in 2016 (Orozco Silva et al., 2006; Ministerio de Educación Nacional, 2006, 2017). Public higher education institutions were primarily responsible for this surge, as evidenced by the near-zero or negative enrollment growth rates at private institutions in the early 2000s.

With the introduction of SPADIES by the MEN, timely information on enrollment, academic performance, peer quality, dropout rates, and graduation rates in the higher education market became available. The primary objective of SPADIES was to reduce dropout rates by identifying at-risk students and providing targeted aid suggestions to higher education institutions (HEIs). Additionally, SPADIES facilitated data collection on the higher education system and HEI performance indicators, which were publicly accessible. SPADIES became the standard for certifying program quality and measuring improvements in access to resources in Colombia's higher education institutions.

2.3.2 Program Design

Developed by the MEN since 2005, SPADIES is an information system that collects data on higher education students. It provides data visualizations, statistical analysis, and reports on at-risk students. The system has a module that is publicly accessible, allowing stakeholders to compare the performance of HEIs at different levels.

To improve data quality, each institution was required to have regular contact with the MEN, and data underwent monthly audits before being stored in the database. Duration Models and focus group analyses were conducted nationwide using the SPADIES database to identify patterns among dropout students. The results were integrated into the application to assist HEIs in identifying and supporting at-risk students.

The MEN visited each HEI to install the software, provide training, and explain the results. Aid programs were categorized into financial, academic, and other types of support. Males with low household income and academic skills, particularly in associate or math-related programs, were identified as the most at-risk population, and HEIs were encouraged to design aid programs mainly targeting this group.

Follow-up visits were conducted 12 to 18 months after installation to verify outcomes and ensure proper use of the application. By 2013, all institutions had installed the software and received follow-up visits. The installation and training of SPADIES were not randomly assigned, and variations in the digital gap and information quality among HEIs existed. Five rounds (list in Figure 2.1) were conducted to collect information from each institution, and the application of SPADIES was not differentiated based on student characteristics or HEI attributes (Figure 2.2). In 2008, the official dropout rate measured by SPADIES was mandated to be included in program quality certificates, and a web portal was established for public access to key statistics on individual HEIs and the higher education system.

Finally, the total development and operational cost of SPADIES from 2005 to 2017 was approximately USD 4.5 million, adjusted for inflation to 2022 prices.

2.4 Data and Variables

This section provides an overview of the SPADIES database and its characteristics, followed by a description of the variables used in the program's definitions, including the new variables created for this study. Finally, the section explains the final database used in the empirical analysis.

2.4.1 SPADIES Database

The SPADIES database merges data from the MEN's SNIES database, the ICFES database, and HEIs' semestral reports. SNIES data provides HEI and program characteristics, while ICFES data captures the Saber 11 exam information. HEIs' reports are updated every semester. SPADIES focuses on students who started college from 1998 onwards. This study utilizes SPADIES data from 1998 to 2017, including around 8 million students. The dataset consists of an unbalanced panel per individual-program semester.

2.4.2 Variables

This subsection explains the variables that SPADIES has, how they are measured, and the new variables I created.

SPADIES Time-Invariant Variables

The ICFES data obtained from the Saber 11 exam includes variables such as exam score, gender, birth year, and household income. The exam score is mandatory for higher education enrollment, ensuring all enrolled students have a score. At the time of test application, students complete a form for characterizing and analyzing data in this study.

In order to standardize the Saber 11 test scores, since the ICFES has employed different score ranges over time, each student's percentile on the exam is assigned, resulting in a variable that ranges from 1 to 100. Additionally, a dummy variable is created to indicate whether a student has achieved a high score (above 90), based on the methodology of the Ministerio de Educación Nacional (2008; 2006; 2017).

Gender and birth year data are available in both the ICFES database and the Freshmen report, while household income is reported to the ICFES by students. In case of discrepancies, SPADIES gives priority to the ICFES data, unless it is missing.

For HEI and program characteristics, SPADIES utilizes data from the SNIES database. HEI characteristics include sector (public or private), category (university -for 4 year programs- or community college -for 2 or 3 year programs-), and location. Program characteristics encompass the level (bachelor's - 4 or 5-year programs - or associate - 2 or 3-year programs) and field of knowledge.

SPADIES also evaluates the quality of data reported by each HEI, grading it as A, B, or C based on the number of reported semesters and the level of detail provided. A dummy variable indicates whether an HEI received an A grade in the 2017 report, as recorded in the database.

SPADIES Time-Variant Variables

SPADIES receives three main reports per semester from each higher education institution (HEI): Freshmen, Graduates, and Enrolled. These reports contain essential information about students, such as their student ID, academic performance, details of financial or academic aid received, and the program of study they are enrolled in.

To track students' progress, SPADIES utilizes the Freshmen and Enrolled reports. It identifies students who enroll in a program and continues to track them until they are reported as Graduates. If a student is not found in the Enrolled report or the Graduated report, their status is updated to reflect their absence or dropout, depending on the number of semesters they have not been enrolled.

SPADIES defines the following statuses for students:

1. **Graduated:** This category includes all individuals who have successfully completed their higher education program. It is further divided into two subcategories:

   (a) Graduated on time (those who graduated within one year of the expected time of graduation).

   (b) Graduated late (those who graduated more than 1 year after their expected time of graduation).

2. **Dropout:** SPADIES categorizes students as dropouts if they have not been reported in the system or if they have not graduated after two or more consecutive semesters, as of 2017.

3. **Absent** students are those who missed only one semester and are not reported as Graduated.

4. **Active** is any student taking classes as of 2017.

Additionally, I have introduced three new variables to measure transitions between statuses. By comparing the time since students first enrolled in college and the number of semesters reported in SPADIES, I can identify if a student has left or returned to school. The two types of transitions are as follows:

1. **Transition from absent,** means that in a period “T” the student was “Absent” and in “T+1” he became “Active” or “Graduated”. The dummy variable following these transitions takes the value of 1 in “T+1” and 0 otherwise.

2. **Transition from dropout**, means that in a period “T” the student was a “Dropout” and in “T>=3” he became “Active” or “Graduated”. The dummy variable following these transitions takes the value of 1 when the transition ends and 0 otherwise.

Finally, I created the variable **Time gap of transition** that counts the amount of semesters during each transition; in the case of transitions from “Absent", it is always 1, but for the transitions from dropout, it is always 2 or more semesters.

2.4.3 Final Data

SPADIES combines time-invariant data from the ICFES database and SNIES with semesterly reported Freshmen, Enrolled, and Graduates data. The individual refers to a student who has ever been enrolled as a freshman in an HEI program between 1998 and 2012.

To mitigate bias in dropout and graduation rates, two changes were made to the original database of 8 million students:

1. Only data for students reported as "active" since 2002 were included. This means they were reported as a freshman before 2002 and appeared in the 2002 enrollment report.

2. Data was limited to students who were freshmen up to 2012 to focus on cohorts with sufficient time to graduate. Data beyond 2017 was not considered for these students.

After these adjustments, the database comprised a population of 6,143,537 students.

A dummy variable was created to identify the time of SPADIES installation in each HEI. The semester of installation marked the start of SPADIES' "treatment" for an HEI, as it provided the first report on at-risk students and allowed the institution to access grants for anti-dropout programs from the MEN. The dataset also includes the unemployment rate estimated by DANE by HEI department and year.

The final dataset is an unbalanced panel per individual-program-HEI and time, consisting of 4,131,302 students. It includes variables such as gender, year of birth, household income, ICFES test score, a dummy for bachelor-level programs, dummies indicating assistance program receipt, public institution dummy, HEI certification dummy, main campus dummy, data quality dummy, region dummies (Bogotá, Valle del Cauca, Antioquia, Atlántico), round of implementation dummies, enrollment period dummy, unemployment rate by department and year, status in the system, transition variables (absent, dropout, time gap), dropout dummy, graduation dummy, and the number of transitions during the program.

2.5 Model Specification

To measure the impact of SPADIES in different outcomes, I use the following equation:

$Y = \beta_0 + \gamma_0 \ SPADIES_{it} + \gamma_1 \ SPADIES_{it} \times AA_{it} + \gamma_2 \ SPADIES_{it} \times \ PFA_{it} + \gamma_3 \ SPADIES_{it} \times \ AA_{it} + \beta \ X_{it} + \epsilon_{it}$ (2.1)

where Y_{it} is a dummy that, depending on the model, measures one of five outcomes that are probabilities: the probability of dropping out, graduation, graduation on time, transitioning from absent, or transitioning from dropout. In the case of Time gap (the sixth outcome), Y_{it} is a continuous numerical variable equal to the number of semesters of the transition. The variable of interest is $SPADIES_{it}$, it takes the value of one (1) if the individual is enrolled in a HEI in a period t when SPADIES was already installed and zero (0) otherwise. I also include three interactions with a variable that takes the value of one (1) if student received aid in a specific time and zero (0) otherwise: SPADIES and academic aid (AA), SPADIES and public financial aid (PFA), and SPADIES and private financial aid (FA). I include these interactions to understand if the combination of SPADIES and the academic or financial aid improved education outcomes. The vector of controls X_{it} is comprised of time variant and time invariant variables. The time variant variables include academic performance, occurrence of assistance if received and type (financial or academic), time that the student has been enrolled in the HEI (tenure), and departmental unemployment rate. The time invariant variables include a dummy for females, the year of birth, a dummy if the Saber 11 exam score is over the 90th percentile, a categorical variable for household income, and a set of dummies to indicate the region of the HEI that the student attends.

I will present two sets of results for Equation 2.1 using a Fixed Effects (FE) and Random Effects (RE) framework. Both models offer different advantages. The RE model enables comparisons with previous literature ICFES (2002); Ministerio Nacional de Educacion (2008) and Ministerio de Educación Nacional (2006), the FE model provides more consistent results to use as benchmark before the Differences in Differences (DiD) approach.

To properly identify the causal effect of SPADIES on the six outcomes, recent literature (Sant'Anna and Zhao, 2020; Callaway and Sant'Anna, 2021; Goodman-Bacon, 2021; Sun and Abraham, 2021) provides relevant tools. Previously, it was challenging to isolate the impact of SPADIES due to its five rounds, non-random assignment, and potential population differences across rounds. Treating each round as a separate treatment that overlapped in time posed a significant problem. I employ the framework proposed by Callaway and Sant'Anna (2021) to estimate causal inference. This DiD approach allows identification, estimation, and inference for multiple time periods, considering up to 6 semesters after the treatment was applied. It also accounts for variation in treatment timing (rounds) and potential differences in treatments, while holding the "parallel trends assumption" (PTA) after conditioning on observed covariates in the pre-treatment period.

2.5.1 Canonical DiD

The basic DiD approach in the canonical format considers two periods and two groups (model 2X2). In the first period (T=0), the two groups are the same in terms of the treatment, as they do not receive any. In the second period (T=1), some of the individuals did receive the treatment creating the group called “treated" (SPADIES=D=1), while those that did not receive any treatment are called “controls" (SPADIES=D=0). So, if we assume that the treated group would follow its predetermined path given by its trend, in case of an absence of the treatment, any deviation from this trend is a causal effect of the treatment on the group. This deviation or difference is the Average Treatment effect on the Treated (ATT) (Equation 2.2). However, the $Y_{i,1}(0)|D_i = 1$ component is never observed, as it is unknown how the treated group would be in T=1 in the absence of treatment.

$$ATT = E(\tau_i|D_i) = E(Y_{i,1}|D_i = 1) - E(Y_{i,1}(0)|D_i = 1) \qquad (2.2)$$

As I do not know the path the treated group will follow in the absence of treatment, my best approach is to check the path of the control group. I assume that the path followed by the treated groups is parallel to the path followed by the control group. This assumption is known as the “Parallel Trend Assumption" (PTA) (Equation 2.3). This assumption is very strong and will be debated later, but by using it, we can re-estimate the ATT (Equation 2.4).

$$E(Y_{i,1}(0) - Y_{i,0}|D_i = 1) = E(Y_{i,1} - Y_{i,0}|D_i = 0) \qquad (2.3)$$

$$\widehat{ATT} = E(Y_{i,1}|D_i = 1) - \widehat{E}(Y_{i,1}(0)|D_i = 1) \qquad (2.4)$$

However, in practice, the empirical research usually faces designs with more than two periods or more than two treated groups. According with Callaway and Sant'Anna (2021), the solution has been to generalize the canonical approach by adding the groups and fixed effects to the specification. The debate about the correct specification has been growing in recent years, but the literature agrees that the standard Two-Way Fixed Effects (TWFE) approach may not be appropriate for the identification of treatment effects, in particular interpreting its results (Callaway and Sant'Anna, 2021) . As mentioned above, the PTA is hard to achieve, as treated and control groups are often not similar enough. To solve this, Sant'Anna and Zhao (2020) proposed to hold PTA for groups with the same pre-treatment characteristics X (Equation 2.5). Where $\theta(X)$ is the ΔY_i if there was no treatment conditional to X. With this new assumption, the new DiD estimator becomes $\widehat{ATT_*}$ (Equation 2.6).

$$E(Y_{i,1}(0) - Y_{i,0}|D_i = 1, X) = E(Y_{i,1} - Y_{i,0}|D_i = 0, X) = \theta(X) \qquad (2.5)$$

$$\widehat{ATT_*} = E(Y_{i,1}|D_i = 1) - \Big[E(Y_{i,0}|D_i = 1) + \widehat{E}(\theta(X)|D_i = 1)\Big] \qquad (2.6)$$

2.5.2 Robust DiD Estimators for ATT

Now, using Rios-Avila et al. (2021)'s CSDID command in Stata, four types of DiD estimators are analyzed using Equation 2.6 as they present four different approaches to estimate the component $\widehat{E}(\theta(X)|D_i = 1)$:

1. Regression Approach (OR). This approach estimates $E(\theta_i|D_i = 1)$ in two steps. The first step models $E(\theta_i|X) = \theta(X)$ as a function of X with data from the control group only. The second step uses the predicted outcome for $\widehat{\theta}(x_i)$to estimate $E(\theta_i|D_i = 1)$. The ATT for this estimator is:$\widehat{ATT_{OR}} = E(\Delta Y_i|D_i = 1) - E(\widehat{\theta}(x_i)|D_i = 1)$ (2.7)

2. Inverse Probability Weights (IPW) from Abadie (2005). In this method, the distribution of characteristics X for the control group is reorganized, so that the control group becomes more similar to the treated group. To do so, it estimates a propensity score using a binomial model and then, using the predicted scores, estimates the inverse probability weights $\omega\,(x)$. The dependent variable in the propensity score is a marker for if the observation is part of the treatment group as a function of X.

   $P(D_i = 1|X) = F(X) \to \widehat{\pi}(X) = \hat{F}(X)$

   $\omega(x_i) = (\widehat{\pi}(x_i))/(1 - \widehat{\pi}(x_i)) \Rightarrow \widehat{E}(\theta_i|D_i = 1) = (E(\omega(x_i)\theta_i|D_i = 0))/E(D_i)$ (2.8)

   $\widehat{ATT_{IPW}} = E(\Delta Y_i|D_i = 1) - (E(\omega(x_i)\theta_i|D_i = 0))/E(D_i)$ (2.9)

3. Doubly Robust Estimator (DRI) from Sant'Anna and Zhao (2020). The doubly robust estimators are a combination of the previous two estimators (OR and IPW). The model first uses the regression approach, and then it reshapes the groups using a propensity score estimation similar to the IPW approach. A propensity score is estimated using Equation 2.8, then $E(\theta_i|X)$ is modeled as a function of X and estimated using the weights obtained from Equation 2.8. See Equation 2.10.

   $\theta_\omega(X) = Min \sum_{i|D_i=0} \omega_i(x_i)(\theta_i - \theta(X_i))^2$ (2.10)

   $\widehat{ATT_{DRI}} = E(\Delta Y_i|D_i = 1) - E(\widehat{\theta_\omega}(x_i)|D_i = 1)$ (2.11)

4. Improved Doubly Robust Estimator (IMP) from Sant'Anna and Zhao (2020). This estimator uses in the first step an approach similar to OR by estimating $E(\theta(X)|D_i = 1)$ using only control data a no weights. Then, it adds a correction Λ, calculating the weighted difference between the predicted and the observed outcome in the control group. See Equation 2.12.

$$\widehat{ATT_{IPW}} = E(\Delta Y_i|D_i = 1) - E(\widehat{\theta_\omega}(x_i)|D_i = 1) - \lambda \qquad (2.12)$$

$$where\ \lambda = E(\omega(x_i)\Delta Y_i|D_i = 0)/E(\omega(x_i)|D_i = 0) - E(\omega(x_i)\widehat{\theta}(x_i)|D_i = 0)/E(\omega(x_i)|D_i = 0)$$

2.5.3 Empirical Framework

Callaway and Sant'Anna (2021) expanded what was proposed previously by Sant'Anna and Zhao (2020) and Abadie (2005). In particular, Callaway and Sant'Anna (2021) debated the application of DiD estimators when a variation in the timing of treatment existing, and they consider a natural generalization of the ATT to be a setup with multiple treatment groups and time periods. Callaway and Sant'Anna (2021) used the average treatment effect for units who are members of a particular group g at a particular time period t, that expressed in terms of the canonical form (Equation 2.2) is:

$$ATT(g,t) = E[Y_t(g) - Y_t(0)|G_g = 1] \qquad (2.13)$$

The framework of Callaway and Sant'Anna (2021) incorporated OR, IPW DRI, and IPW, fixes a group g and allow variation in t, to understand how the proposed ATT evolves in time for a specific group. When this process is extended to all groups, they present the "group-time average treatment effect". In fact, the estimation performed by Rios-Avila et al. (2021) disaggregated the combinations of groups and times in multiple 2X2 models than then are aggregated per the fixed group g. After the process, an ATT and weights per period group allow consolidation of the ATT not only by group-time, but also by time (similar to an event analysis), by group (to analyze impacts per group and compare) and using a single robust consolidated estimator.

2.6 Results

The results section contains three parts: The first part presents the "Parallel Trend Assumption" (PTA) charts per outcome. I find similar trends in all the periods before SPADIES was installed for the first 4 Rounds; although Callaway and Sant'Anna (2021) assumptions only require PTA stability in the pre-treatment period. The second part shows the event analysis for the four proposed estimates in the left panel and the total and group ATT per output in the right panel. For the main findings, I will focus my reading on the outcomes obtained by the IMP methodology (all the results are shown), which is the most robust and has the most conservative results. Finally, in the third part, I present the total ATT per sub-sample (according to time-invariant characteristics) per estimator type. The students in Round 5 account for only 0.3% of the entire system, they are not included in the PTA nor in the results analyses.

2.6.1 Regression Analysis

Table 2.3 reports the results for Equation 2.1, focusing on the variable of interest, SPADIES. It reveals a significant impact of SPADIES in reducing the probability of dropping out and increasing the probability of graduating and graduating on time. These results hold in both the fixed effects (FE) and random effects (RE) frameworks, with the FE framework showing greater impacts. It is important to note that these results are preliminary and do not establish causality. Causal results will be presented later in this section.

However, these initial findings provide valuable insights into the determinants of dropout probability, graduation probability, and transitions. They align with previous literature

(ICFES, 2002; Ministerio Nacional de Educacion, 2008; Ministerio de Educación Nacional, 2006; SPADIES, 2008) and suggest the following expected results:

1. An increase in program tenure and receiving academic or financial aid is associated with a decrease in the probability of dropping out or an increase in the probability of graduating and graduating on time.

2. An increase in the share of failed classes is associated with an increase in the probability of dropping out or a reduction in the probability of graduating and graduating on time.

Economic theory and literature suggest that a high unemployment rate may lead to a lower probability of dropping out, as students prefer to stay in school while waiting for favorable labor market conditions. However, in countries like Colombia, where new college attendees come from low-income households, a higher unemployment rate may also indicate a negative impact on household income, prompting some students to withdraw from school to support their families.

Regarding program tenure, an additional semester in the system is associated with a decreased probability of dropping out and an increased probability of graduation and graduating on time. Conversely, an increase in the share of failed classes in the last semester is linked to an increased probability of dropping out and a decreased probability of graduating and graduating on time, consistent with previous research (Herrera-Prada, 2013; Ministerio Nacional de Educacion, 2008; Ministerio de Educación Nacional, 2006).

Receiving tutoring and mentoring has a positive effect in reducing the probability of dropping out. However, the interaction of tutoring and mentoring with SPADIES increases the probability of dropping out. This can be attributed to the fact that SPADIES targeted students with high academic vulnerabilities who were already at a higher risk of dropping out, and the assistance they received was insufficient to prevent dropout. Interactions of SPADIES with academic aid and public financial aid indicate an increase in transitions from absent and dropout status. In contrast, the interaction of SPADIES with private financial aid shows a reduction in transitions from absent status. This suggests that financial aid provided by the HEI effectively helps prevent students from leaving school due to financial constraints. However, the results suggest that both financial and academic aids, while well-targeted, were not sufficient. The positive sign for the transition from dropouts receiving public financial aid suggests that these students could eventually obtain their degrees. However, the increase in the transition gap and the positive signs in the interactions with academic aid and public financial aid indicate that SPADIES facilitated the reengagement of students who had dropped out, incentivizing them to return to school.

Consistent with previous studies (ICFES, 2002; Ministerio Nacional de Educacion, 2008; SPADIES, 2008; Universidad Nacional de Colombia, 2007), females exhibit a lower probability of dropping out and a higher probability of graduating on time compared to males. Younger students have a lower probability of dropping out but also lower probabilities of graduating or graduating on time. This is because younger students have fewer transitions per year, resulting in a larger transition gap compared to older students. Higher household income and higher scores on the secondary school exit exam are associated with a lower probability of being a dropout student and an increased probability of graduating and graduating on time. Females, students with high scores on the secondary school exit exam, and high-income students have a lower probability of transitioning from absent or dropout status.

2.6.2 Parallel Trends

The assumption for unbiased estimation of the ATT in SPADIES treatment is parallel trends, which requires similar trajectory of dependent variables prior to treatment in treated and control groups. Figure 2.4 presents average outcomes in the six semesters before SPADIES treatment. Due to dynamic nature of indicators and their construction depending on database cutoff, estimating parallel trends using the final database is challenging. To address this, I created a new database including students from all phases of their programs who had no contact with SPADIES, allowing a time horizon of up to 5 years before SPADIES. These "never treated" students were assigned statuses using SPADIES definitions prior to its implementation. Placebo treatments, referred to as SPADIES-Placebo, were then assigned 1, 2, or 3 years prior to actual SPADIES implementation in each HEI. HEI assignment for each cohort was randomly done while maintaining the proportion of HEIs assigned to rounds 1 (27.3%) and 2 (19.6%) observed in reality. All HEIs were assigned, so the remaining 53.1% from the first two rounds were counted in Round 3 of the placebo. A total of 100 placebos were performed, and the SPADIES-Placebo effect was estimated using classical DiD model and Callaway and Sant'Anna (2021) methodology. Coefficients are presented in Figures 2.5 and 2.6. Results from the classical DiD model indicate null effect of SPADIES-Placebo for all variables of interest, suggesting parallel trends assumption is met. In the Callaway and Sant'Anna (2021) placebo exercise, there is evidence of a null placebo effect for drop-outs and graduation, while for transitions, there is a small positive effect.

2.6.3 Main Results

In this section, I present the causal results for SPADIES estimated using Equations 2.7, 2.9, 2.11 and 2.12 by using the command CSDID from Rios-Avila et al. (2021). Results will be presented by the outcome in two panels: the left panel with the event analysis figure, and the right panel will present the coefficients of interest for the total of the program per Round.

Drop-Out

Figure 2.7 demonstrates that SPADIES reduced dropout probability by 70 bps. The DiD estimators consistently showed a positive impact of SPADIES on dropout likelihood, resulting in an 80 bps reduction. Round 1 had the most significant effect, with dropout reductions of 0.90 to 1.20 percent (90-120 bps). Even after six semesters, SPADIES continued to decrease dropout probability by 2 to 8 percent (200-800 bps).

Graduation

Figure 2.8 shows that SPADIES increased graduation probability by at least 60 bps. Round 1 had the highest effectiveness, with graduation probability increases of 90-110 bps (0.9 to 1.1 percent). After six semesters, SPADIES led to graduation probability increases of 210-790 bps (2.1 to 7.9 percent).

On-time Graduation

Figure 2.9 indicates that SPADIES increased the probability of graduating on time by 40 bps. The DiD estimators confirmed the positive impact of SPADIES on timely graduation, particularly in Round 1. After six semesters, SPADIES increased the probability of graduating on time by 180 and 540 bps (1.8 to 5.4 percent).

Transition from Absent

Figure 2.10 shows that SPADIES increased the probability of transitioning from absent status by 30 bps. There were no significant differences in the probability of dropping out, graduating, or graduating on time across the different rounds. This may be due to the short gap period required for the transition from absent status, which is only one semester. The number of transitions for students marked as absent increased consistently across all rounds, with no notable variation between rounds. The impact was consistent across all semesters, indicating a comparable increase in transitions for students marked as absent after the installation of SPADIES.

Transition from Drop-out

In Figure 2.11, SPADIES demonstrated a positive impact on the probability of transitioning from drop-out status, resulting in a 20 bps increase. Round 4 exhibited the most favorable results among the rounds, possibly due to the shorter time since treatment and the HEIs' proactive approach of calling back drop-out students. Although other rounds implemented similar strategies, their effects diminished over time as the number of recalled students decreased and new drop-out students required two semesters to be marked and tracked. It is worth noting that once students are marked as absent, HEIs have the ability to bring them back, making it challenging for them to transition to drop-out status after the implementation of SPADIES. Therefore, the analysis conducted after SPADIES installation shows improvement in drop-out transitions as HEIs become more proficient in tracking and recalling drop-out students. While the round-specific results indicate a diminishing effect of SPADIES over time, the semester-specific results indicate that HEIs were able to enhance their efficiency in tracking and reintegrating drop-out students. However, this effect weakened over time, as evidenced by the results six semesters after SPADIES implementation.

Time Gap of Transition

The results in Figure 2.12 show that SPADIES increased the time gap during the transition by 0.6 semesters. This finding is closely linked to the transitions from dropouts, as the differential increase found in the results of Figure 2.11 can be attributed to the return of students who were previously marked as dropouts but brought back by the HEI with the help of SPADIES. These students had been outside the system for a significant period and would have remained so if not for using SPADIES. Therefore, the reported increase in the time gap is a positive outcome as it represents the return of students who may not have returned to school otherwise.

Students who had transitioned from dropout status before the implementation of SPADIES had already planned to return to school regardless of SPADIES; so it makes sense that the gap was shorter for them. On the other hand, the students who returned because of HEIs' targeting through SPADIES had been out of the system for many years, and the influx of these returning students led to the increase in the time gap. This gap is more significant in Round 4 because of the short time to incorporate new dropouts into the average of all transitions. Overall, the increase in the time gap during the transition reflects the successful re-entry of previously marked dropouts. The longer time gap is an indicator of the effectiveness of SPADIES in bringing these students back to the education system.

2.6.4 Disaggregated Results

The analysis of time-invariant variables revealed notable results, as shown in Figures 2.13 to 2.16. SPADIES proved to be particularly effective in reducing the probability of dropping out for males with low income attending public HEIs. The difference in results was significant when considering the sector of the HEI, with public institutions experiencing a more substantial reduction in dropout rates.

Examining the probability of graduating, males from low-income backgrounds in public HEIs showed an increased likelihood of obtaining their degree. However, a significant difference was observed between HEIs with and without quality certifications. Similar results were found for the probability of graduating on time, with no significant differences among subpopulations.

Analyzing the probability of transitioning from absent status, students from non-certified HEIs and those in associate programs displayed a higher likelihood of transitioning from absent status. Conversely, male students in associate programs at public HEIs were more likely to transition from dropout status.

Regarding the time gap during transitions, the most significant benefits from SPADIES were observed among males with low income attending public HEIs, particularly those in associate programs. This subset of students was also more likely to have already dropped out and may not have returned to school without the targeted intervention of SPADIES. These findings align with the Ministry of National Education's directive for HEIs to focus their aid programs on this specific subset of students.

2.7 Conclusions

The new millennium posed a significant challenge for Colombia's education and labor market as higher education students in the 2000s outnumbered their predecessors but were less prepared and had lower skills. This vulnerability led to higher dropout rates and a decline in the overall quality of the higher education system, affecting enrollment and graduation rates.

Despite limited economic resources, the Ministry of Education in Colombia (MEN) designed an action plan to address this education crisis. Various studies have emphasized the severity of the situation and the need for urgent intervention. As part of this plan, MEN introduced SPADIES, a software application that collects student data and helps institutions target at-risk students for dropout prevention. This initiative modernized protocols and records, providing real-time statistics to education authorities, institutions, students, employers, and the public.

The flow of information was crucial as HEIs received training on operating the dashboard and learned about successful strategies used by other institutions to reduce dropout rates. Students were aware of their school's standing and dropout rates, which were widely reported in the media, and this information influenced program and institution selection. Students in need became the main beneficiaries of new programs and aid. The analysis demonstrates that SPADIES directly impacted educational outcomes and addressed the poor quality of pre-SPADIES information.

SPADIES enabled the Colombian higher education system to achieve a more efficient and socially desirable equilibrium. Equilibrium models, such as Epple et al. (2006) and MacLeod and Urquiola, 2015 explain the path to this equilibrium. High-quality institutions used SPADIES to selectively accept students with lower dropout risk, while low-quality institutions utilized SPADIES as a promotional tool to attract more students. The increased information on HEI quality triggered competition, leading to more targeted aid programs and improved student tracking. The analysis highlights the increase in transitions and reduced barriers to graduation, particularly in low-quality institutions.

The findings reveal that SPADIES significantly improved student retention, on-time graduation, and overall graduation rates. It notably reduced dropout rates for all students, particularly males from low-income backgrounds in public institutions. show that SPADIES reduced the probability of students becoming dropouts on average by 70 basis points (bps), equivalent to saving about 14,000 students from dropping out of the system, and up to 2.1 percent (210 bps) after 5 years of SPADIES installed (4.4% of the control mean). This is an impressive figure, considering that the average size of a higher education institution (HEI) in Colombia is 8,000 students. SPADIES also increased the probability of students earning their degree on time on average by 40 bps and earning their degree overall by 60 bps (up to 1.4 and 1.9 percent after 5 years of SPADIES' installation or 6.2% and 7.3% of the control mean, respectively). SPADIES helped approximately 12,000 students earn their degrees, with 8,000 making them on time.

Furthermore, SPADIES increased the number of transitions from absent and dropout statuses, and the average duration of a transition. Non-certified HEIs effectively used SPADIES to bring back absent students, while certified public HEIs successfully targeted dropout students. The increased time gap during transitions indicated the successful re-entry of former dropout students who may not have returned without SPADIES.

Although many HEIs implemented aid programs based on SPADIES information, the analysis revealed that tutoring and mentoring, while effective in reducing dropout rates, had an adverse interaction with SPADIES for at-risk students. These students already faced high academic vulnerabilities, and SPADIES and aid support alone were insufficient to prevent them from dropping out.

Overall, SPADIES improved the efficiency of the higher education system by reducing the time to graduation and alleviating the burden of overpopulation on HEIs. It facilitated the selection process for students and allowed HEIs to address the challenges posed by the increasing student population entering higher education.

SPADIES also had two unintended spillovers: it helped dropout students return to school and brought the higher education system into the digital era. Many HEIs were still using paper records before the implementation of SPADIES. In summary, SPADIES was instrumental in breaking down the musical chairs game that was the higher education system in Colombia. SPADIES helped improve the efficiency and effectiveness of the higher education system in Colombia and changed the future of many students who would have otherwise dropped out. These findings suggest that the implementation of a data-driven software dashboard can have significant positive effects on the quality and efficiency of higher education systems in developing countries.

Table 2.1: Students Description

Variable	Obs.	Mean	Standard Deviation	Min	Max
Drop-out rate	4,131,302	.478	.5	0	1
Graduation rate	4,131,302	.275	.447	0	1
On-time graduation rate	4,131,302	.222	.416	0	1
Transitions	4,131,302	.055	.228	0	1
Transitions from absent	4,131,302	.029	.167	0	1
Transition from drop-out	4,131,302	.026	.159	0	1
Transitions time gap	4,131,302	.932	1.15	0	29
Tenure in program	4,131,302	4.97	3.42	1	35
Share of failed classes	4,131,302	.119	.244	0	1
Received tutoring or mentoring	4,131,302	.121	.327	0	1
Received financial aid	4,131,302	.254	.435	0	1
Female	4,131,302	.502	.5	0	1
Year of birth	4,131,302	1988	5.95	1960	1998
Secondary test score	4,131,302	61.8	28.4	1	100
Students with secondary test score >= 90	4,131,302	.202	.402	0	1
Household income	4,131,302	1.74	1.33	0	9
Unemployment rate	4,131,302	11	2.58	5.87	22.3

Source: ICFES-HEIs. The unemployment rate from DANE (National Statistical Office).

Table 2.2: Higher Education System Data Description

Variable	Obs.	Mean	Standard Deviation	Min	Max
Public institution	4,131,302	.421	.494	0	1
High quality institution	4,131,302	.308	.462	0	1
Main campus	4,131,302	.656	.475	0	1
Good data report	4,131,302	.964	.187	0	1
Institution located in Bogota	4,131,302	.415	.493	0	1
Institution located in Valle del Cauca	4,131,302	.065	.247	0	1
Institution located in Antioquia	4,131,302	.15	.357	0	1
Institution located in Atlantico	4,131,302	.056	.229	0	1
HEI from round 1 of implementation including sub-locations	4,131,302	.273	.445	0	1
HEI from round 2 of implementation including sub-locations	4,131,302	.196	.397	0	1
HEI from round 3 of implementation including sub-locations	4,131,302	.231	.421	0	1
HEI from round 4 of implementation including sub-locations	4,131,302	.297	.457	0	1
HEI from round 5 of implementation including sub-locations	4,131,302	.003	.058	0	1

Source: ICFES-HEIs.

Table 2.3: SPADIES Rounds Description

	Round					Total
	(1)	(2)	(3)	(4)	(5)	
Drop-out rate	.457	.482	.464	.505	.485	.478
Graduation rate	.313	.245	.274	.262	.187	.275
On-time graduation rate	.231	.2	.227	.226	.18	.222
Transitions = 1	.055	.047	.053	.062	.042	.055
Transitions from absent	.031	.027	.029	.029	.025	.029
Transition from drop-out	.024	.02	.023	.034	.017	.026
Transitions time gap	.976	.908	.92	.919	.78	.932
Received tutoring or mentoring	.085	.3	.087	.064	.079	.121
Any kind of Aid = 1	.228	.351	.244	.218	.451	.254
Female = 1	.474	.513	.519	.507	.452	.502
Year of birth	1988	1988	1987	1987	1988	1988
Secondary test score	73	60	60.8	53.5	56.5	61.8
Students with secondary test score >= 90	.362	.168	.177	.099	.119	.202
Household income	1.99	1.68	1.73	1.56	1.65	1.74
Unemployment rate	11.7	10.5	10.8	11	10.2	11
HEI public sector = 1	.547	.423	.34	.371	.198	.421
High quality institution = 1	.541	.314	.324	.08	.007	.308
Main institution (campus) = 1	.717	.617	.559	.702	.455	.656
Good data report = 1	.987	1	.99	.909	0	.964
Institution located in Bogota	.337	.43	.398	.493	.226	.415
Institution located in Valle del Cauca	.117	.006	.068	.055	0	.065
Institution located in Antioquia	.196	.181	.101	.123	.266	.15
Institution located in Atlantico	.082	.068	.015	.052	.229	.056
Observations	1,092,960	783,459	925,378	1,315,802	13,703	4,131,302
Share of total	26.46	18.96	22.40	31.85	.33	100

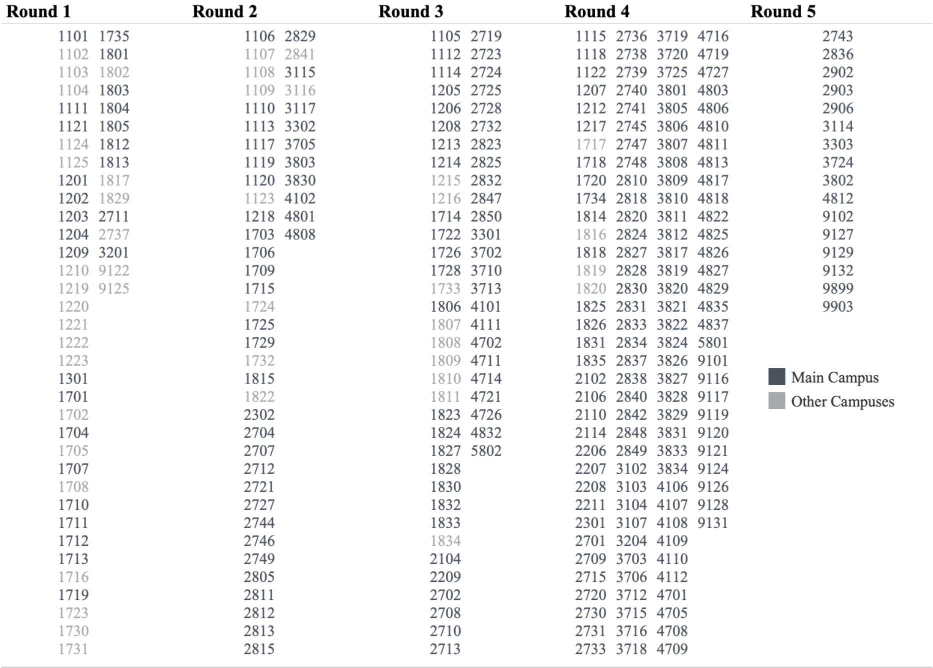
Source: ICFES-HEIs. The unemployment rate from DANE (National Statistical Office).

Table 2.4: Main Results Panel Regression Approach

	Fixed Effects						Random Effects					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Drop-out	Graduated	On-time graduation	Transition from absent	Transition from drop-out	Time gap during transition	Drop-out	Graduated	On-time graduation	Transition from absent	Transition from drop-out	Time gap during transition
SPADIES	-0.014***	0.007***	0.007***	-0.012***	-0.018***	-0.158***	-0.013***	0.001***	0.004***	-0.003***	-0.000***	-0.097***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Tenure in program	-0.000***	0.002***	0.001***	0.001***	0.003***	-0.029***	-0.002***	0.003***	0.002***	-0.001***	-0.001***	-0.020***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Share of failed classes	0.010***	-0.008***	-0.007***	0.007***	0.010***	-0.083***	0.013***	-0.011***	-0.009***	0.031***	0.025***	-0.012***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Received tutoring or mentoring	-0.023***	0.032***	0.017***	-0.010***	-0.008***	-0.019***	-0.021***	0.029***	0.016***	-0.004***	0.005***	-0.015***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.005)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.004)
Received financial aid	-0.002***	0.005***	0.002***	-0.005***	-0.005***	0.010***	-0.003***	0.005***	0.002***	-0.004***	-0.005***	0.012***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Unemployment rate	0.004***	0.002***	0.001***	0.002***	0.006***	0.042***	0.004***	0.004***	0.002***	0.000***	0.001***	0.025***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Institution located in Bogota							0.010***	-0.025***	-0.020***	0.003***	0.001***	-0.042***
							(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Female							-0.092***	0.076***	0.075***	-0.003***	-0.004***	0.014***
							(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Year of birth							-0.003***	-0.015***	-0.011***	-0.001***	-0.001***	0.001***
							(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Students with secondary test score >= 90							-0.132***	0.142***	0.083***	-0.001***	-0.003***	0.062***
							(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)
Household income							-0.015***	0.003***	0.005***	-0.001***	-0.001***	-0.004***
							(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Academic aid x SPADIES	0.027***	-0.025***	-0.014***	0.009***	0.008***	0.062***	0.026***	-0.023***	-0.014***	0.004***	-0.005***	0.035***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.005)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.004)
Private financial aid x SPADIES	-0.002***	-0.004***	-0.001***	-0.002***	-0.000	0.007***	-0.002***	-0.006***	-0.002***	0.002***	-0.001***	0.003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Public financial aid x SPADIES	0.003***	0.007***	0.007***	0.004***	0.007***	0.095***	0.003***	0.011***	0.010***	0.001***	-0.000	0.086***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Constant	0.236***	0.434***	0.313***	0.016***	-0.041***	0.756***	5.722***	30.369***	21.880***	1.243***	1.382***	-2.156***
	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.003)	(0.082)	(0.081)	(0.074)	(0.015)	(0.013)	(0.113)
Observations	22,453,352	22,453,352	22,453,352	22,453,352	22,453,352	22,453,352	22,453,352	22,453,352	22,453,352	22,453,352	22,453,352	22,453,352
Number of id	4,131,302	4,131,302	4,131,302	4,131,302	4,131,302	4,131,302	4,131,302	4,131,302	4,131,302	4,131,302	4,131,302	4,131,302
R^2	0.0142	0.103	0.0179	0.000554	0.000754	0.00676	0.0307	0.103	0.0417	0.00294	0.00453	0.00716
Dependent variable mean	0.478	0.275	0.222	0.055	0.029	0.026	0.478	0.275	0.222	0.055	0.029	0.026

Note: The Table shows the results for the estimations using Equation 2.1. SPADIES is a dummy of 1 if the student was enrolled in a semester after the SPADIES was installed in its HEI. Tenure in the program is expressed in semesters. Share of failed classes is estimated in t+1 as the ratio of the failed classes and total classes reported by the HEI in t. The HEI reports academic and Private financial aid to SPADIES. The unemployment rate is the semestral average for the region where the HEI is located. MEN reports the location of the HEI in the HEI's directory. Female is a dummy that takes the value of 1 if the student report being female at the moment of the secondary test score. Year of birth is reported by the student at the moment of the secondary test. Students with a secondary test score >=90 is a dummy that is 1 if the student has a score higher than 90. Household Income is an increasing categorical variable reported by the student at the moment of the secondary test. Source: SPADIES.

Figure 2.1: HEIs Distribution by Round



Round 1

1101 1735
1102 1801
1103 *1802*
1104 1803
1111 1804
1121 1805
1124 1812
1125 1813
1201 *1817*
1202 *1829*
1203 2711
1204 *2737*
1209 3201
1210 *9122*
1219 *9125*
1220
1221
1222
1223
1301
1701
1702
1704
1705
1707
1708
1710
1711
1712
1713
1716
1719
1723
1730
1731

Round 2

1106 2829
1107 *2841*
1108 3115
1109 *3116*
1110 3117
1113 3302
1117 3705
1119 3803
1120 3830
1123 4102
1218 4801
1703 4808
1706
1709
1715
1724
1725
1729
1732
1815
1822
2302
2704
2707
2712
2721
2727
2744
2746
2749
2805
2811
2812
2813
2815

Round 3

1105 2719
1112 2723
1114 2724
1205 2725
1206 2728
1208 2732
1213 2823
1214 2825
1215 2832
1216 2847
1714 2850
1722 3301
1726 3702
1728 3710
1733 3713
1806 4101
1807 4111
1808 4702
1809 4711
1810 4714
1811 4721
1823 4726
1824 4832
1827 5802
1828
1830
1832
1833
1834
2104
2209
2702
2708
2710
2713

Round 4

1115 2736 3719 4716
1118 2738 3720 4719
1122 2739 3725 4727
1207 2740 3801 4803
1212 2741 3805 4806
1217 2745 3806 4810
1717 2747 3807 4811
1718 2748 3808 4813
1720 2810 3809 4817
1734 2818 3810 4818
1814 2820 3811 4822
1816 2824 3812 4825
1818 2827 3817 4826
1819 2828 3819 4827
1820 2830 3820 4829
1825 2831 3821 4835
1826 2833 3822 4837
1831 2834 3824 5801
1835 2837 3826 9101
2102 2838 3827 9116
2106 2840 3828 9117
2110 2842 3829 9119
2114 2848 3831 9120
2206 2849 3833 9121
2207 3102 3834 9124
2208 3103 4106 9126
2211 3104 4107 9128
2301 3107 4108 9131
2701 3204 4109
2709 3703 4110
2715 3706 4112
2720 3712 4701
2730 3715 4705
2731 3716 4708
2733 3718 4709

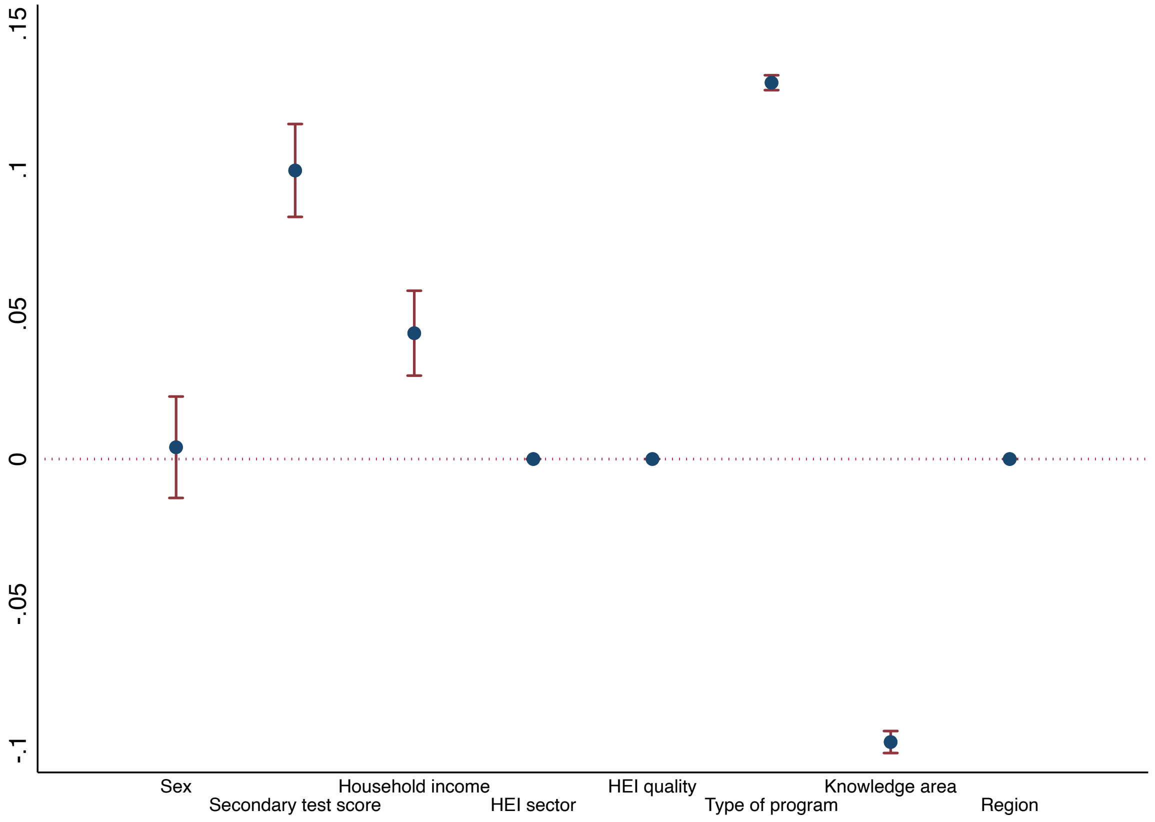
Round 5

2743
2836
2902
2903
2906
3114
3303
3724
3802
4812
9102
9127
9129
9132
9899
9903

■ Main Campus
▨ Other Campuses

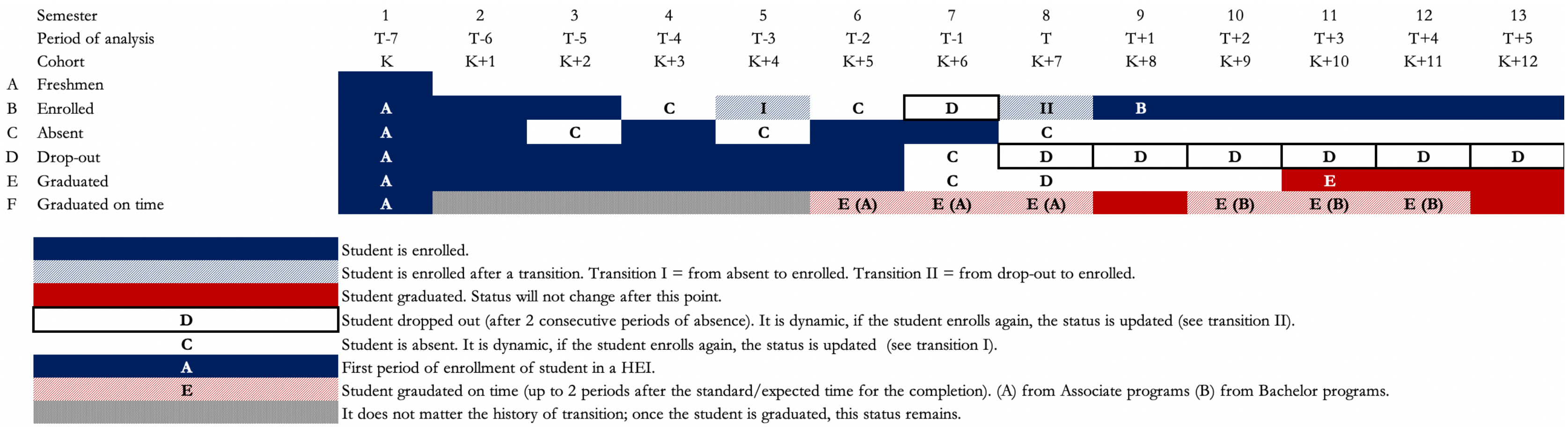
Notes: Five rounds -Round 1 (2005-06), Round 2 (2006-07), Round 3 (2007), Round 4 (2008-09), and Round 5 (2010-11)- were necessary to complete all the HEIs' information into the system. The MEN visited and installed the dashboard in the main HEI; MEN expected that the main HEI shared SPADIES with its other campuses. "Other campuses" were included in the same round that its parent HEIs; they are an extension of one of the main HEIs in other regions (e,g. Universidad Nacional de Colombia code 1101 is the main public national university, located in Bogotá, and it is the parent of 1102 that is the campus located in Medellín. In some cases, the "Other campuses" administration is autonomous, and in other cases, it administration depends directly on the main campus. There is not a rule about this).

Figure 2.2: SPADIES Assignation Balance



Notes: The Figure 2.2 shows the regression coefficients explaining the treatment variable using the students' main characteristics and HEIs. Whiskers show the 95% confidence interval. Sex is a dummy that is 1 if the student is female. The secondary test score is a dummy that is 1 if the student is in the top 10

Figure 2.3: SPADIES Status Analysis

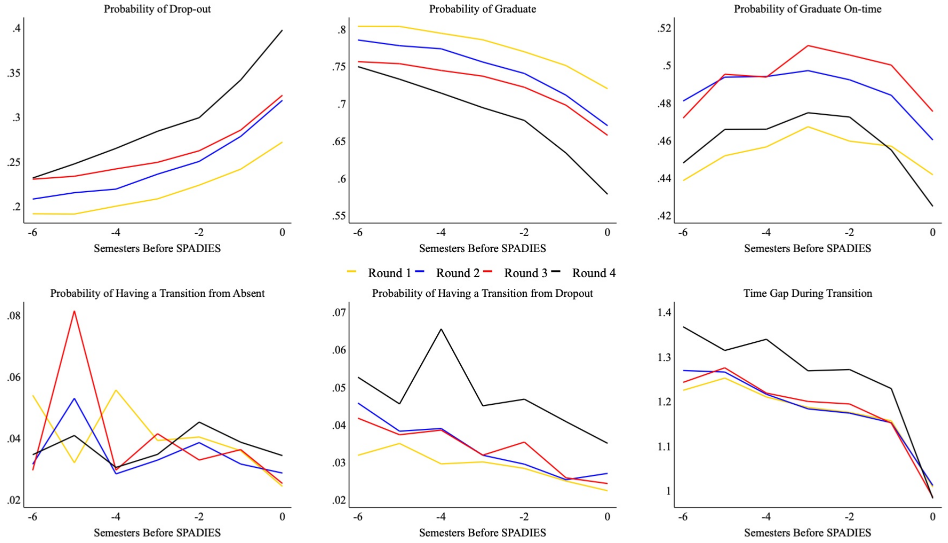


	Semester	1	2	3	4	5	6	7	8	9	10	11	12	13
	Period of analysis	T-7	T-6	T-5	T-4	T-3	T-2	T-1	T	T+1	T+2	T+3	T+4	T+5
	Cohort	K	K+1	K+2	K+3	K+4	K+5	K+6	K+7	K+8	K+9	K+10	K+11	K+12
A	Freshmen													
B	Enrolled	A			C	I	C	D	II	B				
C	Absent	A		C		C			C					
D	Drop-out	A						C	D	D	D	D	D	D
E	Graduated	A						C	D			E		
F	Graduated on time	A					E (A)	E (A)	E (A)		E (B)	E (B)	E (B)	

	Student is enrolled.
▨	Student is enrolled after a transition. Transition I = from absent to enrolled. Transition II = from drop-out to enrolled.
	Student graduated. Status will not change after this point.
D	Student dropped out (after 2 consecutive periods of absence). It is dynamic, if the student enrolls again, the status is updated (see transition II).
C	Student is absent. It is dynamic, if the student enrolls again, the status is updated (see transition I).
A	First period of enrollment of student in a HEI.
E	Student graudated on time (up to 2 periods after the standard/expected time for the completion). (A) from Associate programs (B) from Bachelor programs.
▨	It does not matter the history of transition; once the student is graduated, this status remains.

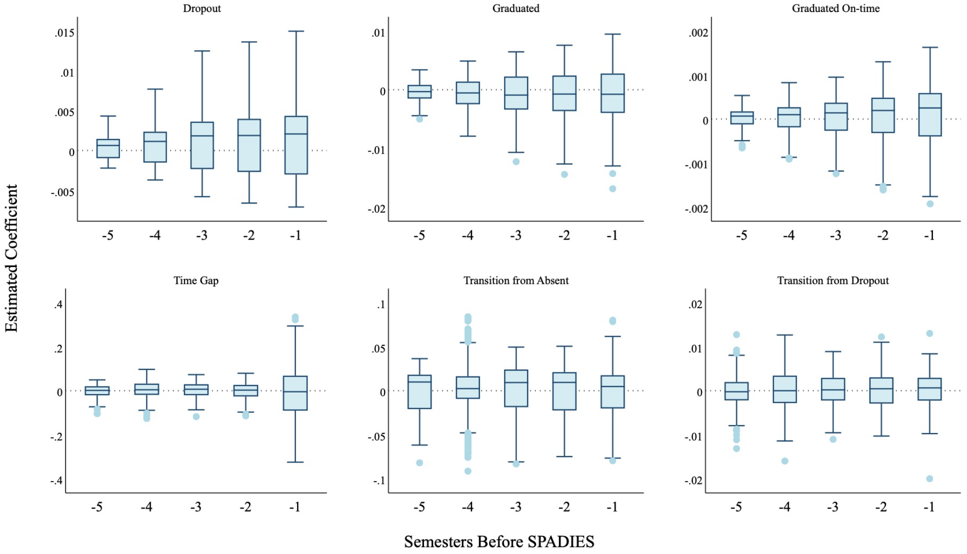
Notes: All the statuses in SPADIES are dynamic in time except by Graduated. Transition I shows an enrolled status once the student was absent per 1 semester and returned to the system. Transition II shows that after being considered a dropout student the status was updated to "enrolled" once the student returned to the system. Letters from "A" to "F" show the status of the student. Each row shows an example of the status in T. "E(T)" if graduated from a tech and "E(P)" from a professional program. Letters in cells are reference to examples of dynamic statuses in time. Time is recorded as the times an student is enrolled; i.e. for cohort K, if our analysis is in period "T", an student who has been enrolled all the semesters will be in 8th semester, but none of the examples account this. In the example "B Enrolled" the student will account 5th, as the student has been enrolled in 5 semesters and considered absent in 2 semesters and dropout in 1. In the example "E Graduated" the student will account 6 semesters (1 as absent and 1 as dropout).

Figure 2.4: Parallel Trends Averages



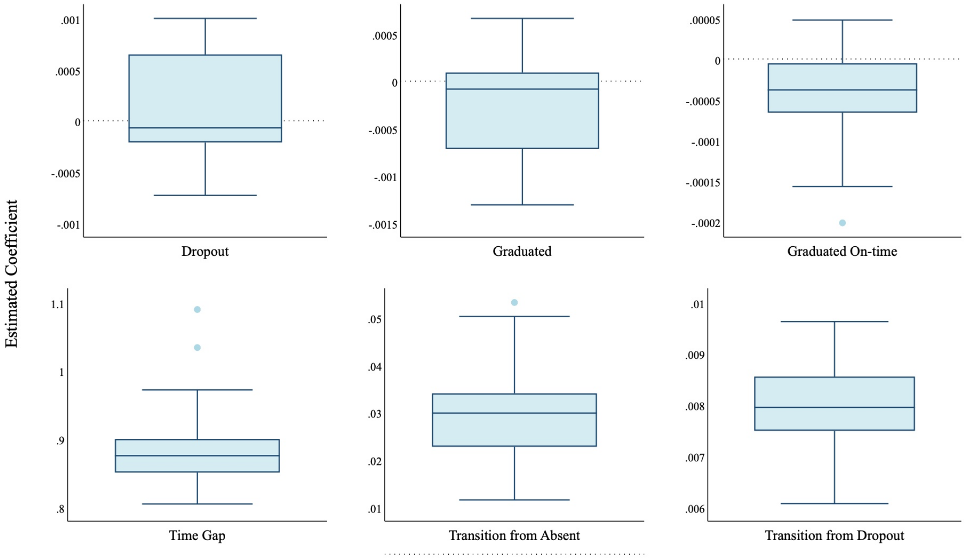
Notes: Chart reports the averages per round for the out put variables before SPADIES. Averages estimates using the full sample 4,131,302 individuals.

Figure 2.5: Parallel Trends - Classic DiD



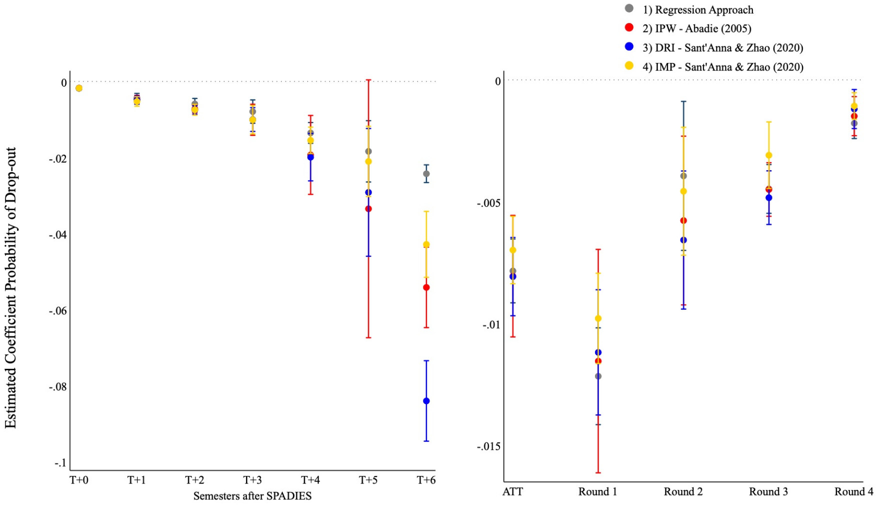
Notes: Chart reports estimated coefficients for the placebo test using the classic DiD methodology by using the placebo sample with 2,015,868 individuals.

Figure 2.6: Parallel Trends - Callaway and Sant'Anna (2021) ATT Estimation



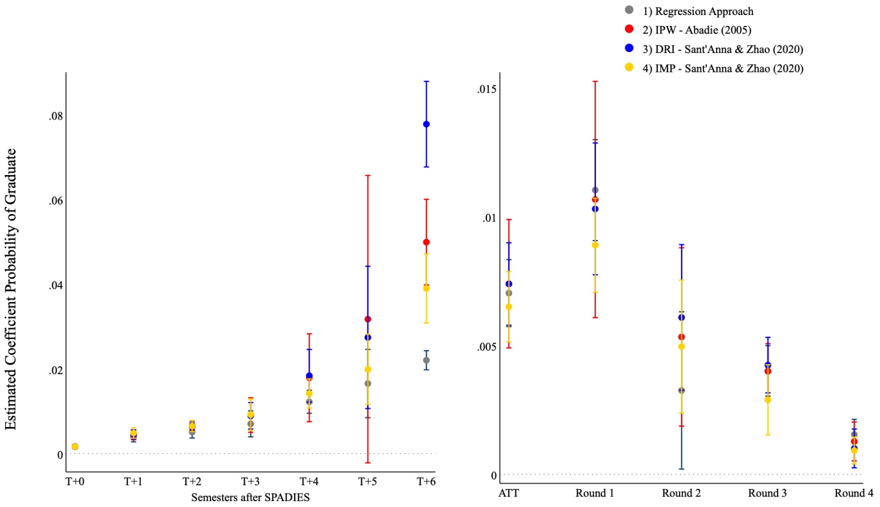
Notes: Chart reports estimated coefficients for the placebo test using the Callaway and Sant'Anna (2021) methodology by using the placebo sample with 2,015,868 individuals.

Figure 2.7: SPADIES ATT for the Probability of Dropping Out



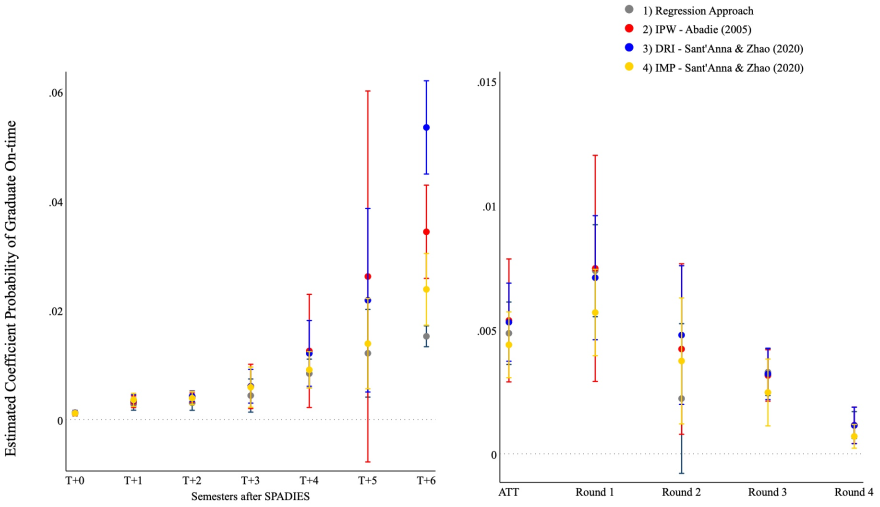
Notes: Chart reports the estimated coefficients for the probability of dropping out using Equations 2.7 to 2.12 -OR, IPW, DRI, IMP- (whiskers at 95%) using the full sample (4,131,302 individuals). The Chart is divided into two panels. On the left is an event analysis report and on the right panel a comparison of coefficients for the total and per round. Coefficients were estimated using Callaway and Sant'Anna (2021)'s framework using Rios-Avila et al. (2021) CSDID command in Stata.

Figure 2.8: SPADIES ATT for the Probability of Graduation



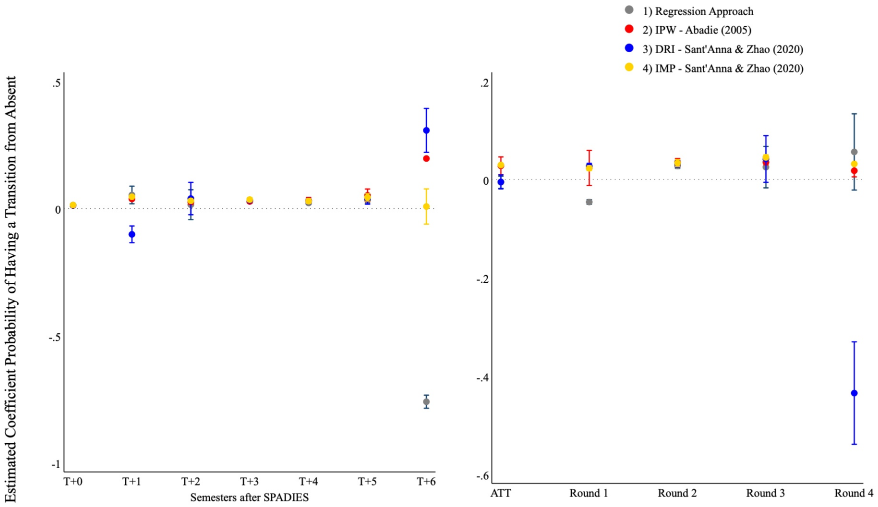
Notes: Chart reports the estimated coefficients for the probability of graduating using Equations 2.7 to 2.12 -OR, IPW, DRI, IMP- (whiskers at 95%) using the full sample (4,131,302 individuals). The Chart is divided into two panels. On the left is an event analysis report and on the right panel a comparison of coefficients for the total and per round. Coefficients were estimated using Callaway and Sant'Anna (2021)'s framework using Rios-Avila et al. (2021) CSDID command in Stata.

Figure 2.9: SPADIES ATT for the Probability of Graduating On-time



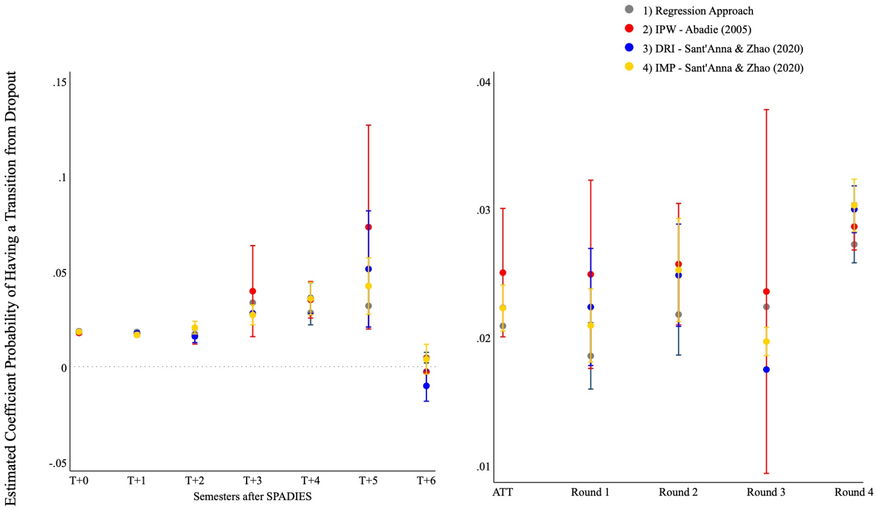
Notes: Chart reports the estimated coefficients for the probability of graduating on time using Equations 2.7 to 2.12 -OR, IPW, DRI, IMP- (whiskers at 95%) using the full sample (4,131,302 individuals). The Chart is divided into two panels. On the left is an event analysis report and on the right panel a comparison of coefficients for the total and per round. Coefficients were estimated using Callaway and Sant'Anna (2021)'s framework using Rios-Avila et al. (2021) CSDID command in Stata.

Figure 2.10: SPADIES ATT for the Probability of Having a Transition (Absent)



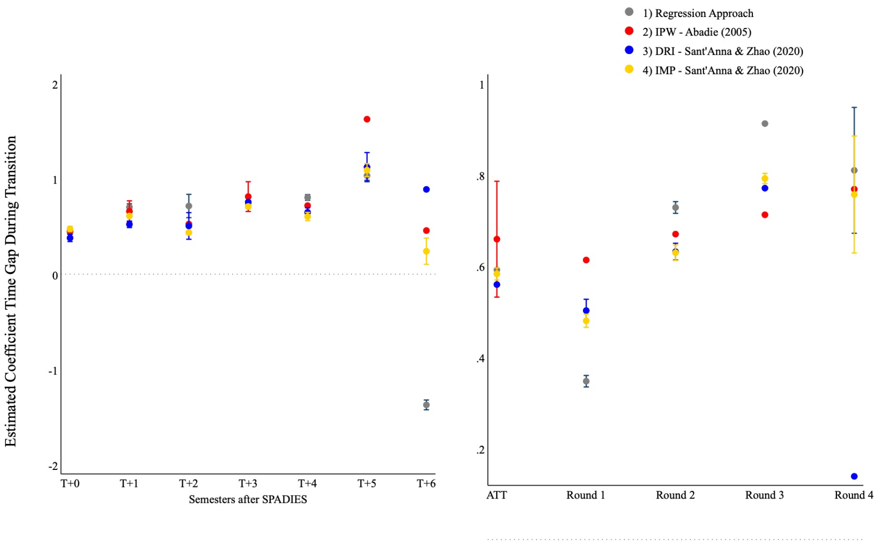
Notes: Chart reports the estimated coefficients for the probability of having a transition from Absentusing Equations 2.7 to 2.12 -OR, IPW, DRI, IMP- (whiskers at 95%) using the full sample (4,131,302 individuals). The Chart is divided into two panels. On the left is an event analysis report and on the right panel a comparison of coefficients for the total and per round. Coefficients were estimated using Callaway and Sant'Anna (2021)'s framework using Rios-Avila et al. (2021) CSDID command in Stata.

Figure 2.11: SPADIES ATT for Probability of Having a Transition (Drop-out)



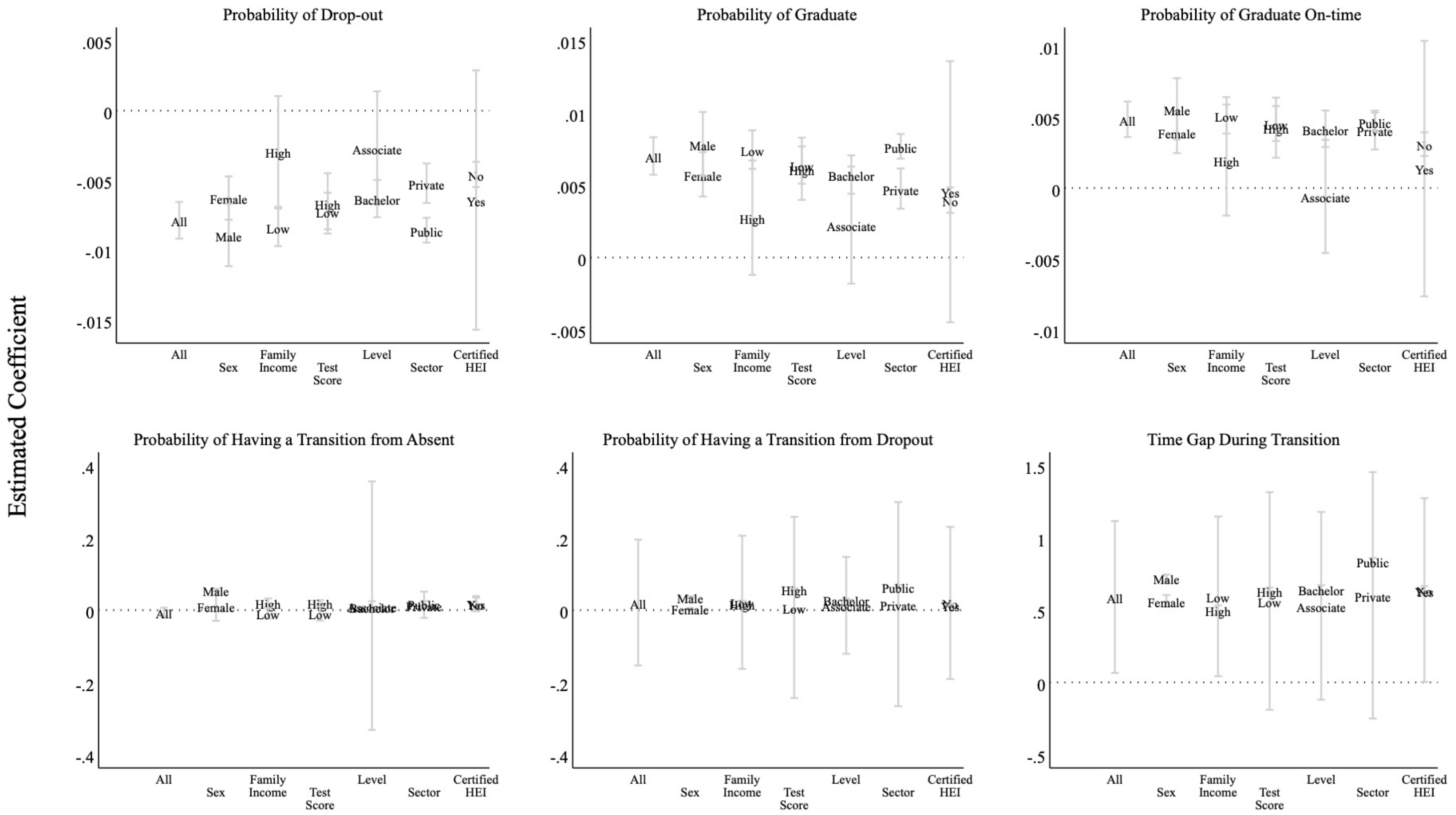
Notes: Chart reports the estimated coefficients for the probability of having a transition from drop-out using Equations 2.7 to 2.12 -OR, IPW, DRI, IMP- (whiskers at 95%) using the full sample (4,131,302 individuals). The Chart is divided into two panels. On the left is an event analysis report and on the right panel a comparison of coefficients for the total and per round. Coefficients were estimated using Callaway and Sant'Anna (2021)'s framework using Rios-Avila et al. (2021) CSDID command in Stata.

Figure 2.12: SPADIES ATT for the Time Gap During the Transition



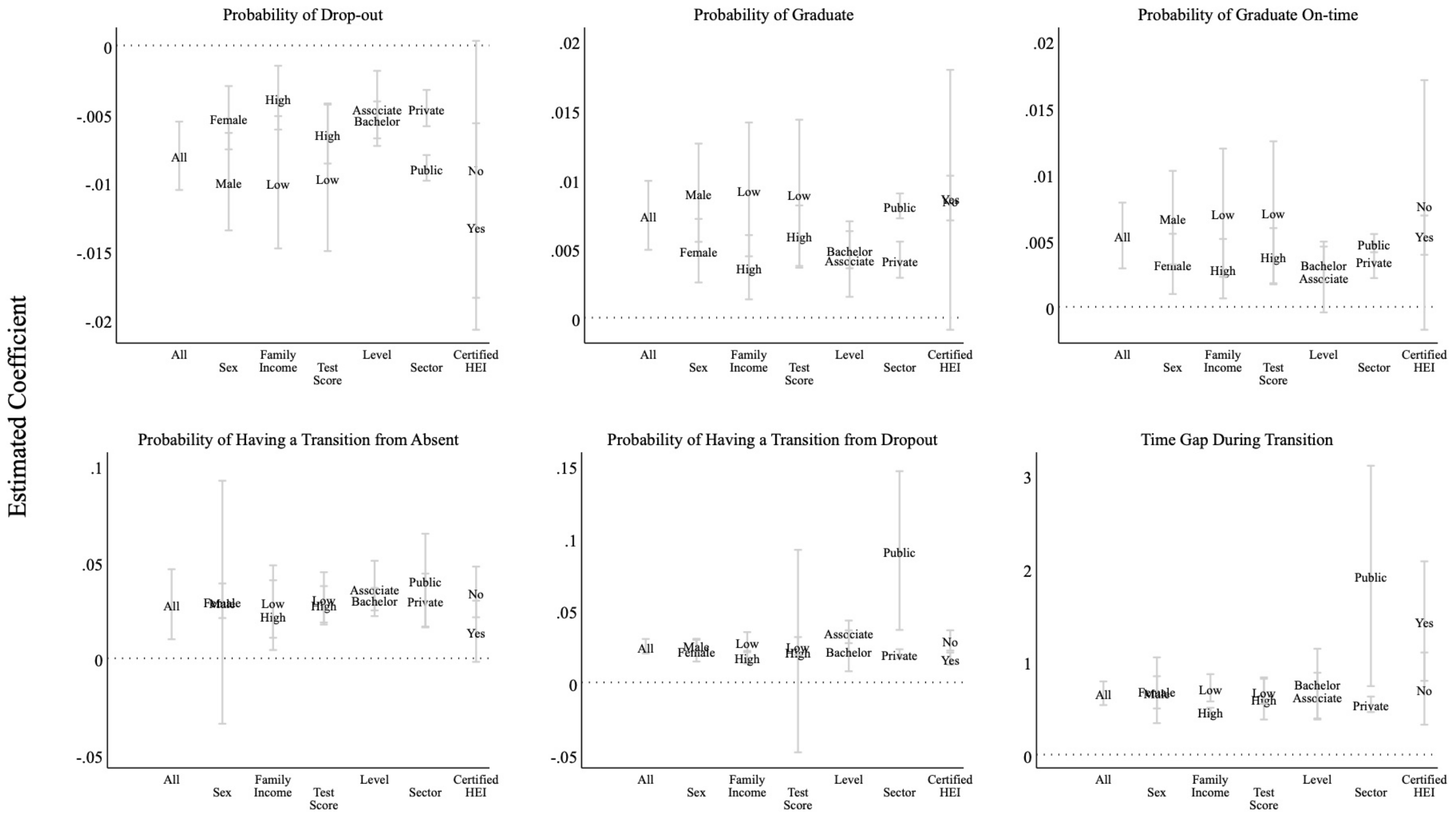
Notes: Chart reports the estimated coefficients for the time gap during transition using Equations 2.7 to 2.12 -OR, IPW, DRI, IMP- (whiskers at 95%) using the full sample (4,131,302 individuals). The Chart is divided into two panels. On the left is an event analysis report and on the right panel a comparison of coefficients for the total and per round. Coefficients were estimated using Callaway and Sant'Anna (2021)'s framework using Rios-Avila et al. (2021) CSDID command in Stata.

Figure 2.13: Disaggregated ATT Results Using OR Approach



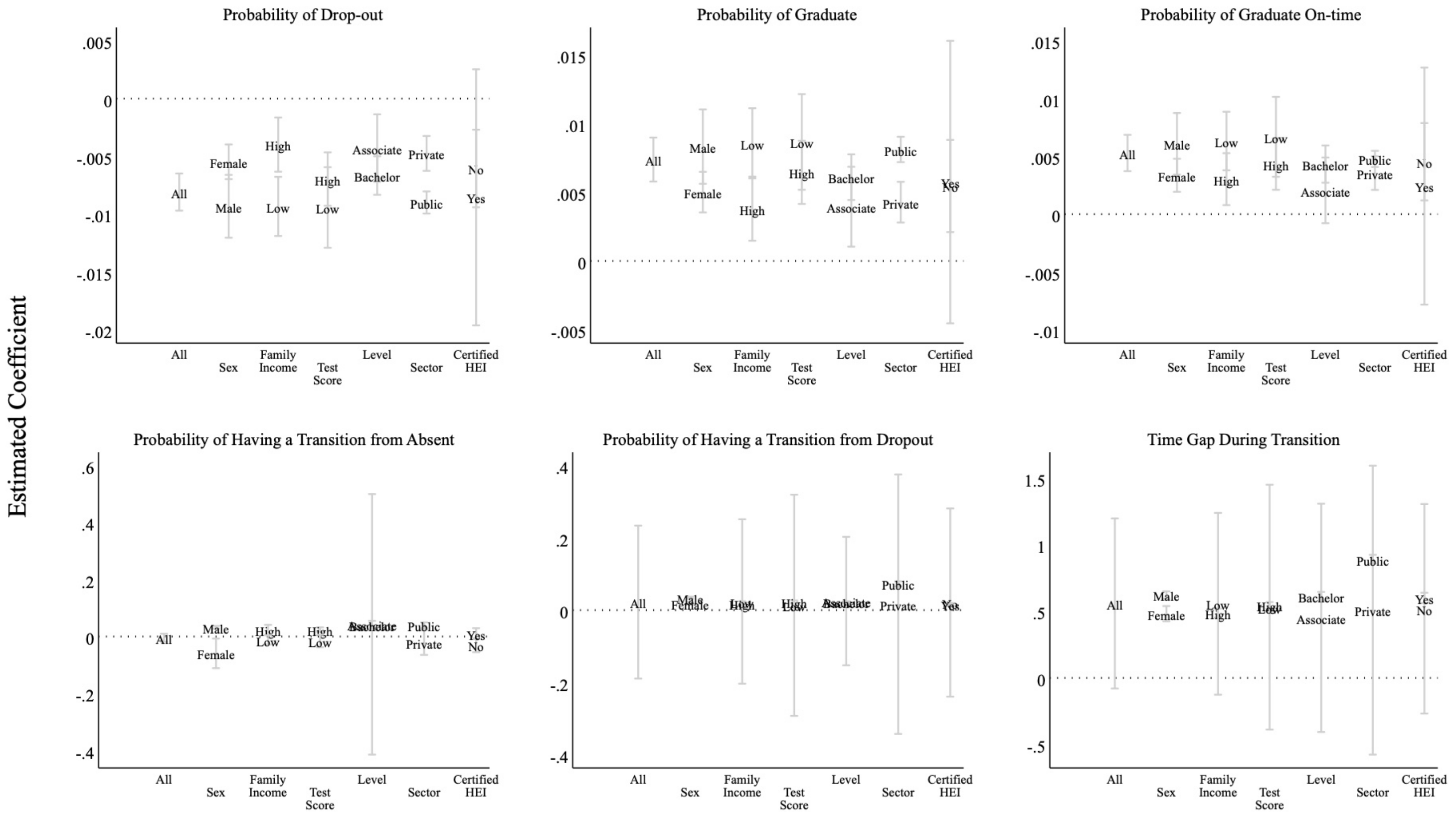
Notes: Chart reports the coefficients using Equation 2.7 -OR- (whiskers at 95%) for the estimations dividing the main sample into groups created using the time-invariant variables in the six outputs analyzed.

Figure 2.14: Disaggregated ATT Results Using IPW Approach



Notes: Chart reports the coefficients using Equation 2.9 -IPW- (whiskers at 95%) for the estimations dividing the main sample into groups created using the time-invariant variables in the six outputs analyzed.

Figure 2.15: Disaggregated ATT Results Using DRI Approach



Probability of Drop-out

Probability of Graduate

Probability of Graduate On-time

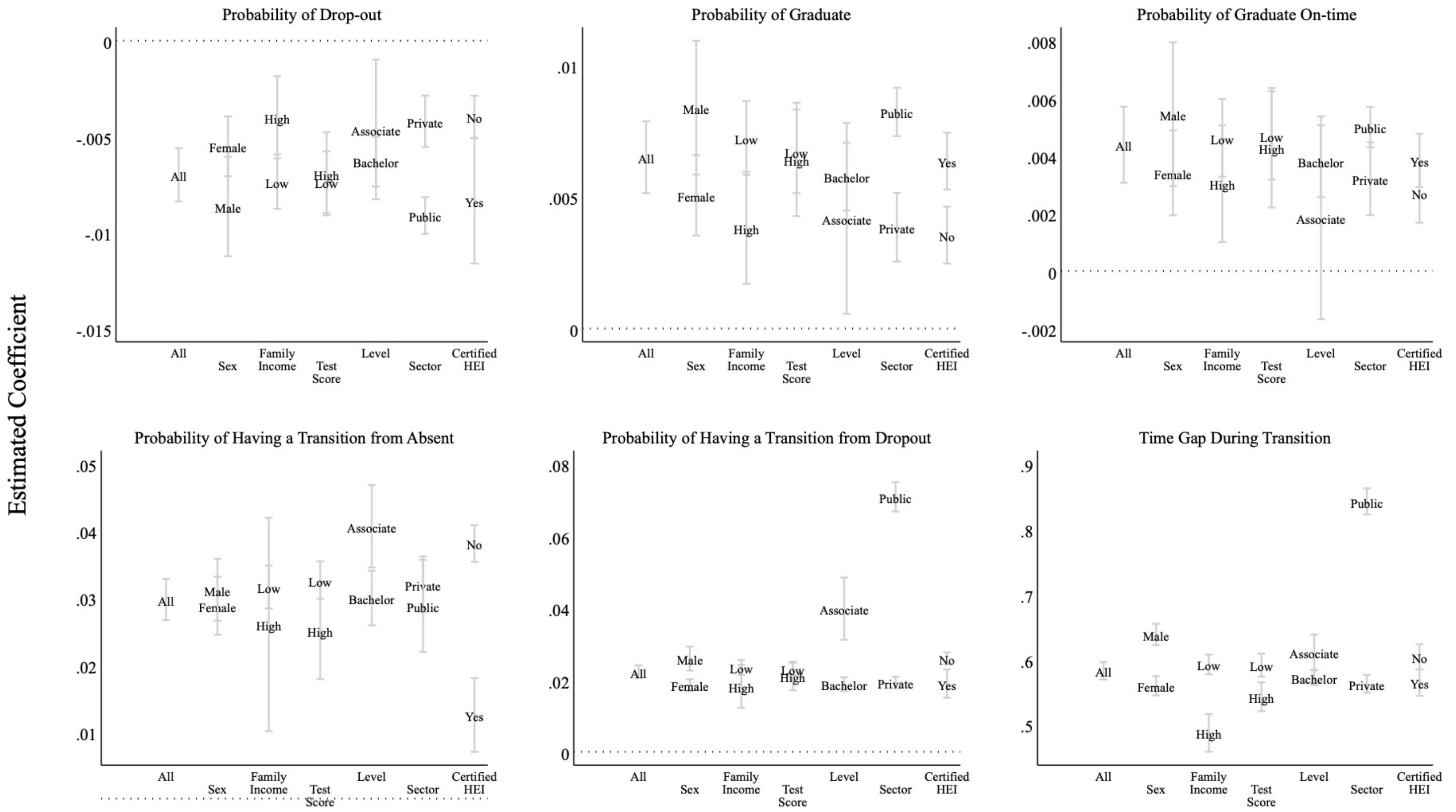
Probability of Having a Transition from Absent

Probability of Having a Transition from Dropout

Time Gap During Transition

Notes: Chart reports the coefficients using Equation 2.11 -DRI- (whiskers at 95%) for the estimations dividing the main sample into groups created using the time-invariant variables in the six outputs analyzed.

Figure 2.16: Disaggregated ATT Results Using IMP Approach



Notes: Chart reports the coefficients using Equation 2.12 -IMP- (whiskers at 95%) for the estimations dividing the main sample into groups created using the time-invariant variables in the six outputs analyzed.

References

Abadie, Alberto, “Semiparametric difference-in-differences estimators,” *Review of Economic Studies*, jan 2005, *72* (1), 1–19.

Adams, Jack and Lois VanderWaerdt, “Affirmative Action in Higher Education: A Sourcebook,” *The Journal of Higher Education*, 1984, *55* (1), 113.

Bank, Barbara, Ricky Slavings, and Bruce Biddle, “Effects of Peer, Faculty, and Parental Influences on Students’ Persistence,” *Sociology of Education*, 1990, *63* (3), 208.

Bean, John, “Dropouts and turnover: The synthesis and test of a causal model of student attrition,” *Research in Higher Education*, 1980, *12* (2), 155–187.

_ , “Interaction Effects Based on Class Level in an Explanatory Model of College Student Dropout Syndrome,” *American Educational Research Journal*, 1985, *22* (1), 35–64.

Becker, Gary, “Investment in Human Capital: A Theoretical Analysis,” *Journal of Political Economy*, 1962, *70* (5, Part 2), 9–49.

Berens, Johannes, Kerstin Schneider, Simon Görtz, Simon Oster, and Julian Burghoff, “Early Detection of Students at Risk – Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods,” *SSRN Electronic Journal*, 2021.

Bound, John and Sarah Turner, “Cohort crowding: How resources affect collegiate attainment,” *Journal of Public Economics*, 2007, *91* (5-6), 877–899.

_ **, Brad Hershbein, and Bridget Terry Long**, “Playing the admissions game: Student reactions to increasing college competition,” in “Journal of Economic Perspectives,” Vol. 23 2009, pp. 119–146.

Callaway, Brantly and Pedro H.C. Sant’Anna, “Difference-in-Differences with multiple time periods,” *Journal of Econometrics*, 2021, *225* (2), 200–230.

Cárdenas, Ernesto, "Estudio de la deserción estudiantil en programas de ingeniería de la Universidad Nacional de Colombia." Tesis de maestria en dirección universitaria, Universidad de los Andes 1996.

Castaño, Elkin, Santiago Gallón, Karoll Gómez, and Johanna Vásquez, "Análisis de los factores asociados a la deserción y graduación estudiantil universitaria," *Lecturas de Economia*, 2006, *65* (65), 9–35.

Córtes, Hernán, Luis Gallego, and Gerardo Rodríguez, "The engineering faculty today: an approach towards consolidating academic indicators," *Ingeniería e Investigación*, 2011, *31* (1), 74–90.

Epple, Dennis, Richard Romano, and Holger Sieg, "Admission, tuition, and financial aid policies in the market for higher education," *Econometrica*, 2006, *74* (4), 885–928.

Facundo Díaz, Ángel Humberto, "Análisis sobre la deserción en la educación superior a distancia y virtual: el caso de la UNAD - COLOMBIA," *Revista de Investigaciones UNAD*, 2009, *8* (2), 117.

Ferreyra, Maria Marta, Ciro Avitabile, Javier Botero Álvarez, Francisco Haimovich Paz, and Sergio Urzúa, *At a Crossroads: Higher Education in Latin America and the Caribbean*, World Bank, Washington, DC, may 2017.

Gomis-Porqueras, Pedro, Jürgen Meinecke, and José A. Rodrigues-Neto, "New Technologies in Higher Education: Lower Attendance and Worse Learning Outcomes?," *Agenda - A Journal of Policy Analysis and Reform*, 2011, *18* (01).

Goodman-Bacon, Andrew, "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, dec 2021, *225* (2), 254–277.

Guzmán Ruiz, Carolina, Diana Durán Muriel, and Jorge Franco Gallego, "Deserción estudiantil en la educación superior colombiana. Metodología de seguimiento, diagnóstico y elementos para su prevención," Technical Report, Bogotá 2009.

Herrera-Prada, Luis Omar, "Determinantes de la tasa de graduación y de la graduación a tiempo en la educación superior de Colombia 1998-2010," *Coyuntura económica: investigación económica y social*, 2013, *43* (1), 143–177.

ICFES, "Estudio de la deserción estudiantil en la educación superior en Colombia.," Technical Report, Universidad Nacional - ICFES, Bogotá 2002.

Lacka, Ewelina and Tse Chiu Wong, "Examining the impact of digital technologies on students' higher education outcomes: the case of the virtual learning environment and social media," *Studies in Higher Education*, 2021, *46* (8), 1621–1634.

Leong, Lam Wai and Othman Ibrahim, "Role of Information System (IS), Social Networking Technology (SNT) and WEB 2.0 for Improving Learning Outcomes: A Case of Malaysian Universities," *Procedia - Social and Behavioral Sciences*, 2015, *211*, 111–118.

López-Pérez, M. Victoria, M. Carmen Pérez-López, and Lázaro Rodríguez-Ariza, "Blended learning in higher education: Students' perceptions and their relation to outcomes," *Computers and Education*, 2011, *56* (3), 818–826.

Lucio, Ricardo and Mariana Serrano, *La Educación Superior: Tendencias y Políticas Estatales.* 1992.

MacLeod, William Bentley and Miguel Urquiola, "Reputation and school competition," *American Economic Review*, 2015, *105* (11), 3471–3488.

McGill, Tanya J. and Jane E. Klobas, "A task-technology fit view of learning management system impact," *Computers and Education*, 2009, *52* (2), 496–508.

Ministerio de Educación Nacional, ““ La Revolución Educativa 2002 – 2006 ” Informe De Gestión a 7 De Agosto De 2006,” Technical Report, Bogotá 2006.

_ , “Boletín Educación Superior,” Technical Report, Bogotá 2017.

Ministerio Nacional de Educacion, “Analisis de determinates de la desercion en la Educacion Superior Colombiana con base en el SPADIES. Primera parte. Factores socioeconómicos. Factores académicos e institucionales,” Technical Report, Ministerio de Educación Nacional - Universidad de los Andes, Bogotá 2008.

Olaya, Diego, Jonathan Vásquez, Sebastián Maldonado, Jaime Miranda, and Wouter Verbeke, “Uplift Modeling for preventing student dropout in higher education,” *Decision Support Systems*, 2020, *134* (May), 113320.

ONU, “Objetivos de Desarrollo del Milenio,” *Naciones Unidas*, 2013, p. 64.

Orozco Silva, Luis Enrique, *La Política de Cobertura: eje de la revolución educativa, 2002-2008.*, Bogotá: Ediciones Uniandes, 2010.

_ , Alberto Roa Valero, and Luis Carlos Castillo Gómez, “La Educación Superior en Colombia,” Technical Report 2011.

_ , Javier Medina Vásquez, María Pérez Piñeros, and Alberto Roa Valero, “Informe Colombia,” in “Proyecto Informe Sobre la Educación Superior en Iberoamérica” 2006.

Rahman, Muhammad Mofizur, “Impact of digital technology in higher education,” *International Journal of Research in Business and Social Science (2147- 4478)*, 2020, *9* (5), 318–325.

Rios-Avila, Fernando, Brantly Callaway, and Pedro H.C. Sant'Anna, “csdid: Difference-in-Differences with Multiple Time Periods in Stata,” 2021.

Sailer, Michael, Florian Schultz-Pernice, and Frank Fischer, "Contextual facilitators for learning activities involving technology in higher education: The Cb-model," *Computers in Human Behavior*, 2021, *121.*

Sánchez Torres, Fabio José, "Equidad social en el acceso y permanencia en la universidad publica determinantes y factores asociados," 2002, *7191*, 1–48.

Sant'Anna, Pedro H.C. and Jun Zhao, "Doubly robust difference-in-differences estimators," *Journal of Econometrics*, nov 2020, *219* (1), 101–122.

Sari, Arif, "WITHDRAWN: Influence of ICT Applications on Learning Process in Higher Education," *Procedia - Social and Behavioral Sciences*, 2014, *116*, 4939–4945.

SPADIES, "Reporte Modelo SPADIES al MEN," Technical Report, Universidad de los Andes, Bogotá 2008.

Sun, Liyang and Sarah Abraham, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, dec 2021, *225* (2), 175–199.

Tinto, Vincent, "Dropout from Higher Education: A Theoretical Synthesis of Recent Research," *Review of Educational Research*, 1975, *45* (1), 89–125.

_ , "Limits of Theory and Practice in Student Attrition," *The Journal of Higher Education*, 1982, *53* (6), 687.

Tongkaw, Aumnat, "Multi Perspective Integrations Information and Communication Technologies (ICTs) in Higher Education in Developing Countries: Case Study Thailand.," *Procedia - Social and Behavioral Sciences*, 2013, *93*, 1467–1472.

Trow, Martin, "Problems in the Transition from Elite to Mass Higher Education. Carnegie Commission on Higher Education," *International Review of Education*, 1973, *18*, 57.

Universidad Nacional de Colombia, *Cuestión de Supervivencia. Graduación Deserción y Rezago*, Bogotá: Beta Impresores Ltda, 2007.

Valentín, Alberto, Pedro Mateos, María González-Tablas, Lourdes Pérez, Estrella López, and Inmaculada García, "Motivation and learning strategies in the use of ICTs among university students," *Computers and Education*, 2013, *61* (1), 52–58.

Abstract Chapter 3

EN

This study examines college attendance's impact on wages in Colombia, a country with high informality and youth unemployment. Analyzing data for all secondary school graduates entering the labor market, we find a 38.4% Local Average Treatment Effect for attendance. Employing a differences-in-differences framework, we estimate a 50.6% Average Treatment Effect on the Treated for college graduates. We observe a 68.1% wage differential between graduates and dropouts with over 90% coursework completion. Results highlight college graduation's role in reducing income inequality and narrowing the gender wage gap, suggesting opportunities for policymakers to promote higher education for improved labor market outcomes.

DE

Diese Studie untersucht die Auswirkungen des College-Besuchs auf die Löhne in Kolumbien, einem Land mit hoher Informalität und Jugendarbeitslosigkeit. Bei der Analyse der Daten für alle Sekundarschulabsolventen, die in den Arbeitsmarkt eintreten, finden wir einen durchschnittlichen lokalen Treatmenteffekt von 38,4% für den Schulbesuch. Unter Verwendung eines Differenzen-in-Differenzen-Ansatzes schätzen wir für Hochschulabsolventen einen durchschnittlichen Behandlungseffekt von 50,6%. Wir beobachten einen Lohnunterschied von 68,1% zwischen Hochschulabsolventen und Personen, die ihre Ausbildung abgebrochen haben obwohl sie bereits mehr als 90% der Kurse abgeschlossen hatten. Die Ergebnisse unterstreichen die Rolle der Hochschulbildung bei der Verringerung der Einkommensungleichheit und des geschlechtsspezifischen Lohngefälles und zeigen Möglichkeiten für politische Entscheidungsträger auf, die Hochschulbildung zu fördern und damit die Arbeitsmarktergebnisse zu verbessern.

Chapter 3

Returns to Education in Colombia: New Empirical Evidence with a Comprehensive Dataset

3.1 Introduction

The decision to pursue higher education significantly influences access to quality employment opportunities, particularly in developing countries with substantial educational disparities. Colombia's unique economic and labor challenges provide a compelling case study. In the early 2000s, the country grappled with a severe economic crisis, heightening income inequality, informal labor markets, and high self-employment rates, especially among young workers. Government policies aimed to increase college enrollment and attendance rates to address these issues, but college dropout rates remain a concern. This chapter explores the impact of higher education on future earnings, examining both graduates and dropouts.

A substantial body of literature has delved into the consequences of pursuing higher education, with foundational contributions harking back to Becker (1962),Spence (1973), Mincer (1974), and Hungerford and Solon (1987). The Mincer's equation (Mincer, 1974), a pivotal tool for estimating the relationship between education and income, is frequently employed in this context, often drawing from household surveys. While generally, higher education levels correspond to increased income, an ongoing debate persists, particularly in developing countries characterized by unequal access to education (Card, 2001; Duflo, 2001). Recent research, however, suggests that the returns on education are reasonably consistent across both developed and developing nations, though outcomes can vary based on factors like geographic region, ethnicity, and the educational sector (Patrinos and Psacharopoulos, 2020; Peet et al., 2015). Brazil, Chile, and Colombia have conducted longitudinal surveys to obtain more precise insights into these returns, with particular interest in Brazil and Colombia due to their secondary and tertiary education exit exams (MacLeod et al., 2017; Manacorda et al., 2007; Melguizo and Wainer, 2016). Notably, prior research in Colombia has omitted self-employed workers from the analysis of education returns despite OECD statistics indicating that they constitute approximately 53.1% of the formal Colombian labor force. Furthermore, the Sheepskin Effect[1] , an aspect yet to be explored, remains uncharted territory, as previous studies have primarily relied on household surveys comparing college students to those at lower educational levels.

This chapter addresses these gaps by leveraging an administrative database encompassing comprehensive records for 5.4 million students who completed secondary school between 2002 and 2012 in Colombia. These records utilize information about college attainment, non-formal education achievements, and performance in the formal Colombian labor market, encompassing self-employed workers in the analysis. The main research questions are twofold: first, whether attending higher education makes a difference in future formal earnings, particularly in a country with a high level of informality in the labor market; and second, I aim to evaluate the real value of a college degree, including the so-called Sheepskin Effect.

[1] The sheepskin effect is a term used to denote when people with an academic degree earn a higher income than those with an equivalent schooling level but without the credential. This effect was first described by Hungerford and Solon (1987), and analyzed in Colombia by (Mora, 2003; Mora and Muro, 2008)

The primary empirical approach entails estimating a modified Mincer equation, which incorporates a binary variable distinguishing individuals who attended college (assigned the value one) from those who did not (assigned the value zero). In addition to panel estimations, I utilize the framework proposed by Callaway and Sant'Anna (2021) to estimate the Average Treatment Effect on the Treated (ATT), and I also employ an Instrumental Variables (IV) approach that leverages the distance between high school and college as an instrument for college attendance. While panel estimations offer valuable context consistent with existing research, the ATT and LATE estimations enable a more precise assessment of the causal impact of attending college on earnings. Additionally, by incorporating instrumental variables, we can address endogeneity concerns.

The IV estimations reveal that attending college leads to a 38.4% increase in earnings compared to those who did not pursue higher education. Regarding the ATT estimations, I observe that attending college yields a premium of 6% for the overall population, which aligns with the results from the panel analysis. However, for those who obtain a degree, the premium rises significantly to 50.6%, with specific premiums of 53.1% for bachelor's degree graduates and 92.83% for diploma (a graduate degree between bachelor and masters, it is explained in detail later in the document) holders. To calculate the Sheepskin Effect, I computed the difference in ATT for the after-college period between graduates (50.6%) and individuals who completed over 90% of the coursework but did not receive a degree, resulting in a Sheepskin Effect of 68.1%.

The empirical findings demonstrate that timely graduation from higher education results in increased earnings, irrespective of an individual's socioeconomic background. Graduates enjoy significantly higher incomes than college dropouts and those who did not attend college. Notably, the wages of female college graduates exhibit faster and more significant growth than those of male college graduates, indicating a gradual reduction in the gender wage gap.

The Colombian labor market places a similar value on college graduates with no work experience as it does on workers with six or seven years of work experience but without higher education, depending on whether the degree is obtained on time. These findings align with the conclusions drawn by Jaeger and Page (1996), emphasizing the significant importance of academic preparation in the early stages of a professional career. Additionally, our study underscores the presence of substantial income premiums for individuals with high cognitive skills or high-household-income levels. Income and the type of high school sector attended are strong predictors of future salary, reflecting the highly segregated educational system. However, higher education slightly narrows the income gap among students from different backgrounds once they obtain their degrees. We also note a stagnation of self-employed individuals, although the study does not establish the underlying cause for this group's reduced or null premiums.

Finally, the returns to higher education in the medium to long run are positive and not statistically different across various education levels, including apprenticeship, professional, or associate programs. This result differs from previous studies conducted by Busso et al. (2020); and González-Velosa et al. (2015), associate degrees were reported to have negative higher education premiums. However, we attribute this difference to the control group used in our study, which consisted of peers with similar characteristics to those who attended community college but did not pursue higher education.

The Sheepskin Effect presents a challenge for college dropouts, as they face financial investments in tuition, potential debt, and time spent in college without recognition in the labor market. They lack working experience and earn less than peers who never attended college. This highlights the importance of completing a college degree once started. Graduates enjoy significant earnings advantages over dropouts and non-college attendees, emphasizing the value of persistence in obtaining a degree.

To address this issue, policies and interventions should aim to reduce college dropout rates, as Colombia did in the past two decades. Colombia has reported an increase in higher education enrollment rates and a decline in dropout rates, but graduation rates have not kept up with these trends (Ferreyra et al., 2017; Herrera-Prada, 2013; Ministerio de Educación

Nacional, 2017). This study emphasizes in the costs incurred by students who drop out after completing over 90% of their program. Our results indicate a positive impact of higher education on career outcomes, as evidenced by the "Stairway to Heaven" effect of graduating from college, but also a “Highway to hell" effect for those who drop out, particularly in the last part of their program.

The following section presents the literature review. Section 3.3 describes the data and variables, and Section 3.4 discusses the conceptual framework and the models. Section 3.5 presents the results, and section 3.6 provides conclusions and discussion.

3.2 Theoretical Framework and Empirical Literature Review

In this section, in the first part, we conduct an in-depth review of pertinent literature to establish the theoretical framework. In the second part, our objective is to discern the existing research gaps, emphasizing studies within the purview of developing countries, with a specific spotlight on the Colombian scenario.

3.2.1 Theoretical Framework

Extensive research explores the intricate link between education and human capital development, which includes intangible assets like knowledge and skills. This investment in human capital significantly influences earnings, habits, and overall health. According to Becker (1962), education plays a pivotal role in augmenting human capital, thereby boosting earnings—up to a certain point. In essence, individuals earn more as they accumulate more years of education until the costs outweigh the benefits, as illustrated by Mincer (1974).

Subsequent literature, however, has challenged the relationship between years of schooling and labor market earnings. Researchers argue that education is a signal of qualification or even a filter, and employers often face asymmetric information that makes it challenging to select the best employee (Phelps, 1972; Arrow, 1973; Spence, 1973).

Phelps (1972); Arrow (1973); Spence (1973) propose that higher education acts as a signal of an individual's quality or skill, streamlining the hiring process. Job seekers use their educational credentials to convey their preparedness and expertise, while employers signal their preference for educated candidates, thus simplifying selection. Even individuals with existing skills pursue further education to signal qualifications to employers. Wood (2009) adds that education is rewarded with higher salaries but also comes with increased opportunity costs. This motivates highly skilled workers to advance more quickly in the education system, while less skilled individuals may drop out due to the high cost of continued enrollment.

In summary, the interplay between education, human capital, and earnings involves signaling and screening mechanisms that shape individuals' career trajectories. Collins (1979) observed that pursuing social mobility drove students to acquire degrees and credentials, ultimately increasing the pool of graduates. In response, employers raised job requirements, seeking specific degrees, grade point averages, or coursework as selection criteria. This surge in qualified workers also triggered unintended consequences, such as grade inflation and escalating education costs. Some institutions began charging extra fees for degrees, exacerbating inequalities. Access to these credentials became a privilege, disadvantaging those unable to afford multiple degrees, tutors, and additional fees.

The value of a credential has eclipsed the knowledge and skills it represents. In the human capital framework, two individuals with the same education should earn identical salaries, irrespective of whether one possesses a degree. However, research reveals that earnings increase faster for individuals with a degree. Hungerford and Solon (1987) introduced the "Sheepskin Effect" concept when estimating a Mincer Equation with a notable discontinuity in years of higher education. Their findings demonstrate substantial salary gains compared to workers with one less year of education. Data limitations initially obscured this phenomenon, as researchers only had basic information on individuals' backgrounds, education, and earnings from the late 1960s to the early 1980s. In the 1980s and 1990s, degree completion data became available. More sophisticated information on cognitive skills, earnings, and degrees emerged after 2000.

The availability of more sophisticated data has enabled researchers to examine the credentials theory and the Sheepskin Effect rigorously. This progress facilitated case studies across different countries, using these theories as analytical frameworks. For instance, Shabbir (1991) measured the impact of master's and primary education on resource allocation. Belman and Heywood (1991, 1997) and Jaeger and Page (1996) investigated how the Sheepskin Effect affected women and men in minority groups, observing that the signal weakens over time as workers accumulate experience.

Bilkic et al. (2012) explored how the opportunity cost of pursuing a credential becomes relevant for workers, influencing their decision to continue studying or enter the labor market. Numerous case studies encompassed various countries, including Gibson (2000) for New Zealand, Ferrer and Riddell (2002) for Canada, Mora (2003), García-Suaza et al. (2014) and Bacolod et al. (2021) for Colombia, Schady (2003) and Olfindo (2018) for the Philippines, Bauer et al. (2005) for Japan, Calonico and Ñopo (2007) for Peru, Crespo and Reis (2009) for Brazil, Son (2013) for Indonesia, and Yunus (2017)for Malaysia. This emerging body of research yields two primary insights. Firstly, the labor market views years of schooling as valuable work experience. Secondly, the returns for each additional year of schooling are relatively modest compared to the disparities in returns between degree holders and non-degree holders.

3.2.2 Empirical Literature About Colombia

Since the late 1960s, extensive research has examined the complexities of Colombia's labor market, particularly the interplay between education and workforce performance. This substantial work can be categorized by the specific labor market dimensions studied and the data quality.

Before 1980, research mainly centered on evaluating returns to education in the expanding primary and secondary systems, with some few cases reporting findings for higher education. Using data for the most prominent public university, the first reports of returns to higher education in Colombia were 5%, according to Selowsky (1969) or 7.5%, according to Dougherty (1971).

Also, in the mid-1970s, research revealed significant class segregation within the Colombian education system, which persists today. It was first documented by scholars like Fields (1977), Kugler (1974), and Urrutia (1974). This research emphasizes the intrinsic connection between Colombia's educational framework and its societal class structure. Importantly, it highlights that the impact of schooling endures even when socioeconomic factors are considered. Colombian higher education witnessed a transformative shift in 1978, evolving from limited supply and demand to substantial expansion, as Orozco Silva (2010)documented. However, it was not until the 1980s that studies began reporting returns on higher education exceeding 10%.

Numerous studies using household surveys have explored education's impact on earnings. Tenjo Galarza (1993) and Tenjo Galarza et al. (2015) found stable returns to education of around 20% for post-secondary education between 1976 and the present. Arias and Chávez (2002) reported that in 1991, each year of higher education increased male salaries by 3.71% (18.6% if completed) and female salaries by 0.95% (4.75% if completed). By 1999, these figures had risen to 5.54% for males (27.7% if completed) and 4.3% for females (21.5% if completed). Prada (2006) observed increasing returns for females, from 12% in 1985 to 19% in 2000, while male returns remained steady at around 17%. Mora and Muro (2008) reported a 26% return to college from 1996 to 2000.

Other household survey studies delved into the labor market's demand side, examining salaries across education levels. For instance, Núñez and Sánchez (1998)analyzed relative wages from 1976 to 1995, noting that the supply of highly educated workers initially increased, leading to a decline in their relative salaries. However, between 1982 and 1991, a surge in demand for highly educated workers improved their salaries. This demand shift was attributed to advancing technology in the production sector, for which highly educated workers were better prepared.

The year 1991 marked a pivotal moment when Colombia significantly increased its openness to international trade, enhancing the available technology in the productive sector. Consequently, research by Mesa and Gutiérrez (1996) and Santamaría (2001) reported an escalation in salary disparities between highly and less educated workers due to the new trade policies. Zárate (2005) analyzed relative salaries from 1991 to 2000 using quantile regression; he found that education was driving salaries for high-income workers, while experience was driving salaries for low-income workers. The persistence of income and education segregation continued to channel higher salaries toward individuals from high-income backgrounds, exacerbating income inequality. This enduring trend was corroborated by Cárdenas and Bernal (1999).

A common problem faced in the papers mentioned above was reliance on household survey data that was inaccurate and incomplete Farné and Vergara 2006. Household surveys in Colombia did not contain enough detailed information to examine the question they were investigating.

In late 2004, the Colombian Ministry of Education (MEN) established the Observatory for Educated Labor (OLE), intending to gather detailed data to provide accurate insights into the relevance of tertiary education and the returns on education (Orozco Silva et al., 2011). The OLE collects individual-level information regarding job positions and salaries for college graduates in a systematic survey post-graduation. It has been pivotal in advancing research on returns to education, with Forero and Ramirez-Gomez (2008) among the first to utilize OLE data.

While their study did not specifically calculate returns to education by comparing college graduates with non-graduates, Forero and Ramirez-Gomez (2008) shed light on crucial determinants of salaries. Their findings indicated that gender (with males earning more), age (older graduates having higher incomes), location (individuals from Bogota earning more), field of study, public service employment, and having an open-ended employment contract were all factors associated with higher earning potential. Additionally, the study revealed that higher parental education correlated with higher income, reaffirming the persistence of labor market segregation.

Initially, the sample size of the OLE survey was small and biased towards institutions with digital capacity before 2010. Over the years, however, the availability of new data allowed for significant improvements in the analysis of returns to higher education. Hernández (2010) studied the returns to education using the first release of the enhanced OLE database, which was the best approach at the time but highlighted several structural problems with the data, including a lack of data on the self-employed. Hernández (2010) found that the returns to education were 5.2% for associate programs, 34.0% for Bachelors, 70.6% for diplomas, 75.0% for masters, and 128.5% for PhD. Herrera-Prada and Caballero (2013) estimated the expected time to recover the investment in college using the OLE database and aggregated tuition and expenses data. They found that graduates from public colleges recovered their investment in the first year of employment, but those from private colleges could take up to 5 years to recover the cost.

Recent research using OLE data provides valuable insights into the Colombian labor market. González-Velosa et al. (2015) found a 26% return for bachelor's degrees, with varying returns for associate programs from -33% to 25%, depending on the program. Busso et al. (2020) reported a 3.9% higher return for Bachelor programs than associate programs, with private school graduates earning 5.9% more than their public-school counterparts. Fields like engineering and medicine boasted even higher earnings. Ferreyra et al. (2020) noted that longer, in-person programs in major cities and certified institutions yielded better employment prospects and wages. MacLeod et al. (2017) highlighted the significance of college reputation on future income. Finally, de Roux and Riehl (2022) explored the impact of academic breaks between secondary school and higher education, revealing lower future income for high-performing students who took breaks.

There are still three main gaps in the literature that this chapter aims to address. First, this chapter tracks both secondary school graduates who attend college and those who do not, which has particular importance given that Colombia has a 52% attendance rate for college. Second, this chapter includes self-employed workers, which constitute 53.1% of Colombia's formal labor force and which was a group omitted in the OLE database. Third, this chapter explores the Sheepskin Effect, a phenomenon not extensively examined in prior household survey-based research. Detailed data enables a comparison of post-college employment between graduates and students who completed over 90% of their program but did not receive a degree. This chapter sheds light on the significance of degree completion.

3.3 Stylized Facts, Data, and Variables

This section presents some stylized facts of the higher education system in Colombia. We will then describe the six databases we use, the criteria to adjust them, and the variables we created and used from each. Finally, we will describe the data management and how we matched the data.

3.3.1 The Colombian Education System

In 2021, Colombia's pre-college education system had approximately 9.7 million students, with around 80% attending public institutions. It consists of primary education (5 years) and secondary education (6 years), divided into lower secondary (years 6 to 9) and upper secondary (years 10 and 11). The upper secondary stage offers different tracks, such as academia, military, or teaching. In the last year, all students took the mandatory Saber 11 exam, required to graduate secondary school and be admitted to college.

In 2018, the Ministry of Education reported 2.3 million students in higher education, of which 52.9% are women, 50% are enrolled in public institutions, and 93% are enrolled in an undergraduate program. The higher education system offers two undergraduate levels: associate degrees (2 or 3-year programs) and bachelor's degrees (4 or 5-year programs). Among undergraduates, 70.1% are pursuing bachelor's degrees. Graduate programs, which require a bachelor's degree, include diplomas (6 to 18 months), master's degrees (2 years), and PhDs (about 5 years). In 2018, 77.1% of graduate students were pursuing a diploma, 21.5% a master's degree, and 1.2% a PhD degree.

In 2018, the higher education system included 314 higher education institutions (HEIs), with 52 of them holding high-quality certification. These institutions are distributed across 28 out of the 32 departments and across 70 municipalities out of a total of 1,121. Every institution offers different programs, which are majors or concentrations (e.g., mathematics, sociology, political science). Across the 5,592 active programs, 72.7% were in-person.

As of March 2020, before the COVID-19 pandemic, Colombia's labor market included 20.5 million workers, with an employment rate of 51.7% and an unemployment rate of 12.6% (DANE, 2023) Additionally, self-employed individuals constituted 53.1% of the total population in Colombia, according to (OECD, 2021). However, both household surveys and social security data report only 23%.

3.3.2 Databases Description

For our empirical analysis, the Saber 11 test database is the primary dataset, which offers comprehensive insights into students who completed their final year of high school, including certain socioeconomic characteristics. We merged the Saber 11 database with two additional datasets: the SPADIES and Social Security (PILA). The SPADIES database provides information on secondary school graduates who proceeded to higher education, including their status and program details. Secondary graduates who pursued higher education form our "treated" group, while those who did not attend college constitute the "control" group. The Social Security database furnishes data on formal labor market income and other job-related characteristics.

The Colombian Institute for the Evaluation of Education (ICFES) administers the Saber 11 exam required to graduate from secondary school in Colombia since 1968. The empirical analysis uses data from the ICFES database for all students who took the Saber 11 exam between 2002 and 2012. The database contains individual-level information, including the student's school, gender, household income, and exam score. This information is merged with the 2016 Census on Schools collected by the Ministry of Education. Census data includes the schedules, shifts, school coordinates, school level, and school sector (i.e., public or private). In Colombia, multiple schools can operate in the same building, so a secondary school is identified according to their shift and sector. For our purpose, if the government does not operate the school, the school is classified as a private sector. A school administered by a private entity under a contract with the government is also considered private.

Throughout the analyzed period, the ICFES changed the test score range. To address this, we employed the established standardization procedure the Ministry of Education utilized in its database management. This procedure involves assigning each student a percentile ranking based on their exam performance in relation to the scores of their contemporaneous test-takers.

Given the inconsistent collection of household income data across periods and each school's relatively stable market niche, we used the same imputation approach for missing data that the Ministry of Education uses. Specifically, for periods lacking this information, household income for each student was imputed using the mode of household income within the same school during comparable periods. In cases where multiple modes were present, the highest value was selected. It is important to note that the ICFES standardizes income levels into nine ascending categories, adding a layer of structure to the imputation process[2]. Additionally, we incorporated the departmental unemployment rate per year, as reported by the Colombian Statistical Office (DANE), based on secondary school location and panel time.

We also use data from SPADIES. The SPADIES database furnishes information on secondary school graduates who pursued higher education from 1998 to 2017. This dataset comprises details such as the institution attended, academic performance, field of study, program level, duration of studies, and status within the system (e.g., dropout, graduate, or active). The status is determined by the system based on the database cutoff date, which in this case is September 2017. We refined the status categories in the SPADIES dataset to enhance the specificity of the final dataset utilized for empirical analysis.

In the SPADIES database, a “Graduate" is defined as an individual who successfully completed a higher education program and received a degree. We divided this category into two groups: "Graduated on time" for those who graduated within or up to 1 year of their expected graduation date, and "Graduated late" for those who graduated more than 1 year after their expected date. The expected graduation year is assumed to be five years for bachelor's programs and three years for associate degree programs.

The SPADIES database defines a "dropout" as a student who has not been enrolled for two or more consecutive semesters at the moment of the cut-off. To study the Sheepskin Effect, we refined this definition. We divided dropout students into two categories: "Candidates" and "Incompletes." "Candidates" refer to students who completed over 90% of their coursework but are classified as dropouts by SPADIES due to not graduating and being unenrolled for two or more consecutive semesters after reaching the 90% coursework completion at the moment of the cut-off. "Incompletes" encompass the remaining dropouts who left with less than 90% of the coursework complete. To examine the impact of dropping

[2] 0 "[0-1) minimum wages" 1 "[1-2) minimum wages" 2 "[2-3) minimum wages" 3 "[3-5) minimum wages" 4 "[5-7) minimum wages" 5 "[7-9) minimum wages" 6 "[9-11) minimum wages" 7 "[11-13) minimum wages" 8 "[13-15) minimum wages" 9 "[15-∞) minimum wages"

out at different times on income, we further subdivided the "Incompletes" into "Early Incomplete" for those who left within the first year of college and "Late Incomplete" for those who departed after the first year without becoming “Candidates."

We adopted the SPADIES definition for the "Active" status, which includes students enrolled in the system as of 2017. Students pursuing studies in natural sciences, engineering, and mathematics disciplines are identified as STEM students using a binary variable. This variable takes the value of one if the student belongs to these programs and zero otherwise. The program level is categorized as "bachelor" for those pursuing a bachelor’s degree and "associate" for those in associate degree programs. Our empirical analysis focuses on students whose expected year of college graduation falls within the range of 2001 to 2017. An "active" status in 2017 implies that the student was enrolled for a minimum of 3 years, having started their studies no later than 2014.

Furthermore, to accommodate variations in the timing of college enrollment relative to secondary school graduation, we introduce a variable that distinguishes between "early enrollees" (those who enrolled within 3 semesters after secondary graduation) and "late enrollees" (those who took more than 3 semesters to commence their college studies).

We utilized data from the SNIES (National System of Information for Higher Education) to acquire information about the Ministry of Education’s quality certification and the addresses of higher education institutions (HEIs). A binary variable was established, taking the value of one if a student is enrolled in a high-quality institution. Additionally, we computed the orthodromic distance between secondary schools and HEIs in kilometers. Since SPADIES exclusively encompasses data for undergraduates, we augmented the information regarding SPADIES graduates by integrating data from the Ministry of Education’s labor observatory (OLE). This augmentation allowed us to determine whether a student pursued and successfully completed graduate school. Consequently, SPADIES’ graduates were categorized into four distinct groups: "Bachelors," representing those who did not pursue graduate education; "Diploma," encompassing individuals who pursued specialized post-graduate degrees known locally as specializations, which are shorter and more flexible than master's degrees; "Master," for those with a master's degree; and "PhD," for those engaged in doctoral studies.

The income data utilized in this study are derived from the PILA database (Colombian Social Security Records), which contains comprehensive records of Social Security payments for all individuals in the formal sector, including their employment type (including self-employed individuals), and the location of their jobs. For this chapter, we extracted the aggregated annual sum of contribution payments made by all formal Colombian workers to the health system from 2008 to 2014. While constituting an unbalanced panel, we balanced it by assuming that if a gap in income appears for those with income in any period, it signifies zero income in the formal labor market during those missing periods. For our empirical analysis, we employed the natural logarithm of income, and each year with income greater than zero was counted as one year of experience.

The social security records classify self-employed workers based on their contributor type[3]. This classification also allows us to distinguish public servants and apprentices. It's important to note that not all employees of the State are reported in the PILA, particularly cases such as military personnel and public teachers. Regarding apprentices, these individuals are secondary school graduates who participate in training programs offered by the National Services of Apprenticeships (SENA), which is not considered higher education. The SENA's apprenticeship programs consist of both theoretical lectures and practical components, spanning from 12 to 36 months in total duration. During the practical phase of the program, participating companies take on these apprentices as interns and are required to provide payment ranging from 50% to 75% of a minimum monthly salary. These apprentices are also reported to the pension and health funds (PILA) as SENA apprentices. While we establish a variable for apprentices based on this data, it's essential to recognize that we can

[3] The students whose statuses in the PILA are self-employed workers, self-employed workers in an association, self-employed workers without regulations to contribute and other codes for the transition from employee to self-employed (code 42 and code 49) were marked as self-employed.

solely track secondary school graduates who engage with SENA and are in the final stages of their programs, which correspond to timelines similar to associate programs. Additionally, our information does not encompass details about other apprentice statuses beyond this scope. In summary, we use the following variables from the SABER 11 database: reported gender, age when the test was taken, Saber 11 standardized test score, household income, school location, school shift, and school sector. The SPADIES database is the source for the variables of students' status in the system, the program, and the higher education institution of enrollment. Finally, we use the annual income from 2008 to 2014 and the information related to the public servants, self-employed workers, and SENA apprentices.

3.3.3 The Administrative Data Matching Process and Final Database

We employed various merging approaches for the administrative databases, contingent on distinct identification variables. The merging process between Saber 11 and SPADIES and SPADIES and OLE utilized the same matching algorithm employed by the Colombian Ministry of Education to combine the Saber 11 and SPADIES databases[4]. The merging of Saber 11 and PILA was executed by the Ministry of Health and Social Protection, employing the national identification number of Colombia. The merge between Saber 11 and the Census of Schools was achieved using the ICFES' school code. Lastly, the merging of SPADIES and SNIES was conducted using the identification code of the respective Higher Education Institutions (HEIs).

The Saber 11 dataset encompasses a total of 5,425,850 secondary graduates. Subsequently, we submitted the identifiers for these 5.4 million secondary graduates to the Ministry

[4] The algorithm takes two key variables, namely the full name and the date of birth, from the databases. Firstly, the algorithm removes the spaces, converts all alphabetic characters to uppercase, and then decomposes the strings into all possible combinations of the characters. For instance, the name "Tom" is transformed into TOM, MOT, OTM, OMT, TMO, MTO. Next, the algorithm compares each discomposed key variable for every observation in each database to all possible observation matches between the databases. If the comparison reaches a certain "trigger" level, the algorithm identifies the observation as a match. The level of match is the percentage of similarity between the discomposed variables. The algorithm is cautious, meaning that if there is more than one potential matching option, it will not execute the matching. In this chapter, the trigger value used is 98%, the same as the value used by the Ministry of Education in the SPADIES-ICFES match.

of Health and Social Protection, which provided PILA information for 418,699 of them. Following this, we merged this data with SPADIES and identified that 99,571 of these secondary graduates pursued higher education. This merged outcome forms the basis for determining our treatment group. Specifically, the secondary graduates from the Saber 11 dataset who matched with the SPADIES database constitute our treatment group, while those who did not match comprise our control group.

After excluding observations with missing values in any of the variables employed in the empirical analysis, the ultimate dataset encompasses information for 393,166 secondary graduates within a balanced panel spanning the years 2002 to 2014. This results in a total of 5,111,158 observations. Descriptive statistics for variables utilized in the empirical analysis are presented in Table 3.1. A description of the college attendants by sector, program, time of enrollment, and quality of HEI where enrolled is shown in Table 3.1.

3.4 Theoretical Framework and Model

This section first presents the theories that conceptualize the returns to education and the Sheepskin Effect. Next, the empirical model specification is discussed. A modified Mincer equation, which is based on the theoretical framework, is used to estimate the returns to education and the Sheepskin Effect.

3.4.1 Model Specification

To estimate the returns to education and the Sheepskin Effect, a modified Mincer equation is used as main framework for the empirical analysis. The basic theoretical foundation for Mincer's earnings regression is:

$$ln(W(s,x)) = \alpha_0 + \rho_s s + \beta_0 x + \beta_0 x^2 + \varepsilon \quad (3.1)$$

In this context, $W(s,x)$ represents the wage for a specific level of schooling s and x denotes work experience, initially calculated by Mincer as the age minus six years of education. The initial specification in Mincer's model assumes that all individuals are relatively similar, with the only distinguishing factor among them being their choice to acquire more years of schooling. This assumption has faced significant criticism because individuals vary in various characteristics and because the non-random assignment of education influences the link between education and earnings.

We incorporate new variables to control for detailed individual characteristics to address the first concern. Additionally, we address the second concern by utilizing other econometric techniques, as explained later in this section.

In our analysis, while maintaining the simplicity of the foundational equation, we introduced the vector X .to represent a set of measurable characteristics that are correlated with an individual's salary, including time, secondary graduation cohort, sex, score in Saber 11 test, school sector, household income, age, age squared, experience, experience squared, school shift, apprenticeship, self-employment, public servant, and unemployment rate controls. This adjustment allows us to consider a broader range of individual attributes when examining wage determination, resulting in a more precise evaluation of the benefits of pursuing higher education.

In line with this approach, considering that all our workers have completed secondary education, the initial variation in schooling is determined by whether or not they attend college. Therefore, we have renamed the variable "s" to "Attend." The new equation to estimate the returns to education is as follows:

$$ln(W_{it}) = \alpha_i + \delta_t + \rho Attend_{it} + {X_{it}}'\beta + \varepsilon_{it} \quad (3.2)$$

In this equation, W_{it} irepresents the salary reported for individual i in the year t. The variable of interest is "Attend," a dummy variable that equals 1 if individual i attended college in the year $\leqq t$ and 0 otherwise. ρ would be the “rate of return for attending college." α_i controls for all the time-invariant characteristics of each individual, including gender, Saber 11 score (used as a proxy for academic ability), household income, school sector, school shift, and graduation cohort from secondary school. δ_t captures time-varying drivers of salary at the national level. The vector X includes observable predictors for the wages per individual as age, work experience, employment characteristics (such as public servant or self-employed status), and the departmental unemployment rate (utilized as a proxy for the opportunity cost of pursuing further education). We will also consider two variations of the "Attend" variable. Since "Attend" is a dummy variable for those who attended college, we can introduce a set of dummy variables to represent different statuses among college attendees. In the initial case, we consider four statuses: Graduate (G), Candidate (C), Incomplete (I), and Active (A). The equation is formally defined as follows:

$$ln(W_{it}) = \alpha_i + \delta_t + \rho_1 G_{it} + \rho_2 C_{it} + \rho_3 I_{it} + \rho_4 A_{it} + {X_{it}}'\beta + \varepsilon_{it} \quad (3.3)$$

In the second scenario, we further categorize individuals who have Graduated (G) from college into a more detailed set of dummy variables to distinguish those with a Ph.D. (D), a Master's (M), a Diploma (E), or solely a bachelor's degree (B). Additionally, we introduce new variables for Incomplete, which can be categorized as Early Incomplete (EI) or Late Incomplete (LI). The final equation for our OLS panel approach is then expressed as follows:

$$ln(W_{it}) = \alpha_i + \delta_t + \rho_1 B_{it} + \rho_2 E_{it} + \rho_3 M_{it} + \rho_4 D_{it} + \rho_5 C_{it} + \rho_6 EI_{it} + \rho_7 LI_{it} + \rho_8 A_{it} + {X_{it}}'\beta + \varepsilon_{it} \quad (3.4)$$

As presented in Equation 3.2, the baseline model allows us to compare our results with the existing literature, extract insights from the data, and progress toward addressing the questions raised in this chapter. We employ Equation 3.3 to estimate our initial approach to the Sheepskin Effect, calculated as the difference between the coefficients associated with Graduates and Candidates. Subsequently, in Equation 3.4, we investigate whether pursuing graduate school or early dropout from college yields distinct outcomes compared to our initial findings.

While consistent with prior research and relevant to our research questions, this model has limitations in establishing causality. Notably, a significant challenge in studies examining the relationship between education and earnings is the non-random assignment of education levels. Individuals make conscious decisions regarding their educational pursuits, considering opportunity cost (as highlighted by Wood (2009)). To address potential econometric issues such as sample selection and endogeneity, we employ an instrumental variables approach to estimate the Local Average Treatment Effect (LATE) of college attendance.

Furthermore, it is important to recognize that effects may vary among different groups. To address this issue, we apply Callaway and Sant'Anna (2021)'s framework to estimate the Average Treatment Effect on the Treated (ATT). This approach allows us to explore heterogeneity in the effects of education on earnings across various subgroups.

Instrumental Variables Approach

Instrumental variables represent an appropriate methodology when we can access suitable instruments for addressing potential endogeneity issues. In this study, we adopt a panel estimation for the two-stage least squares (2SLS) estimator to address these concerns methodically. Specifically, we leverage the distance from secondary school to college as an instrumental variable for quantifying educational attainment. Recognizing the inherent difficulty in locating instruments for the complete set of variables, we concentrate on instrumenting "Attend." This approach allows us to understand the relationship between education and earnings better while addressing potential sources of bias in the data.

Therefore, our approach involves estimating a first step to predict the probability of college attendance, employing the distance from school to college (the instrument) as an independent variable. The vector of control variables X_{it} remains consistent with Equation 3.2. The first step equation is formally specified as follows:

$$Attend_{it} = \alpha_i + \delta_t + \mu Distance + X_{it}{}' \beta + \varepsilon_{it} \quad (3.5)$$

The distance to college has long been utilized in the literature for similar purposes since it was first proposed by Card (1993). The intuition is that proximity to a college can significantly impact a student's decision to pursue higher education. The distance to college may influence the likelihood of attending college, but it is not necessarily related to one's ability or wealth (Card, 1993; Frenette, 2004).

Unlike in other countries, Colombia does not have university cities dependent on college campuses. Therefore, we assume that high schools and colleges are located randomly in the cities, and families of varying income levels can be found within any radius from secondary schools or colleges. Finally, distance to college is not related to wages, as individuals with different income levels can be found anywhere within the same distance radius.

In Equation 3.6, we use the estimated probability of attending college (obtained from Equation 3.5) as an instrument for "Attend." As the distance from secondary school to college satisfies the "exclusion restriction," the exogenous variation provided by the instrument in the Instrumental Variables (IV) approach gives a precise local average treatment effect (LATE). Therefore, the results in Equation 3.6 can be interpreted as the causal effect of attending college on future salaries.

$$ln(W_{it}) = \alpha_i + \delta_t + \rho \widehat{Attend}_{it} + X_{it}{}' \beta + \varepsilon_{it} \quad (3.6)$$

Heterogeneous Difference in Differences (DiD) Approach

The conventional Difference-in-Differences (DiD) approach utilizes a 2X2 model with two periods and two groups. In the initial period (t=0), both groups share the same characteristics and lack exposure to the treatment. In the subsequent period (t=1), some individuals undergo the treatment, forming a "treated" group (Attend=D=1), while others remain "controls" (Attend=D=0) without the treatment. This basic model corresponds to the interpretation of Equation 3.2, where t=0 marks the year of secondary school graduation, and t=1 represents the subsequent year when some individuals attend college while others enter the workforce. Equation 3.7 describes the basic approach for a DiD based on Equation 3.2, forgetting momentarily the $X_{it}\beta$ component that will be incorporated later once the homogeneous is specified.

$$Y_{it}(D) = ln(Y_{it}) = \gamma_i + \theta_i t + \rho_i D \times t + \varepsilon_{it} \qquad (3.7)$$

In Equation 3.7, γ_i is the individual fix effect, θ_i is an individual specific trend. ρ_i is the individual-specific treatment effect. For t=0, Y_{i0} (D=1)=Y_{i0} (D=0). Indeed, each individual in this framework has two potential outcomes, one with treatment and one without treatment. However, our observations are limited to the outcomes corresponding to each group (treated or not treated) in t=1. In theory, these outcomes should differ due to the presence or absence of the treatment and are given by:

$$Y_{it}(D) = D_i Y_{it}(1) + (1 - D_i) Y_{it}(0) \qquad (3.8)$$

Assuming that the treated group would follow a predetermined trajectory in the absence of treatment, any deviation from this path can be attributed to the causal impact of the treatment on this group. This deviation, denoted as the Average Treatment Effect on the Treated (ATT), is described in Equation 3.9.

$$ATT = \underbrace{E(Y_{i,1}(1)|D_i = 1)}_{A=Observed\, outcome\, for\, treated} - \underbrace{E(Y_{i,1}(0)|D_i = 1)}_{B=Unobserved\, outcome\, for\, treated} \qquad (3.9)$$

In Equation 3.9, we have information about the value of part A, as it represents the observed outcome for the treated group in t=1 after the treatment. However, when it comes to part B (as defined in Equation 3.9), the path that the treated group would have followed in the absence of treatment is unknown. To make this estimation, we rely on the assumption that this path would be parallel to the trajectory followed by the control group. This assumption is referred to as the Parallel Trend Assumption (PTA). In simpler terms, we assume that the unobserved path taken by the treated group (B) in the scenario where they did not receive treatment is the same as the observed path in the control group (Equation 3.10).

$$E(Y_{i,1}(0) - Y_{i,0}|D_i = 1) = E(Y_{i,1} - Y_{i,0}|D_i = 0) \quad (3.10)$$

Finally, using Equation 3.10 in Equation 3.9, we can construct a feasible estimator for the ATT that will be given by:

$$\widehat{ATT} = [E(Y_{i,1}(0) - Y_{i,0}|D_i = 1)] - [E(Y_{i,1} - Y_{i,0}|D_i = 0)]$$
$$= E\left(Y_{i,1}|D_i = 1\right) - \hat{E}\left(Y_{i,1}(0)|D_i = 1\right) \quad (3.11)$$

Nonetheless, the PTA assumption can be difficult to fulfill in practice, as the treated and control groups may not possess similar characteristics. As such, Callaway and Sant'Anna (2021) have suggested generalizing the canonical approach by including additional groups and fixed effects in the specification. Moreover, DiD designs often feature more than two periods or more than two treated groups, which can further complicate the PTA assumption. To address this issue, Sant'Anna and Zhao (2020) propose using the PTA for groups with identical pre-treatment characteristics; in our case, this is the vector of controls X from Equation 3.2, thereby reducing the risk of bias due to differences between treated and control groups (Equation 3.12). Where $\theta(x)$ is the ΔY_i if there was no treatment conditional to X_{it}. With this new assumption, the new DiD estimator becomes $\widehat{ATT_*}$ (Equation 3.13).

$$E(Y_{i,1}(0) - Y_{i,0}|D_i = 1, X) = E(Y_{i,1} - Y_{i,0}|D_i = 0, X) = \theta(X) \qquad (3.12)$$

$$\widehat{ATT_*} = E(Y_{i,1}|D_i = 1) - \Big[E(Y_{i,0}|D_i = 1) + \widehat{E}(\theta(X)|D_i = 1)\Big] \qquad (3.13)$$

Various approaches have been proposed in the literature to estimate the component $\widehat{E}(\theta(X)|D_i = 1)$ from Equation 3.13. In this chapter, I adopt the Improved Doubly Robust (IMP) estimator proposed by Sant'Anna and Zhao (2020). The IMP estimator employs a two-step procedure. In the first step, the estimator obtains an estimate of $E(\theta(X)|D_i = 1)$ using only control data and without any weighting by substituting the $\theta(x_i)$ with the predicted outcome $\widehat{\theta}(x_i)$.

In the second step, the estimator adds a correction term Λ, which captures the difference between the predicted and the observed outcome in the control group, weighted by the inverse probability of treatment weights. Specifically, this correction term is calculated as the difference between the predicted and observed outcomes in the control group, weighted by the inverse probability of receiving treatment among the treated individuals.

To define Λ, it is necessary to specify first the weight strategy. The process starts by estimating a propensity score using a binomial model, where the dependent variable is $D(X) = \mathit{Attend}(X)$ as a function of the characteristics X and then use the predicted score to estimate the inverse probability weight $\omega(x)$.

$$P(D_i = 1\Big|X) = F(X) \rightarrow \widehat{\pi}(X) = \widehat{F}(X)$$

$$\omega(x_i) = (\widehat{\pi}(x_i))/(1 - \widehat{\pi}(x_i))$$

The detailed formulation of the IMP estimator is provided in Equation 3.14.

$\widehat{ATT_{IMP}} = E(\Delta Y_i|D_i = 1) - E(\widehat{\theta}(x_i)|D_i = 1) - \Lambda$

$where\, \Lambda = E(\omega(x_i)\Delta Y_i|D_i = 0)/E(\omega(x_i)|D_i = 0) - E(\omega(x_i)\widehat{\theta}(x_i)|D_i = 0)/E(\omega(x_i)|D_i = 0)$

$\widehat{ATT}_{imp} = E(\Delta Y_i|D_i = 1) - E(\widehat{\theta}(x_i)|D_i = 1) - [E(\omega(x_i)\Delta Y_i|D_i = 0)/E(\omega(x_i)|D_i = 0)] - \left[E(\omega(x_i)\widehat{\theta}(x_i)|D_i = 0)/E(\omega(x_i)|D_i = 0)\right] \qquad (3.14)$

However, college enrollment spans across multiple cohorts or times of enrollment, so the treatment occurs at different times and among different groups. To address this complex situation, we adopt the framework proposed by Callaway and Sant'Anna (2021), which extends the work of Sant'Anna and Zhao (2020).This proposed estimator effectively circumvents issues related to negative weights or inappropriate comparison groups by focusing solely on DiD designs useful for identifying the ATT.

In this approach, we designate a specific group as "g," representing the cohort of college enrollment, while allowing for temporal variation denoted by "t." This framework enables us to explore how the proposed ATT evolves over time for a given group, as outlined in Equation 3.15.

In our estimation, we utilize the methodology developed by Rios-Avila et al. (2021). This approach dissects the combinations of groups and times into multiple 2X2 models, which are then aggregated based on "g." By following this approach, we can identify ATTs for every treated group "G" and at every time point "t" $ATT(g, t)$.

$$ATT(g,t) = E\left(Y_{i,t} - Y_{i,g-1}|G_i = g\right) - \left(E\left(Y_{i,t}(0) - Y_{i,g-1}|G_i = g\right)\right) \qquad (3.15)$$

After this process, an ATT and weights are calculated for each period group, allowing us to consolidate the ATT by time (similar to an event analysis as we report in the results section), and by group to analyze impacts per group and make comparisons. As mentioned earlier, the groups can have different times, so under this framework, the previously divided population into two groups (treatment and control) is now sorted into three sets: treated, not yet treated, and control.

The final part is to estimate the expected outcome change in absence of treatment. As students may delay their enrollment in higher education and may need to work prior to enrollment, our control group contains those who have not yet been treated. So, we impose conditional PTA for the not yet treated, secondary graduates that are working but were not enrolled in higher education for first time in time “t." The PTA assumption is given by:

$$E\left(Y_{i,t}(0) - Y_{i,g-1} | G_i = g\right) = E\left(Y_{i,t} - Y_{i,g-1} | G_i = 0\, or\, G_i > t\right) \qquad (3.16)$$

So, Equation 3.15 describes the final ATT to be estimated using the PTA for not yet treated.

$$ATT(g,t) = E\left(Y_{i,t} - Y_{i,g-1} | G_i = g\right) - \left(E\left(Y_{i,t} - Y_{i,g-1} | G_i = 0\, or\, G_i > t\right)\right) \qquad (3.17)$$

This framework enables us to estimate the causal impact of college enrollment for each cohort and explore how this impact evolves over time. We utilize the Event aggregation from Rios-Avila et al. (2021) with different values for “e." Event aggregation is defined as:

$$ATT_{Event} = \frac{\sum\limits_{t+e=g} w_{g,t} ATT\left(g,t\right)}{\sum\limits_{t+e=g} w_{g,t}} \qquad (3.18)$$

We employ aggregations for subsets of periods before treatment to facilitate the evaluation of the PTA. Additionally, we consider a subset of periods post-treatment, specifically ranging from 5 to 10 years after treatment. This is of particular significance in our study because the earnings premiums associated with attending college are expected to materialize after individuals have obtained their degrees. Depending on the program, this could take more than three years. Aggregated ATTs for the entire post-treatment period in our panel can be substantially biased downward, considering that we anticipate college students may not work or have a fraction of the income of their peers who did not attend college but are working during their college years (which covers at least half of the timeline in our panel). Furthermore, we implement intervals of one unit to simulate an event study. This approach provides valuable insights into the dynamic nature of the ATT over time, allowing us to observe distinct trajectories for statuses and covariates.

3.5 Results

In this section, we first analyze the outcomes derived from Equation 3.2. Subsequently, we discuss the results obtained through the Panel Ordinary Least Squares (OLS) model, as expanded upon in Equations 3.3 and 3.4. We proceed to present the findings from the Instrumental Variables (IV) approach, specifically focusing on the Local Average Treatment Effects (LATE) estimation as outlined in Equation 3.6. To conclude, we provide graphical representations of the heterogeneous Difference-in-Differences results (ATT estimation) in accordance with Equation 3.18, encompassing both aggregate data and diverse demographic subgroups.

3.5.1 Main Results

The estimated coefficient for "Attend" in Equation 3.2 is 0.063 (Table 3.2, Column 1), suggesting that individuals who attend college earn 6.5% more than those who do not. However, this estimate, albeit similar to the ATT estimated for the 5 to 10-year post-treatment period using Equation 3.18 (Figure 3.1), may suffer from potential biases arising from omitted variable considerations or sample selection issues. To address these concerns, we applied the instrumental variable approach, utilizing the distance from school to college as the instrumental variable in Equation 3.6. The resulting estimated coefficient (LATE) for college attendance, as per this approach, indicates a substantial 38.4% increase in income for individuals who attended college compared to those who completed secondary school without pursuing higher education (32.5% in Ln from Table 3.2, Column 5).

3.5.2 Disaggregated Results

In this section, we present the results for Equation 3.3 and Equation 3.4 in Table 3.2, disaggregated by different status categories. It's important to note that these estimations are not causal but provide valuable descriptive insights into how students who attend college achieve an income 38.4% higher than those who never attended. The panel analysis results can be considered as a conservative lower-bound estimate.

In Table 3.2, specifically in columns (2) and (3), we present the findings on returns to education for students who attended college compared to students who completed secondary school but did not pursue higher education, using Equation 3.3 and Equation 3.4. All regressions shown in Table 3.2 include controls for various factors such as time, secondary graduation cohort, gender, Saber 11 test scores, school sector, household income, age, age squared, experience, experience squared, school shift, apprenticeship, self-employment, public servant status, and the departmental unemployment rate. The complete output can be found in the appendix.

From column (2), the results indicate that the returns to education for students who graduated from college amount to 18.6% (from the 0.171 in Ln). Candidates who attended college exhibit a return to education of 3.8%, suggesting that the Sheepskin Effect, as estimated under this approach, stands at 14.8%. In contrast, students with an incomplete status earn 2.2% less than their peers who did not enroll in higher education. Active students earn an income on par with their peers who did not attend college.

In column (3), we analyze the results for postgraduates and earlier or late incomplete statuses. We find that the returns for students who only hold a bachelor's degree are 17.7%

(from the 0.163 in Ln), while wages for students who earn a Diploma or Master's degree are 64.8% and 67.6%, respectively. Although we included the PhDs graduates in the regression, we do not consider them in the analysis as the secondary school graduates that we track and reach this level are small in number. Conversely, while Early incomplete does not report different earnings from their peers who did not attend college, the late incompletes earn 2.2% less than their peers who did not go to college.

In column (4), we present the results of the first step in the Instrumental Variables (IV) analysis, Equation 3.5, which reveals a significant and negative association between the distance from high school to college. This aligns with the expectation that greater distance from secondary school to college results in reduced exposure and, consequently, a lower probability of attending college. In column (5), we report the Local Average Treatment Effect (LATE) as the coefficient for college attendance estimated from Equation 3.6. These findings indicate that the returns to education are approximately 38.4%. Our Panel analysis results align with figures from existing literature, particularly the 19.94% reported by Tenjo Galarza et al. (2015) for the same time frame. Although our LATE results are somewhat higher, they are more in line with post-college values reported by Hernández (2010).

3.5.3 Heterogeneous Difference in Differences (DiD) Analysis

In this subsection, we aim to analyze the yearly dynamics of the higher education premium since high school graduation. To accomplish this, we will employ the heterogeneous difference-in-differences (DiD) results from Equation 3.16 using the CSDID command by Rios-Avila et al. (2021).Our primary focus will be on the values of T>5, located on the right-hand side of the value 5 on the X-axis. We approximate T=5 as the completion of the academic program. Notably, the values reported during college are comparatively lower than those observed in the control group because the students do not work full-time, as illustrated in Figure 3.1. Additionally, the delay in getting the degree can also be attributed to the students studying and working simultaneously. Estimating a benchmark for comparison is possible, as the starting salary for those without formal sector experience (represented by the zero line) should be equivalent to the minimum wage (234.87 US dollars in 2022). At the same time, the earnings of working students would likely come from part-time or hourly work.

The first part of this section will present the general results, while the second part will report the results by differentiating between three groups of covariates. The first group includes variables collected during the Saber 11 test, such as gender, score, household income, school area, and school sector. The second group includes variables collected during reporting to social security. The third group comprises variables collected during tertiary education, such as program level, higher education institutions' quality, and program area. In the first and second groups, the comparison is made against all individuals within the same groups who were not enrolled in higher education. For the third group, the comparison includes all remaining students, both those who attended college and those who did not. For example, women are compared to women who did not enroll in higher education, while STEM majors are compared to all non-STEM majors, including those who never enrolled in higher education.

In an ideal scenario, college students would not be engaged in work activities before and during their academic program. However, the economic context in Colombia and the need to obtain financial resources to pay for their studies may explain the prevalence of employment with lower income compared with the peers that did not attend college during these periods. A clear pattern emerges from the graph during college, indicating that individuals who work less or are less compelled to do so are more likely to complete their academic program (Figure 3.2).

The Sheepskin Effect, measuring the earnings difference between graduates and candidates (highlighted in red in Figure 3.3), consistently increases after college graduation. On average, the Sheepskin Effect stands at 68.1%, calculated by subtracting the post-college average earnings of candidates (-17.5%) from the post-college average earnings of graduates

(50.6%) in Figure 3.1. Furthermore, the Sheepskin Effect amplifies over time, reaching an average of 119.4% at the 10-year mark after secondary graduation in Figure 3.1. Lastly, as displayed in Figure 3.4, the Sheepskin Effect is more pronounced in females, students from public secondary schools, those with high household income, a high Saber 11 score, residing outside Bogota, and non-self-employed individuals (Figure A3.1 shows the same results than Figure 3.4 but for the 10-year mark).

In the Colombian labor market, graduates who complete their education within the expected timeframe enjoy an average premium of 69.6% during the 5 to 10 years following the freshman year, compared to similarly experienced workers without higher education (Figure 3.5). Moreover, for those who graduate on time, their earnings in the fifth year after enrolling in college are similar to those of workers with five years of experience but no higher education (Figure 3.2 and Figure A3.2). These findings confirm Jaeger and Page (1996) previous research, which suggests that the labor market values academic preparation as much, if not more, than experience in the early stages of the professional career (It is easier to see in Figure A3.2 when lines for graduates cut the zero between year 5 and 6). Returns for individuals who drop out of college are the same, whether they dropped out in their first year, later, or were candidates (Figure A3.3).

The medium to long-term returns on higher education do not report statistically significant differences across fields (STEM or non-STEM) or levels (Apprenticeship, Professional, or Associate Programs) compared to individuals who did not attend college. Notably, students in Apprenticeships experience a rapid recovery in earnings due to their concurrent work and study arrangement, which lasts until year 3. However, after year 7, all program levels demonstrate statistically similar income premiums compared to their non-college-educated peers. There is no statistically significant difference in earnings based on the level of the program. However, graduates from STEM programs tend to experience higher returns, as there is a decline in returns in the medium term for non-STEM graduates (refer to Figure 3.6 and Figure 3.7). Additionally, the quality of higher education institutions significantly influences the economic benefits of attending college. Empirical evidence suggests that students who attended certified institutions report similar earnings in the years following college. Still, those from certified institutions exhibit a more favorable trend and higher returns at year 10 than those who attended non-certified institutions (see Figure 3.8).

The results from Table 3.2 indicate the presence of a gender wage gap, with men earning approximately 20% more than women. However, the premium for higher education is notably higher for women compared to men, especially in the early stages of their careers (Figure 3.9). Interestingly, the long-term returns for males resemble the early premiums for females, with females experiencing a substantial increase in their premiums by year 10 (reaching 162.1%, as seen in Figure 3.8, and Figure A3.1). In contrast, males experience similar or even lower returns than their peers who did not attend college until year 9 (Figure 3.9). Moreover, the study reveals that the Sheepskin Effect, which represents the income increase associated with completing a college degree, is more pronounced for women than men. This can be attributed to the fact that even women who did not complete their college degree tend to have a higher premium than men who did complete their degree (Figure 3.4 and Figure 3.9). Conversely, men who graduated from college tended to earn the same as their peers who did not attend college, while male candidates exhibited negative returns (Figure 3.4, Figure 3.9, and Figure A3.1).

Figure 3.4, Figure 3.10, and Figure A3.5 examine the relationship between Saber 11 test scores and earnings, they show that graduates with high scores experience a premium of 114.9% a decade after completing their enrollment in higher education. Conversely, those graduates with low scores receive a premium of 113.7% (Figure A3.1). Additionally, the Sheepskin Effect is more pronounced for high-skilled students, with a Sheepskin Effect of 77.8% for high-skilled individuals and 61.3% for low-skilled individuals in the years post-college (Figure 3.4 and Figure 3.10).

Students with high household incomes who attended college but did not obtain a degree consistently earn significantly less than their peers who did not attend college, and their premium is higher than those with low incomes. This phenomenon may be attributed to social influence, where unsuccessful students with high household incomes can access preferred job markets (Figure 3.4 and Figure A3.1). We also found that household income has a less significant impact on future premiums than academic skill-based income, and an even less significant impact than gender-based income differences (Compare Figure A3.4, Figure A3.5, and Figure A3.6).

Figure 3.11 highlights that graduates from private secondary schools report significantly higher incomes than those from public secondary schools who attended college. This suggests that social connections or other external factors may influence students' long-term outcomes from private secondary schools. The Sheepskin Effect is higher for graduates from public secondary schools, reaching 70.2%, while private secondary school graduates have a Sheepskin Effect of 64.1% (Figure 3.4). Figure 3.12 shows that the gap in income due to sector of high school remains into college, even after graduation. However, college graduates have a narrower gap than secondary school graduates, showing that graduation from public colleges improves social mobility but perhaps not enough to compensate initial differences.

Figure 3.13 shows no significant difference in the premium among students who did not complete their college program, regardless of the region, until year 8. However, graduates from Bogota experience a faster increase in their premium than those outside the region, although their premium becomes lower than that of graduates from other regions 10 years after starting college. Consequently, the Sheepskin Effect is 55.8% for students from Bogota and 70.9% for the rest of the country (Figure 3.4, Figure 3.13, and Figure A3.1). This can be explained by graduates from outside Bogota potentially migrating to regions with higher income opportunities. At the same time, candidates in Bogota continue to earn the same as their peers who did not attend college. Finally, self-employees report a stagnation, as their income is the same if they graduate or drop out after completing 90% of their coursework (Figure 3.4).

3.6 Discussion

This chapter provides valuable insights into the benefits of pursuing higher education in Colombia. The findings reveal a positive impact on future formal sector income for college graduates compared to their peers who did not attend college. Additionally, the study highlights significant premiums for individuals with high cognitive skills but notes a stagnation of self-employed individuals in the labor market.

These findings raise several questions, such as whether self-employed workers underreport their actual wages and only report the minimum required for health and pension systems or whether the labor market for contractual workers is undervaluing the added value of a college education. Furthermore, it is unclear whether self-employed workers in the formal market are considered part of the informal sector or whether they are informal workers in a formal market. This issue warrants further research.

Our research highlights the crucial role of higher education in Colombia in mitigating educational class segregation and reducing the gender gap. It underscores the need to explore further the intricate relationship between the labor market, education, and its potential impact on future income.

In summary, the study emphasizes the significance of obtaining a higher education degree, particularly on-time, as it increases income levels across diverse socioeconomic backgrounds. Colombia must increase the college graduation rate, especially for on-time graduation, to enhance income opportunities for individuals, irrespective of their socioeconomic background. Increasing college graduation rates will also open opportunities for further academic achievement at the master's and Ph.D. levels, which have demonstrated important improvements in students' future income, albeit with limited data available to track their progress.

To maximize income opportunities, Colombia should prioritize improving the quality and alignment of community college and apprenticeship programs with the labor market. These programs offer returns comparable to professional programs in the medium and long term compared to the outcomes of individuals who did not attend college. Moreover, community college and apprenticeship programs present an attractive opportunity for individuals seeking to enhance their future income, given their relatively short duration, low opportunity cost, high demand, and favorable return on investment.

3.7 Conclusion

Based on the results presented in this chapter, attending higher education can be a life-changing experience in terms of income, especially for those who obtain their degree on time. Furthermore, it represents a significant improvement for female secondary graduates, or graduates with low household incomes or who attended public schools. While it is evident that the higher education system suffers from socioeconomic segregation, the results show that graduating from public colleges reduces the income gap between socioeconomic classes.

The study reveals a consistent increase in the Sheepskin Effect after graduation from college, with higher effects observed for females, individuals with higher income levels, those with better academic skills, those who attended secondary schools outside Bogota, and those who attended private schools. The Sheepskin Effect reaches an average of 119.4% at year 10 after secondary school graduation (Figure 3.3 and Figure A3.1), highlighting the long-term benefits of getting a degree from higher education in Colombia.

The Colombian labor market rewards individuals with a college degree the same as workers who possess five or six years of experience but lack higher education, depending on if the graduation was on time or not. These findings support the research of Jaeger and Page (1996), suggesting that the labor market values academic preparation as much, if not more, than experience in the early stages of a professional career. Interestingly, regardless of the timing of dropout, there is no difference in returns for those who drop out of college. Surprisingly, there is no difference in returns for individuals who drop out of college, regardless of when they do so. This highlights the urgency of reducing dropout rates, and it is better if it is by increasing graduation rates, as dropout students face unfavorable financial prospects, with costs incurred and income comparable to or lower than their peers who did not attend college.

The findings indicate that attending college can lead to modest yet redistributive income improvements, especially for socially disadvantaged individuals. It also contributes to narrowing the gender gap, particularly for women who graduate from college. Low-income students and those from public secondary schools benefit significantly from college education. Regarding gender, females experience a more substantial income premium than their non-college-educated peers, suggesting a gradual reduction in the gender gap. Household income is a critical factor in determining future income premiums, with higher-income students consistently reporting higher premiums. This highlights potential barriers low-income students face in fully realizing the benefits of higher education, becoming an interesting topic for further research on strategies to address these disparities. The Sheepskin Effect remains robust over time, with greater premiums observed for females, high academic skilled students, graduates from private secondary schools or secondary schools outside Bogota.

The returns on higher education in the medium to long term are positive and consistent across program levels, including apprenticeship, professional, or associate programs, compared to individuals who did not attend college. This finding diverges from earlier studies by González-Velosa et al. (2015)and Busso et al. (2020), reported negative premiums for associate degrees. However, this discrepancy can be attributed to our control group, which, in our case, consisted of secondary school graduates who attended community colleges. While individuals pursuing associate programs or apprenticeships may fare better than their peers, their income improvements might not be as substantial as those pursuing professional programs, which constitute most of the system, and can explain the difference with previous literature. Our results highlight the positive impact of higher education on career prospects, characterized by the "Stairway to Heaven" effect of college graduation. Graduates enjoy a substantial increase in income compared to those who do not attend college. However, a significant "Highway to Hell" effect is observed for individuals who drop out or remain in the system for extended periods, incurring increasing costs without reaping the benefits. Considering these results, policymakers should focus on designing policies that encourage higher education and support timely graduation to improve career outcomes for graduates. By prioritizing educational attainment and reducing barriers to graduation, policymakers can help to ensure that all students, regardless of their income or background, have the opportunity to succeed in the labor market.

In conclusion, the findings presented in this chapter have important implications for Colombia regarding reducing barriers to graduation and promoting the benefits of higher education. This can lead to a more highly trained and skilled workforce, which in turn can promote economic growth, reduce the gender gap, and encourage social mobility. Policymakers should therefore, focus on designing policies that encourage higher education and support timely graduation to improve career outcomes beyond the success reported by Ferreyra et al. (2017) and Ministerio de Educación Nacional (2017). In addition, higher education institutions should evaluate how their degree requirements affect their students and take action to reduce barriers to graduation. Obtaining a degree not only sends signals to the labor market and increases students' knowledge and training but also opens the door to new levels of education that can further empower students and contribute to a more prosperous and equitable society.

Table 3.1: Descriptive Statistics for Secondary Graduates

Variable	Mean (1)	Standard Deviation (2)	Did not attend college mean (3)	Attended college group difference (4)
Log Annual income	7.03	7.13	7.043	-.0371***
Attended college	23.93	42.66		
of wich:				
Graduated	6.36	24.40		
Graduated on time	4.62	20.99		
Graduated late	1.74	13.06		
Only bachelor degree	6.21	24.13		
With advanced studies	0.15	3.86		
Diploma	0.14	3.72		
Master	0.01	0.89		
PhD	0.00	0.53		
Dropout	17.03	37.59		
Candidate	1.44	11.92		
Incomplete	15.59	36.27		
Incomplete 1st year	1.71	12.96		
Incomplete after 1st year	13.88	34.57		
Active	0.54	7.33		
Age	22.02	6.13	21.86	.6874***
Age^2	522.54	289.43	515.8	28.33***
Experience	1.64	2.12	1.651	-.0257***
Experience^2	7.20	12.28	7.239	-.1704***
Unemployment rate	11.94	2.75	11.96	-.0555***
Age at the Saber 11 test	19.86	3.96	19.94	-.3052***
Female	41.83	49.33	41.54	1.22***
Saber 11 test score	40.66	26.49	37.04	15.14***
From a public high school	66.21	47.30	66.44	-9.85***
Self-employed	8.32	20.93	7.612	2.954***
Public servant	0.60	5.86	.5763	.1071***
Distance to HEI (in km)	0.18	0.53	.1864	-.0434***
Income categories in mmw				
[0,1)	4.64	21.03	4.663	-.1083
[1,2)	29.04	45.40	29.34	-1.241***
[2,3)	42.92	49.50	42.88	.1694
[3,5)	11.89	32.36	11.81	.3419***
[5,7)	6.71	25.02	6.586	.5161***
[7,9)	1.72	13.02	1.702	.097**
[9,11)	0.98	9.84	.9562	.0888**
[11,13)	1.84	13.44	1.804	.1481***
[13,15)	0.08	2.86	.0822	-.0025
15 and more	0.17	4.18	.1772	-.0092
School shift categories				
Full day	22.75	41.92	22.67	.3422**
Morning	43.68	49.60	43.68	.0165
Afternoon	18.42	38.76	18.44	-.1061
Evening	8.06	27.23	8.074	-.0478
Weekend	2.26	14.87	2.278	-.0726
Other	4.83	21.44	4.861	-.1323*
Observations	5,111,158			
Individuals	393,166			

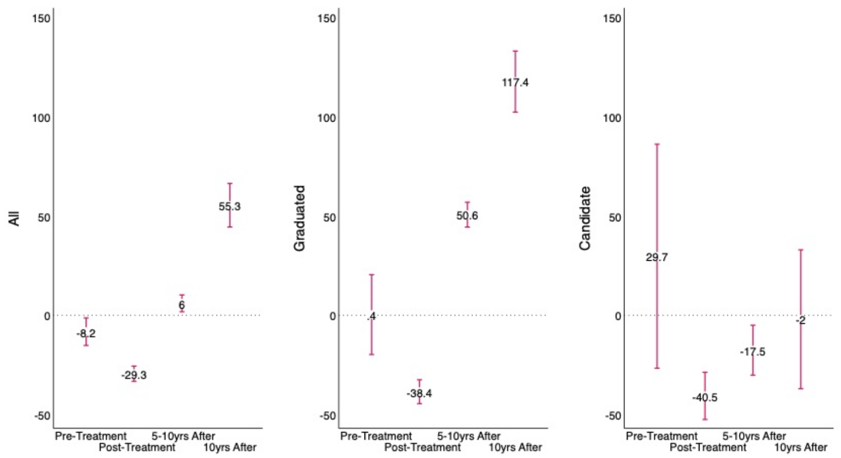
Note: Table shows the mean, standard deviation, minimum, and maximum for the main characteristics of college attendants by their status in tertiary system. Variables in percent, Age and Experience in years, Distance in Kilometers.

Table 3.2: Main Results

	Panel A. Panel Regressions			Pane B. IV Regressions	
	General	Status	Detailed Status	First Step	Second Step
	Log Annual Income			Attend	Log Annual Income
	(1)	(2)	(3)	(4)	(5)
Attend or $\widehat{Attend}$	0.063***				0.325*
	(0.007)				(0.180)
Graduated		0.171***			
		(0.007)			
Bachelor			0.163***		
			(0.008)		
Diploma			0.500***		
			(0.049)		
Master			0.517**		
			(0.227)		
PhD			0.332		
			(0.314)		
Candidate		0.038***	0.038***		
		(0.014)	(0.014)		
Dropout		-0.022***			
		(0.005)			
Dropout Early			-0.016		
			(0.012)		
Dropout Late			-0.023***		
			(0.005)		
Active		0.027	0.027		
		(0.022)	(0.022)		
Distance to HEI (in km)				-0.017***	
				(0.001)	
Apprenticeship	-0.106***	-0.107***	-0.107***	-0.134***	-0.071***
	(0.006)	(0.006)	(0.006)	(0.001)	(0.025)
Observations	5,111,158	5,111,158	5,111,158	5,111,158	5,111,158
Number of id	393,166	393,166	393,166	393,166	393,166
Overall R^2	0.832	0.832	0.832	0.143	0.832
χ^2 p-value	0	0	0	0	0

Notes: The table shows the coefficients of the regressions corresponding to Equation 3.1 and Equation 3.6. Non-shown regression controls include: unemployment rate, age, age squared, experience, experience squared, sex, Saber 11 test score, and dummies for public sector high school, self employe, public servant, household income level, and high school shift. Full regression can be found in the Appendix. In Columns (1), (2), (3), and (5), the dependent variable is expressed in logarithm. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

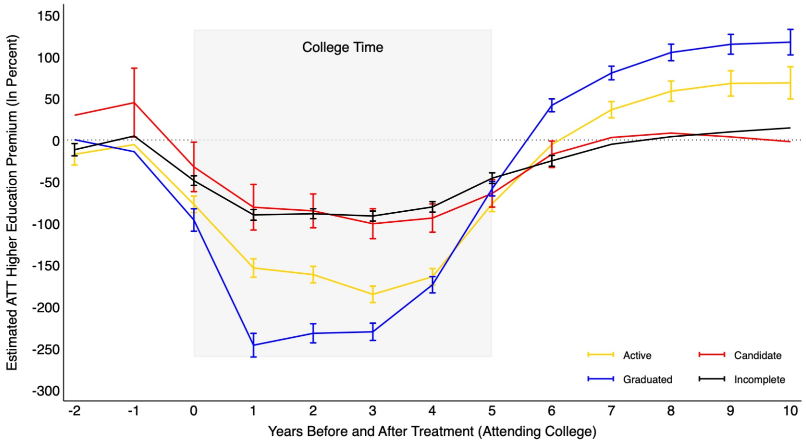
Figure 3.1: Estimated Aggregated ATT in Higher Education Comparing All and Sheepskin Effect



Improved Doubly Robust Estimator IMP - Sant'Anna & Zhao (2020)

Notes: The Figure shows the estimated ATT coefficients obtained through the modified Mincer regression (Equation 3.1) that calculates the returns to education. The treated group consists of individuals who attended college and is further divided into three subgroups: all, graduates, and candidates. Sheepskin effect would be the difference between the coefficient for Graduates and Candidates after treatment. The control group comprises secondary school students who have not yet enrolled in higher education (dotted horizontal line in Y=0). The X-axis presents the pre treatment average, the post treatment average, and the average 10 years after treatment. The whiskers depict the 95 percent confidence intervals. The premiums presented in the Figure 3.1 are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 3.18) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) IMP estimator (Equation 3.14). The model controls for academic skills, household income, gender, program level, program type, and quality of higher education institutions. For further details on the variables used, readers can refer to Table 3.1.

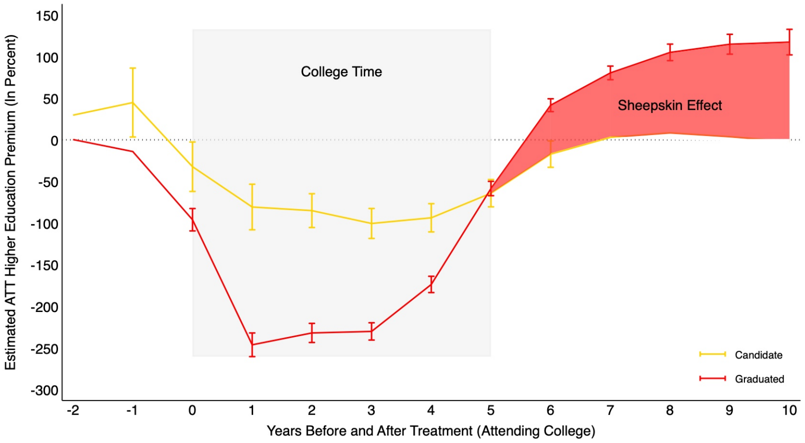
Figure 3.2: Higher Education Premium in Colombia



Improved Doubly Robust Estimator IMP - Sant'Anna & Zhao (2020)
Confidence interval was removed if P-value>0.05

Notes: The Figure shows the estimated ATT coefficients obtained through the modified Mincer regression (Equation 3.1) that calculates the returns to education. The treated group consists of individuals who attended college and is further divided into four subgroups: graduates, candidates, incomplete, and active students. The control group comprises secondary school students who have not yet enrolled in higher education (dotted horizontal line in Y=0). The X-axis represents the years before and after enrollment in higher education, while the whiskers depict the 95 percent confidence intervals. In Colombia, students who discontinue their studies are categorized as incomplete or candidates. The shadowed area indicates the expected duration for completing a professional degree, which is five years after commencing as freshmen, whereas an associate program typically lasts three years. The premiums presented in the Figure 3.2 are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 3.18) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) IMP estimator (Equation 3.14). The model controls for academic skills, household income, gender, program level, program type, and quality of higher education institutions. For further details on the variables used, readers can refer to Table 3.1.

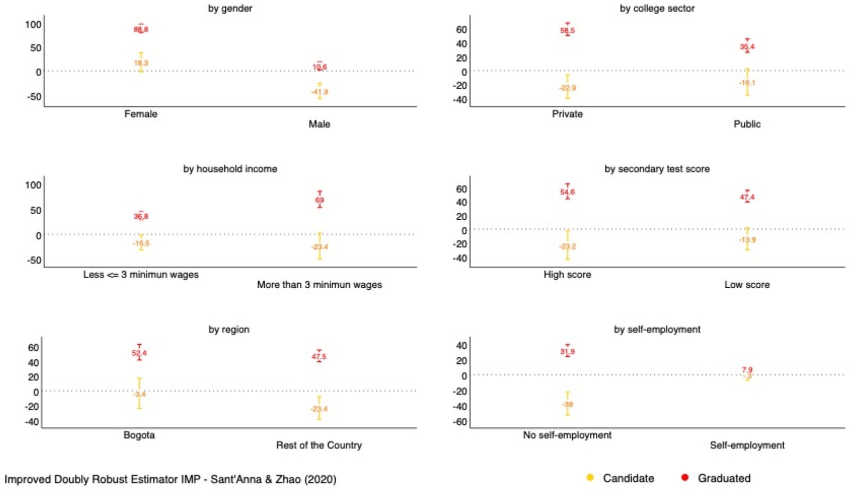
Figure 3.3: Higher Education Degree Sheepskin Effect



Improved Doubly Robust Estimator IMP - Sant'Anna & Zhao (2020)
Confidence interval was removed if P-value>0.05

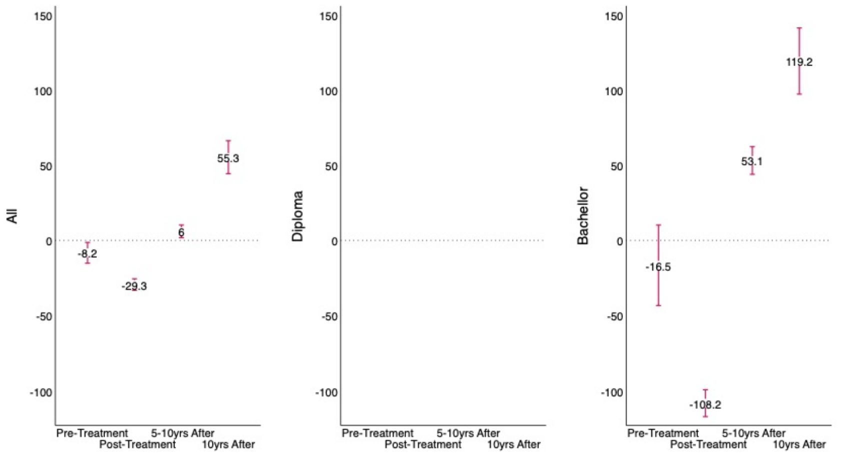
Notes: The Figure shows the estimated ATT coefficients obtained through the modified Mincer regression (Equation 3.1) that calculates the returns to education. The treated group consists of individuals who attended college and is further divided into two subgroups to get the sheepskin effect: graduates and candidates. Sheepskin effect as the area created by the difference between the Graduates and the Candidates. The control group comprises secondary school students who have not yet enrolled in higher education (dotted horizontal line in Y=0). The X-axis represents the years before and after enrollment in higher education, while the whiskers depict the 95 percent confidence intervals. In Colombia, students who discontinue their studies are categorized as incomplete or candidates. The shadowed area indicates the expected duration for completing a professional degree, which is five years after commencing as freshmen, whereas an associate program typically lasts three years. The premiums presented in the Figure 3.3 are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 3.18) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) IMP estimator (Equation 3.14). The model controls for academic skills, household income, gender, program level, program type, and quality of higher education institutions. For further details on the variables used, readers can refer to Table 3.1.

Figure 3.4: Sheepskin Effect by Characteristics 5 to 10 Years After Treatment



Notes: The figure shows the estimated ATT coefficients obtained through Equation 3.2 that calculates the returns to education. The treated group consists of individuals who attended college and is further divided into two subgroups: graduates and candidates. The whiskers depict the 95 percent confidence intervals. The premiums presented in the figure are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 3.18) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) IMP estimator (Equation 3.14). For further details on the variables used as controls, readers can refer to Table 3.1.

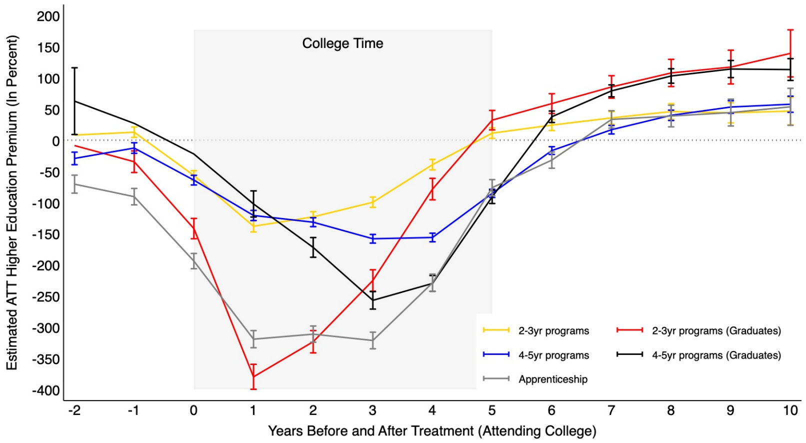
Figure 3.5: Summary of Aggregated ATT for Graduates



Improved Doubly Robust Estimator IMP - Sant'Anna & Zhao (2020)

Notes: The figure shows the estimated ATT coefficients obtained through Equation 3.2 that calculates the returns to education. The treated group consists of individuals who attended college and is further divided into three subgroups: total, graduates late and on time. The control group comprises secondary school students who have not yet enrolled in higher education (dotted horizontal line in Y=0). The X-axis represents the years before and after attending college, while the whiskers depict the 95 percent confidence intervals. The premiums presented in the figure are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 3.18) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) IMP estimator (Equation 3.14). For further details on the variables used as controls, readers can refer to Table 3.1.

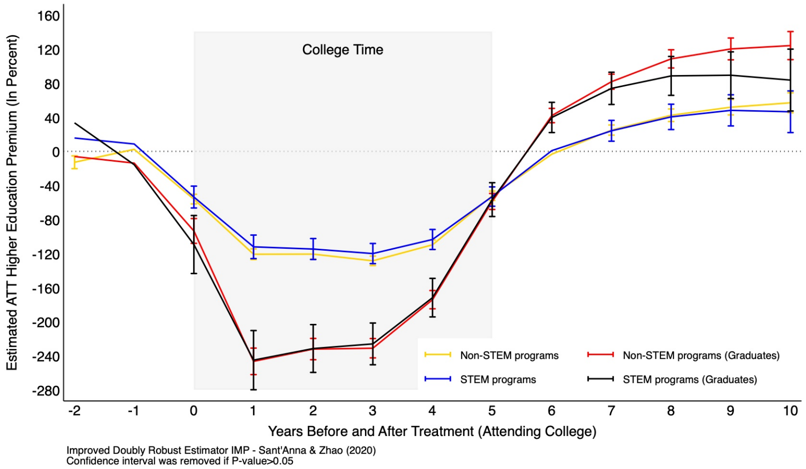
Figure 3.6: Evolution of ATT by Level of Program



Improved Doubly Robust Estimator IMP - Sant'Anna & Zhao (2020)
Confidence interval was removed if P-value>0.05

Notes: The figure shows the estimated ATT coefficients obtained through Equation 3.2 that calculates the returns to education. The treated group consists of individuals who attended college and is further divided into five subgroups: Apprenticeships, Bachelor and Associate programs students, and Bachelor and Associate program graduates. The control group comprises secondary school students who have not yet enrolled in higher education (dotted horizontal line in Y=0). The X-axis represents the years before and after attending college, while the whiskers depict the 95 percent confidence intervals. The shadowed area indicates the expected duration for completing a bachelor's degree, which is five years after commencing as freshmen, whereas an associate program typically lasts three years. The premiums presented in the figure are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 3.18) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) IMP estimator (Equation 3.14). For further details on the variables used as controls, readers can refer to Table 3.1.

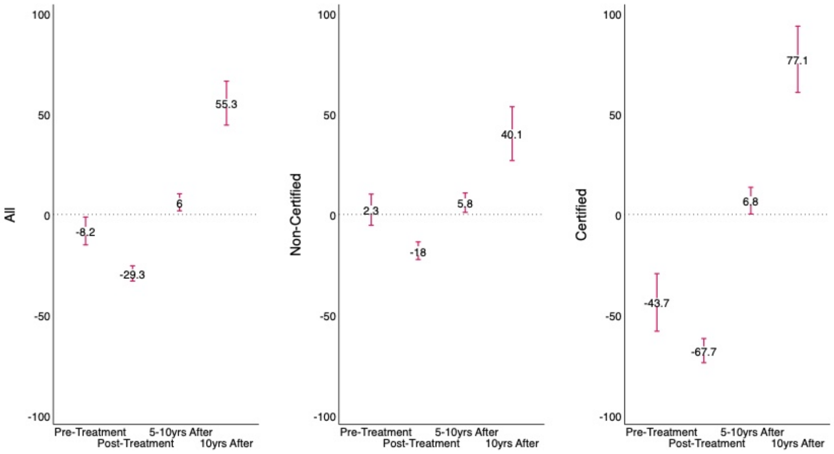
Figure 3.7: Evolution of ATT By Field of the Program



Improved Doubly Robust Estimator IMP - Sant'Anna & Zhao (2020)
Confidence interval was removed if P-value>0.05

Notes: The figure shows the estimated ATT coefficients obtained through Equation 3.2 that calculates the returns to education. The treated group consists of individuals who attended college and is further divided into four subgroups: STEM and non STEM students, and STEM and non STEM graduates. The control group comprises secondary school students who have not yet enrolled in higher education (dotted horizontal line in Y=0). The X-axis represents the years before and after attending college, while the whiskers depict the 95 percent confidence intervals. The shadowed area indicates the expected duration for completing a bachelor's degree, which is five years after commencing as freshmen, whereas an associate program typically lasts three years. The premiums presented in the figure are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 3.18) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) IMP estimator (Equation 3.14). For further details on the variables used as controls, readers can refer to Table 3.1.

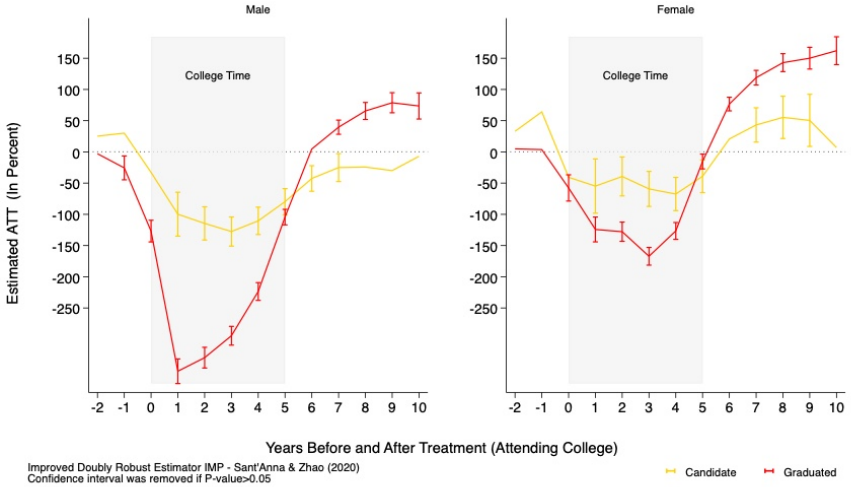
Figure 3.8: Summary of Aggregated Students in Certificate and Non-Certificate HEIs



Improved Doubly Robust Estimator IMP - Sant'Anna & Zhao (2020)

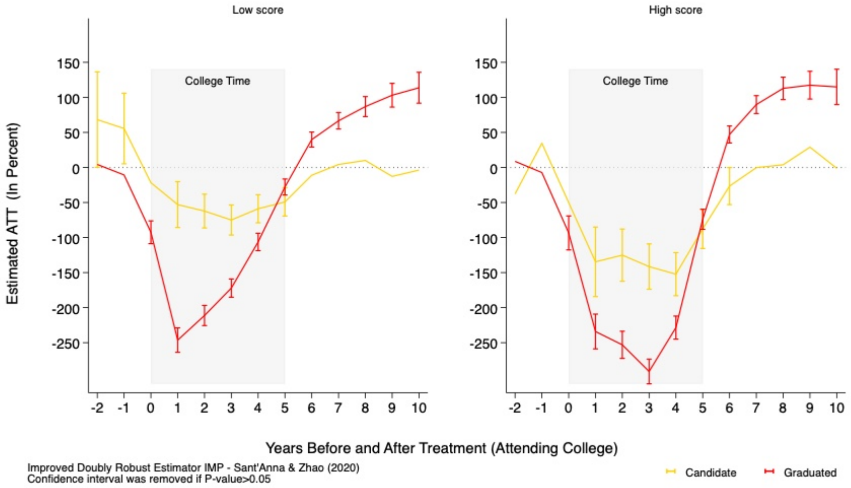
Notes: The figure shows the estimated ATT coefficients obtained through Equation 3.2 that calculates the returns to education. The treated group consists of individuals who attended college and is further divided into two subgroups: College students enrolled in Certified and Non Certified HEIs. The control group comprises secondary school students who have not yet enrolled in higher education (dotted horizontal line in Y=0). The X-axis represents the years before and after attending college, while the whiskers depict the 95 percent confidence intervals. The shadowed area indicates the expected duration for completing a bachelor's degree, which is five years after commencing as freshmen, whereas an associate program typically lasts three years. The premiums presented in the figure are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 3.18) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) IMP estimator (Equation 3.14). For further details on the variables used as controls, readers can refer to Table 3.1.

Figure 3.9: Evolution of ATT for Graduates and Candidates by Gender



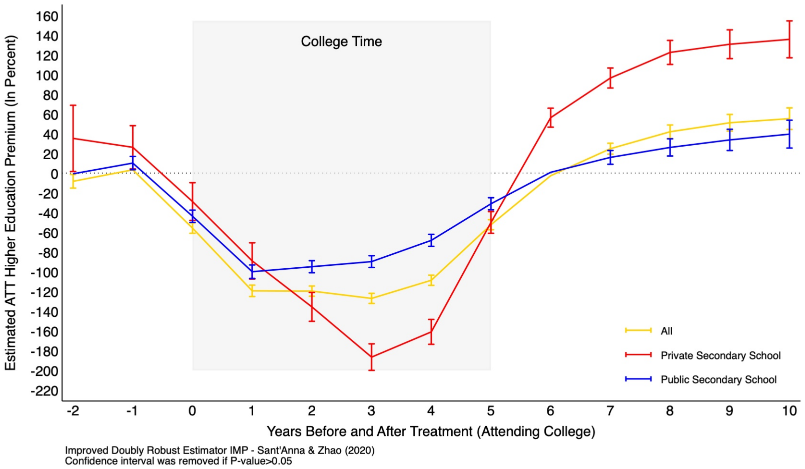
Notes: The figure shows the estimated ATT coefficients obtained through Equation 3.2 that calculates the returns to education. The treated group consists of individuals who attended college and is further divided into two subgroups: Candidates, and Incompletes. The control group comprises secondary school students who have not yet enrolled in higher education (dotted horizontal line in Y=0). The X-axis represents the years before and after attending college, while the whiskers depict the 95 percent confidence intervals. The shadowed area indicates the expected duration for completing a bachelor's degree, which is five years after commencing as freshmen, whereas an associate program typically lasts three years. The premiums presented in the figure are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 3.18) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) IMP estimator (Equation 3.14). For further details on the variables used as controls, readers can refer to Table 3.1.

Figure 3.10: Evolution of ATT for Graduates and Candidates by Saber 11 Test Score



Notes: The figure shows the estimated ATT coefficients obtained through Equation 3.2 that calculates the returns to education. The treated group consists of individuals who attended college and is further divided into two subgroups: Candidates, and Incompletes. The control group comprises secondary school students who have not yet enrolled in higher education (dotted horizontal line in Y=0). The X-axis represents the years before and after attending college, while the whiskers depict the 95 percent confidence intervals. The shadowed area indicates the expected duration for completing a bachelor's degree, which is five years after commencing as freshmen, whereas an associate program typically lasts three years. The premiums presented in the figure are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 3.18) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) IMP estimator (Equation 3.14). For further details on the variables used as controls, readers can refer to Table 3.1.

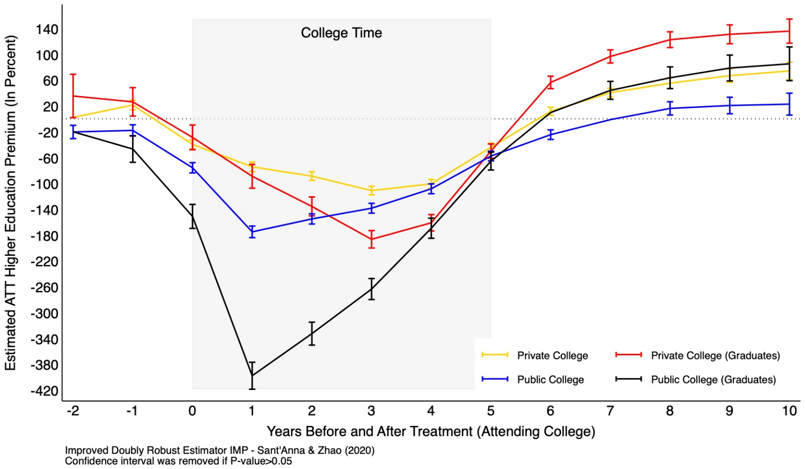
Figure 3.11: Evolution of ATT by Secondary School Sector



Improved Doubly Robust Estimator IMP - Sant'Anna & Zhao (2020)
Confidence interval was removed if P-value>0.05

Notes: The figure shows the estimated ATT coefficients obtained through Equation 3.2 that calculates the returns to education. The treated group consists of individuals who attended college and is further divided into two subgroups: Public and Private Secondary Graduates. The control group comprises secondary school students who have not yet enrolled in higher education (dotted horizontal line in Y=0). The X-axis represents the years before and after attending college, while the whiskers depict the 95 percent confidence intervals. The shadowed area indicates the expected duration for completing a bachelor's degree, which is five years after commencing as freshmen, whereas an associate program typically lasts three years. The premiums presented in the figure are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 3.18) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) IMP estimator (Equation 3.14). For further details on the variables used as controls, readers can refer to Table 3.1.

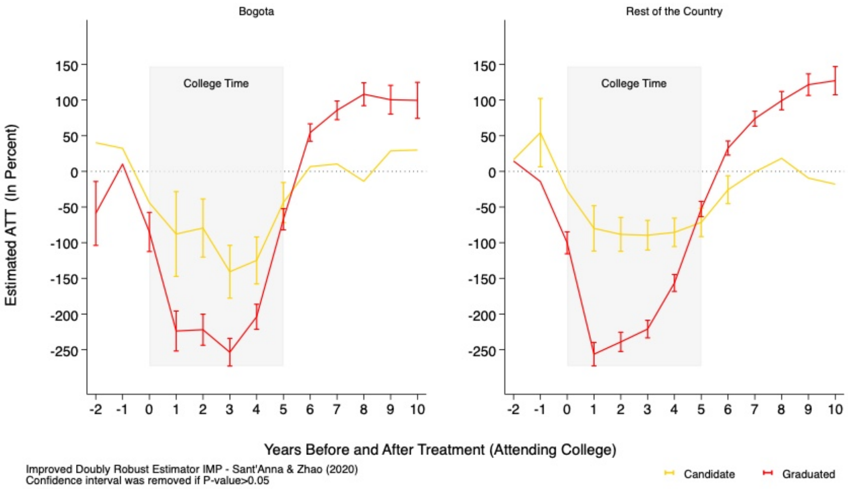
Figure 3.12: Evolution of ATT by College Sector



Improved Doubly Robust Estimator IMP - Sant'Anna & Zhao (2020)
Confidence interval was removed if P-value>0.05

Notes: The figure shows the estimated ATT coefficients obtained through Equation 3.2 that calculates the returns to education. The treated group consists of individuals who attended college and is further divided into four subgroups: Public and Private college students and Public and Private college graduates. The control group comprises secondary school students who have not yet enrolled in higher education (dotted horizontal line in Y=0). The X-axis represents the years before and after attending college, while the whiskers depict the 95 percent confidence intervals. The shadowed area indicates the expected duration for completing a bachelor's degree, which is five years after commencing as freshmen, whereas an associate program typically lasts three years. The premiums presented in the figure are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 3.18) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) IMP estimator (Equation 3.14). For further details on the variables used as controls, readers can refer to Table 3.1.

Figure 3.13: Evolution of ATT for Graduates and Candidates by Region



Bogota

Rest of the Country

Improved Doubly Robust Estimator IMP - Sant'Anna & Zhao (2020)
Confidence interval was removed if P-value>0.05

Notes: The figure shows the estimated ATT coefficients obtained through Equation 3.2 that calculates the returns to education. The treated group consists of individuals who attended college and is further divided into two subgroups: Candidates, and Incompletes. The control group comprises secondary school students who have not yet enrolled in higher education (dotted horizontal line in Y=0). The X-axis represents the years before and after attending college, while the whiskers depict the 95 percent confidence intervals. The shadowed area indicates the expected duration for completing a bachelor's degree, which is five years after commencing as freshmen, whereas an associate program typically lasts three years. The premiums presented in the figure are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 3.18) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) IMP estimator (Equation 3.14). For further details on the variables used as controls, readers can refer to Table 3.1.

References

Arango, Luis Eduardo, Carlos Esteban Posada, and José Darío Uribe, "Cambios en la estructura de los salarios urbanos en Colombia, 1984-2000," *Lecturas de Economia*, 2005, *63* (2), 7–39.

Arias, Helmut and Álvaro Chávez, "Cálculo de la tasa interna de retorno de la educación en Colombia," 2002.

Arrow, Kenneth, "Higher Education as filter," *Journal of Public Economics*, 1973, *2*, 193–216.

Bacolod, Marigee, Jorge De la Roca, and María Marta Ferreyra, "In search of better opportunities: Sorting and agglomeration effects among young college graduates in Colombia," *Regional Science and Urban Economics*, mar 2021, *87*, 103656.

Bauer, Thomas, Patrick Dross, and John Haisken-DeNew, "Sheepskin effects in Japan," *International Journal of Manpower*, jun 2005, *26* (4), 320–335.

Becker, Gary, "Investment in Human Capital: A Theoretical Analysis," *Journal of Political Economy*, oct 1962, *70* (5, Part 2), 9–49.

Belman, Dale and John Heywood, "Sheepskin Effects in the Returns to Education: An Examination of Women and Minorities," *The Review of Economics and Statistics*, nov 1991, *73* (4), 720–724.

_ **and** _ , "Sheepskin Effects by Cohort: Implications of Job Matching in a Signaling Model," *Oxford Economic Papers*, oct 1997, *49* (4), 623–637.

Bilkic, Natasa, Thomas Gries, and Margarethe Pilichowski, "Stay in school or start working? - The human capital investment decision under uncertainty and irreversibility," *Labour Economics*, oct 2012, *19* (5), 706–717.

Binelli, Chiara, "Returns to Education and Increasing Wage Inequality in Latin America," 2008, pp. 1–54.

Busso, Matías, Juan Sebastián Muñoz, and Sebastián Montaño, "Unbundling Returns to Skills: Evidence from Postsecondary Education in Colombia," 2020.

Callaway, Brantly and Pedro H.C. Sant'Anna, "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

Calonico, Sebastian and Hugo Ñopo, "Where did you go to school? Private-Public differences in schooling trajectories and their role on earnings," *Well-Being and Social Policy*, 2007, *3* (2002), 25–46.

Card, David, "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," 1993.

_ , "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, sep 2001, *69* (5), 1127–1160.

Cárdenas, Mauricio and Raquel Bernal, "Changes in the distribution of income and the new economic model in Colombia," *Cepal Naciones Unidas*, 1999, (002116).

Collins, Randall, *The Credential Society: An Historical Sociology of Education and Stratification.*, New York: Academic Press, 1979.

Crespo, Anna and Maurício Cortez Reis, "Sheepskin effects and the relationship between earnings and education: analyzing their evolution over time in Brazil," *Revista Brasileira de Economia*, sep 2009, *63* (3), 209–231.

DANE, "DANE," 2023.

de la universidad colombiana, El Observatorio, "El observatorio de la universidad colombiana," 2020.

de Roux, Nicolás and Evan Riehl, "Disrupted academic careers: The returns to time off after high school," *Journal of Development Economics*, may 2022, *156*, 102824.

Dougherty, Christopher, "El Futuro de la Educación Colombiana: Proyecciones y Prioridades," *Revista de Planeación y Desarrollo*, 1971, *3* (1), 1–149.

Duflo, Esther, "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment," *American Economic Review*, 2001, *91* (4), 795–813.

Epple, Dennis, Richard Romano, and Holger Sieg, "Admission, tuition, and financial aid policies in the market for higher education," *Econometrica*, 2006, *74* (4), 885–928.

Farber, Henry S and Robert Gibbons, "Learning and Wage Dynamics," *The Quarterly Journal of Economics*, nov 1996, *111* (4), 1007–1047.

Farné, Stefano and Carlos Andrés Vergara, "El Mercado de Trabajo de los Profesionales Colombianos," *Observatorio del Mercado de Trabajo y la Seguridad Social*, 2006, *015934.*

Ferrer, Ana M. and W. Craig Riddell, "The role of credentials in the Canadian labour market," *Canadian Journal of Economics/Revue Canadienne d'Economique*, nov 2002, *35* (4), 879–905.

Ferreyra, María Marta, Andrea Franco Hernández, Tatiana Melguizo, and Angélica María Sánchez Díaz, "Estimating the Contribution of Short-Cycle Programs to Student Outcomes in Colombia," 2020.

Ferreyra, Maria Marta, Ciro Avitabile, Javier Botero Álvarez, Francisco Haimovich Paz, and Sergio Urzúa, *At a Crossroads: Higher Education in Latin America and the Caribbean*, World Bank, Washington, DC, may 2017.

Fields, Gary, "Educación y Movilidad Económica en Colombia," Technical Report, Universidad de los Andes, Bogota 1977.

Forero, Nohora and Manuel Ramirez-Gomez, "Determinants of Earnings of College Graduates in 2001-2004 (Determinantes De Los Ingresos Laborales De Los Graduados Universitarios Durante El Período 2001-2004)," 2008.

Frenette, Marc, "Access to college and university: Does distance to school matter?," *Canadian Public Policy*, 2004, *30* (4), 427–442.

García-Suaza, Andrés Felipe, Juan Carlos Guataquí, José Alberto Guerra, and Darío Maldonado, "Beyond the Mincer equation: the internal rate of return to higher education in Colombia," *Education Economics*, may 2014, *22* (3), 328–344.

Gibson, John, "Sheepskin effects and the returns to education in New Zealand: Do they differ by ethnic groups?," *New Zealand Economic Papers*, 2000, *34* (2), 201–220.

González-Velosa, Carolina, Graciana Rucci, Miguel Sarzosa, and Sergio Urzúa, "Returns to higher education in Chile and Colombia," 2015.

Hernández, Gustavo, "Cuán rentable es la educación superior en Colombia?," *Lecturas de Economia*, 2010, *73* (julio-diciembre), 181–214.

Herrera-Prada, Luis Omar, "Determinantes de la tasa de graduación y graduación a tiempo en la educación superior de Colombia 1998-2010," *Coyuntura Económica*, 2013, *XLIII* (1), 143–177.

_ **and Carlos Caballero**, "El financiamiento de la educación superior en Colombia," in "La educación superior:Retos y perspectivas" 2013, chapter 3, pp. 121–175.

Hungerford, Thomas and Gary Solon, "Sheepskin Effects in the Returns to Education," *The Review of Economics and Statistics*, feb 1987, *69* (1), 175–177.

Jaeger, David A. and Marianne E. Page, "Degrees Matter: New Evidence on Sheepskin Effects in the Returns to Education," *The Review of Economics and Statistics*, nov 1996, *78* (4), 733–740.

Kugler, Bernardo, "Influencia de la Educación en los Ingresos de Trabajo: El Caso Colombiano," *Revista de Planeación y Desarrollo*, 1974, *6* (2), 53–74.

_ **, Álvaro Reyes, and Martha de Gómez**, "Educación y Mercado de Trabajo Urbano en Colombia: Una Comparación entre Sectores Modernos y no Modernos," *Corporación Centro Regional de Población*, 1979, (Monografía 10).

MacLeod, William Bentley, Evan Riehl, Juan Saavedra, and Miguel Urquiola, "The big sort: College reputation and labor market outcomes," *American Economic Journal: Applied Economics*, 2017, *9* (3), 223–261.

Manacorda, Marco, Carolina Sánchez-Páramo, and Norbert R. Schady, "Changes in Returns to Education in Latin America : The Role of Demand and Supply of Skills," *Industrial and Labor Relations Review*, 2007, *63* (2), 307–326.

Melguizo, Tatiana and Jacques Wainer, "Toward a set of measures of student learning outcomes in higher education: evidence from Brazil," *Higher Education*, 2016, *72*, 381–401.

Mesa, Fernando and Javier Gutiérrez, "Efectos de la apertura económica en el mercado laboral industrial," *Revista de Planeación y Desarrollo*, 1996, *27* (4), 14–45.

Mincer, Jacob, *Schooling, Experience, and Earnings*, NBER, 1974.

Ministerio de Educación Nacional, "Análisis de determinantes de la deserción en la educación superior colombiana con base en el SPADIES.," Technical Report, Ministerio de Educación Nacional - Universidad de los Andes, Bogotá 2008.

_ , "La Revolución Educativa 2002 - 2010. Informe de gestión.," Technical Report, Bogotá 2010.

_ , "Boletín Educación Superior," Technical Report, Bogotá 2017.

Mogollon Plazas, Monica and Christian Posso, "Ticket to the middle class? Long-Term Effects of Public Universities on Labor Market and Financial Outcomes," 2022.

Mora, Jhon James, "Sheepskin effects and screening in Colombia," *Colombian Economic Journal*, 2003, *1* (1), 95–108.

_ **and Juan Muro**, "Sheepskin effects by cohorts in Colombia," *International Journal of Manpower*, may 2008, *29* (2), 111–121.

Núñez, Jairo and Fabio Sánchez, "Educación y Salarios relativos en Colombia, 1976-1995. Determinantes, Evolución e Implicaciones para la Distribución del Ingreso," 1998.

OECD, "Self-employment rate (indicator)," 2021.

Olfindo, Rosechin, "Diploma as signal? Estimating sheepskin effects in the Philippines," *International Journal of Educational Development*, may 2018, *60*, 113–119.

Orozco Silva, Luis Enrique, *La Política de Cobertura: eje de la revolución educativa, 2002-2008.*, Bogotá: Ediciones Uniandes, 2010.

_ **, Alberto Roa Valero, and Luis Carlos Castillo Gómez**, "La Educación Superior en Colombia," Technical Report 2011.

Patrinos, Harry Anthony and George Psacharopoulos, "Returns to education in developing countries," in "The Economics of Education" 2020, chapter 4, pp. 53–64.

Peet, Evan D., Günther Fink, and Wafaie Fawzi, "Returns to education in developing countries: Evidence from the living standards and measurement study surveys," *Economics of Education Review*, 2015, *49*, 69–90.

Phelps, Edmund S, "The Statistical theory of Racism and Sexism," *American Economic Review*, 1972, *62* (4), 659–661.

Posso, Christian, "Desigualdad salarial en Colombia 1984-2005: cambios en la composición del mercado laboral y retornos a la educación post-secundaria," *Borradores de Economía*, 2008, (529), 1–31.

Prada, Carlos Felipe, "¿Es rentable la decisión de estudiar en Colombia?," *Ensayos sobre Política Económica*, jun 2006, (51), 226–323.

Psacharopoulos, George, "Returns to Education: A Further International Update and Implications," *Journal of Human Resources*, 1985, *20* (4), 583–604.

_ , "Returns to Investment in Education. A Global Update," *World development*, 1994, *22* (9), 1325–1343.

_ **and Harry Anthony Patrinos**, "Returns to investment in education: a further update," *Education Economics*, aug 2004, *12* (2), 111–134.

_ **and** _ , "Returns to investment in education: a decennial review of the global literature," *Education Economics*, sep 2018, *26* (5), 445–458.

Rios-Avila, Fernando, Brantly Callaway, and Pedro H.C. Sant'Anna, "csdid: Difference-in-Differences with Multiple Time Periods in Stata," 2021.

Rodríguez, Eduardo, "Rentabilidad y crecimiento de la educación superior en Colombia 1971- 1978," 1981.

Santamaría, Mauricio, *External Trade, Skill Technology and the Recent Increase of Income Inequality in Colombia* 2001.

Sant'Anna, Pedro H.C. and Jun Zhao, "Doubly robust difference-in-differences estimators," *Journal of Econometrics*, nov 2020, *219* (1), 101–122.

Schady, Norbert R., "Convexity and Sheepskin Effects in the Human Capital Earnings Function: Recent Evidence for Filipino Men," *Oxford Bulletin of Economics and Statistics*, may 2003, *65* (2), 171–196.

Selowsky, Marcelo, "El Efecto del Desempleo y el Crecimiento sobre la Rentabilidad de la Inversión Educacional: Una Aplicación a Colombia," *Revista del Departamento Nacional de Planeación*, 1969, *1* (2), 5–68.

Shabbir, Tayybb, “Sheepskin Effects in the Returns to Education in a Developing Country,” *The Pakistan Development Review*, mar 1991, *30* (1), 1–19.

Son, Hyelim, “Human Capital Investment When Sheepskin Effects Matter : Evidence from Income Shocks in Indonesia,” 2013.

Spence, Michael, “Job Market Signaling,” *The Quarterly Journal of Economics*, aug 1973, *87* (3), 355–374.

Tenjo Galarza, Jaime, “Evolución de los retornos a la inversión en educación 1976- 1989.” PhD dissertation 1993.

_ , Oriana Álvarez Vos, Alejandro Gaviria Jaramillo, and María Jiménez, “Evolution of returns to education in Colombia (1976-2014) Programa de Economía Documentos de Trabajo,” 2015, p. 46.

Urrutia, Miguel, “La distribución del ingreso y la distribución de la educación: el sector financiero y la distribución del ingreso.,” Technical Report, Fedesarrollo, Bogota 1974.

Wood, Tom, “The Sheepskin Effect,” Technical Report July 2009.

Yunus, Norhanishah Mohamad, “Sheepskin effects in the returns to higher education: New evidence for Malaysia,” *Asian Academy of Management Journal*, 2017, *22* (1), 151–182.

Zárate, Héctor Manuel, “Cambios en la estructura salarial: una historia desde la regresión cuanfílica,” *Monetaria*, 2005, *XXVIII* (4), 339–364.

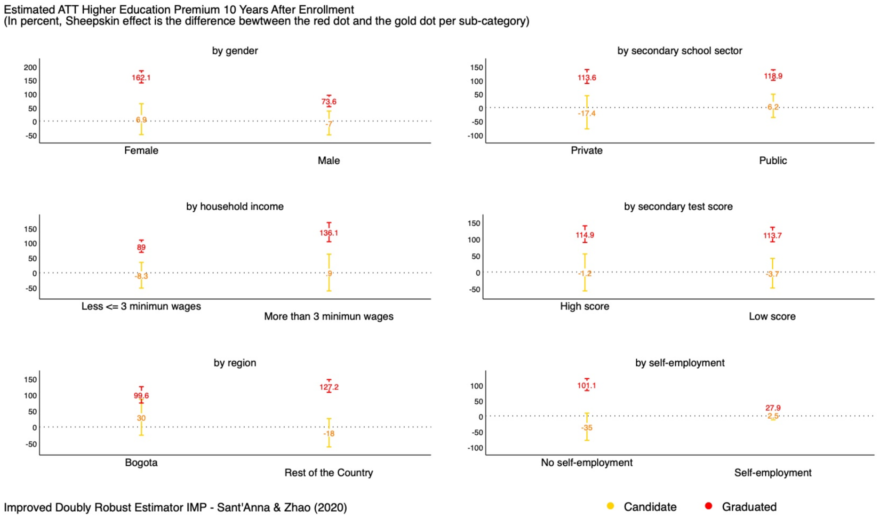
3.8 Appendix

Table A3.1: Descriptive Statistics for Secondary Graduates that Attended College or Apprenticeship

Variable	Mean (1)	Std Dev. (2)	Min (3)	Max (4)
Enrolled in higher education	23.93	42.66	0	100
...of which did enroll in...				
3 or less semesters after secondary graduation	11.71	32.16	0	100
a STEM program	7.26	25.95	0	100
a Certify HEI	6.78	25.15	0	100
a 4yr program	13.04	33.67	0	100
Enrolled in a SENA apprenticeship	8.86	28.42	0	100
Observations	5,111,158			
Individuals	393,166			

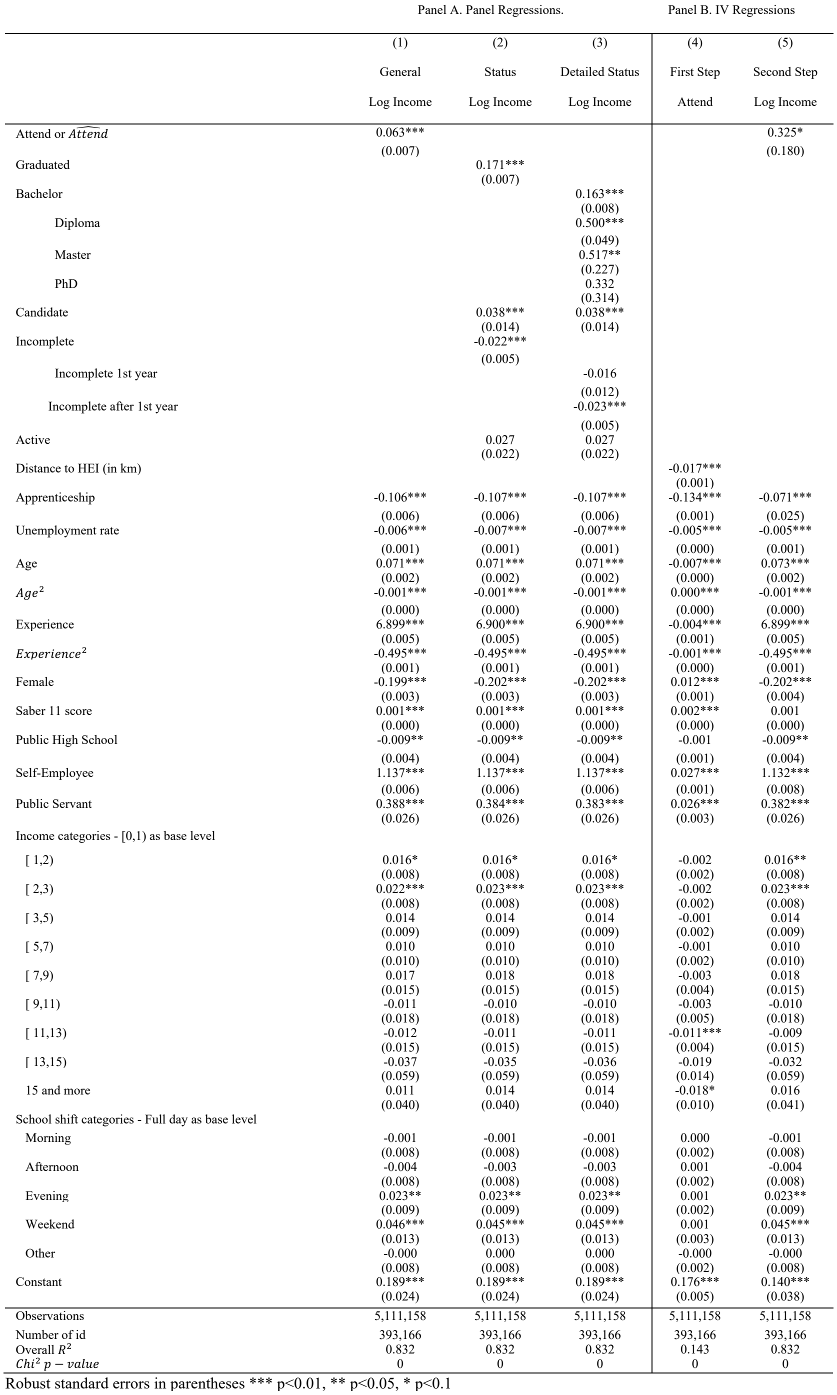
Note: Table shows the mean, standard deviation, minimum, and maximum for the main characteristics of college attendants.

Figure A3.1: Sheepskin Effect by Characteristics 10 Years After Treatment



Notes: The figure shows the estimated ATT coefficients obtained through Equation 3.2 that calculates the returns to education. The treated group consists of individuals who attended college and is further divided into two subgroups: graduates and candidates. The whiskers depict the 95 percent confidence intervals. The premiums presented in the figure are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 3.16) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) IMP estimator (Equation 3.13). For further details on the variables used as controls, readers can refer to Table 3.1.

Table A3.2: Main Results -Full Output-

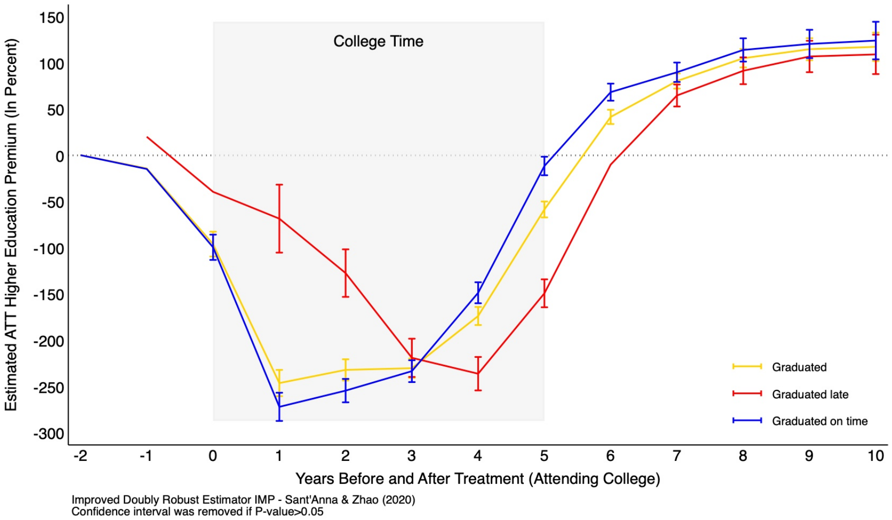


	Panel A. Panel Regressions.			Panel B. IV Regressions	
	(1)	(2)	(3)	(4)	(5)
	General	Status	Detailed Status	First Step	Second Step
	Log Income	Log Income	Log Income	Attend	Log Income
Attend or $\widehat{Attend}$	0.063***				0.325*
	(0.007)				(0.180)
Graduated		0.171***			
		(0.007)			
Bachelor			0.163***		
			(0.008)		
Diploma			0.500***		
			(0.049)		
Master			0.517**		
			(0.227)		
PhD			0.332		
			(0.314)		
Candidate		0.038***	0.038***		
		(0.014)	(0.014)		
Incomplete		-0.022***			
		(0.005)			
Incomplete 1st year			-0.016		
			(0.012)		
Incomplete after 1st year			-0.023***		
			(0.005)		
Active		0.027	0.027		
		(0.022)	(0.022)		
Distance to HEI (in km)				-0.017***	
				(0.001)	
Apprenticeship	-0.106***	-0.107***	-0.107***	-0.134***	-0.071***
	(0.006)	(0.006)	(0.006)	(0.001)	(0.025)
Unemployment rate	-0.006***	-0.007***	-0.007***	-0.005***	-0.005***
	(0.001)	(0.001)	(0.001)	(0.000)	(0.001)
Age	0.071***	0.071***	0.071***	-0.007***	0.073***
	(0.002)	(0.002)	(0.002)	(0.000)	(0.002)
Age^2	-0.001***	-0.001***	-0.001***	0.000***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Experience	6.899***	6.900***	6.900***	-0.004***	6.899***
	(0.005)	(0.005)	(0.005)	(0.001)	(0.005)
$Experience^2$	-0.495***	-0.495***	-0.495***	-0.001***	-0.495***
	(0.001)	(0.001)	(0.001)	(0.000)	(0.001)
Female	-0.199***	-0.202***	-0.202***	0.012***	-0.202***
	(0.003)	(0.003)	(0.003)	(0.001)	(0.004)
Saber 11 score	0.001***	0.001***	0.001***	0.002***	0.001
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Public High School	-0.009**	-0.009**	-0.009**	-0.001	-0.009**
	(0.004)	(0.004)	(0.004)	(0.001)	(0.004)
Self-Employee	1.137***	1.137***	1.137***	0.027***	1.132***
	(0.006)	(0.006)	(0.006)	(0.001)	(0.008)
Public Servant	0.388***	0.384***	0.383***	0.026***	0.382***
	(0.026)	(0.026)	(0.026)	(0.003)	(0.026)
Income categories - [0,1) as base level					
[1,2)	0.016*	0.016*	0.016*	-0.002	0.016**
	(0.008)	(0.008)	(0.008)	(0.002)	(0.008)
[2,3)	0.022***	0.023***	0.023***	-0.002	0.023***
	(0.008)	(0.008)	(0.008)	(0.002)	(0.008)
[3,5)	0.014	0.014	0.014	-0.001	0.014
	(0.009)	(0.009)	(0.009)	(0.002)	(0.009)
[5,7)	0.010	0.010	0.010	-0.001	0.010
	(0.010)	(0.010)	(0.010)	(0.002)	(0.010)
[7,9)	0.017	0.018	0.018	-0.003	0.018
	(0.015)	(0.015)	(0.015)	(0.004)	(0.015)
[9,11)	-0.011	-0.010	-0.010	-0.003	-0.010
	(0.018)	(0.018)	(0.018)	(0.005)	(0.018)
[11,13)	-0.012	-0.011	-0.011	-0.011***	-0.009
	(0.015)	(0.015)	(0.015)	(0.004)	(0.015)
[13,15)	-0.037	-0.035	-0.036	-0.019	-0.032
	(0.059)	(0.059)	(0.059)	(0.014)	(0.059)
15 and more	0.011	0.014	0.014	-0.018*	0.016
	(0.040)	(0.040)	(0.040)	(0.010)	(0.041)
School shift categories - Full day as base level					
Morning	-0.001	-0.001	-0.001	0.000	-0.001
	(0.008)	(0.008)	(0.008)	(0.002)	(0.008)
Afternoon	-0.004	-0.003	-0.003	0.001	-0.004
	(0.008)	(0.008)	(0.008)	(0.002)	(0.008)
Evening	0.023**	0.023**	0.023**	0.001	0.023**
	(0.009)	(0.009)	(0.009)	(0.002)	(0.009)
Weekend	0.046***	0.045***	0.045***	0.001	0.045***
	(0.013)	(0.013)	(0.013)	(0.003)	(0.013)
Other	-0.000	0.000	0.000	-0.000	-0.000
	(0.008)	(0.008)	(0.008)	(0.002)	(0.008)
Constant	0.189***	0.189***	0.189***	0.176***	0.140***
	(0.024)	(0.024)	(0.024)	(0.005)	(0.038)
Observations	5,111,158	5,111,158	5,111,158	5,111,158	5,111,158
Number of id	393,166	393,166	393,166	393,166	393,166
Overall R^2	0.832	0.832	0.832	0.143	0.832
$Chi^2\ p-value$	0	0	0	0	0

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

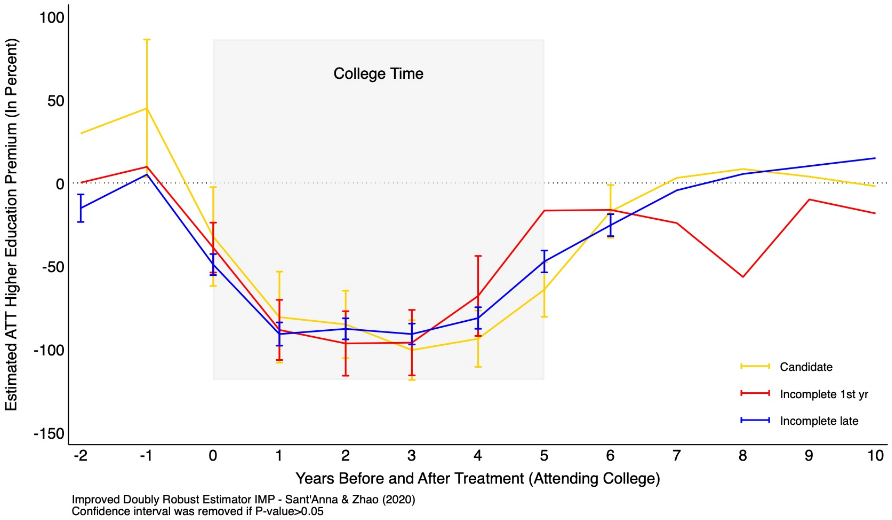
Notes: The table shows the coefficients of the regressions corresponding to Equation 3.1 and Equation 3.6. Unemployment rate in percent; age, age squared, experience and experience squared in years; Saber 11 test score is the percentile. Dummies for females, public sector high school, self employe, public servant, household income level, and high school shift. In Columns (1), (2), (3), and (5), the dependent variable is expressed in logarithm.

Figure A3.2: Evolution of ATT for Graduates



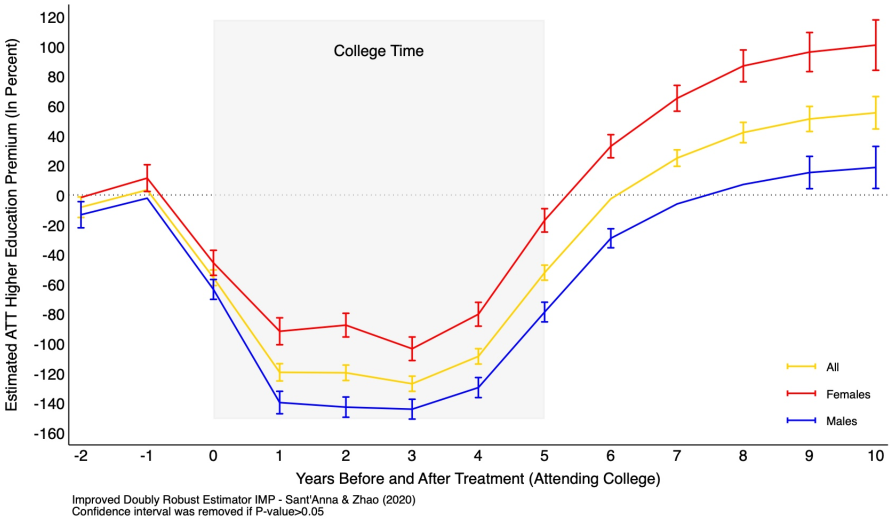
Notes: The figure shows the estimated ATT coefficients obtained through Equation 3.2 that calculates the returns to education. The treated group consists of individuals who attended college and is further divided into two subgroups: Graduated late and on time. The control group comprises secondary school students who have not yet enrolled in higher education (dotted horizontal line in Y=0). The X-axis represents the years before and after attending college, while the whiskers depict the 95 percent confidence intervals. The shadowed area indicates the expected duration for completing a bachelor's degree, which is five years after commencing as freshmen, whereas an associate program typically lasts three years. The premiums presented in the figure are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 3.16) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) IMP estimator (Equation 3.13). For further details on the variables used as controls, readers can refer to Table 3.1.

Figure A3.3: Evolution of ATT for Incompletes



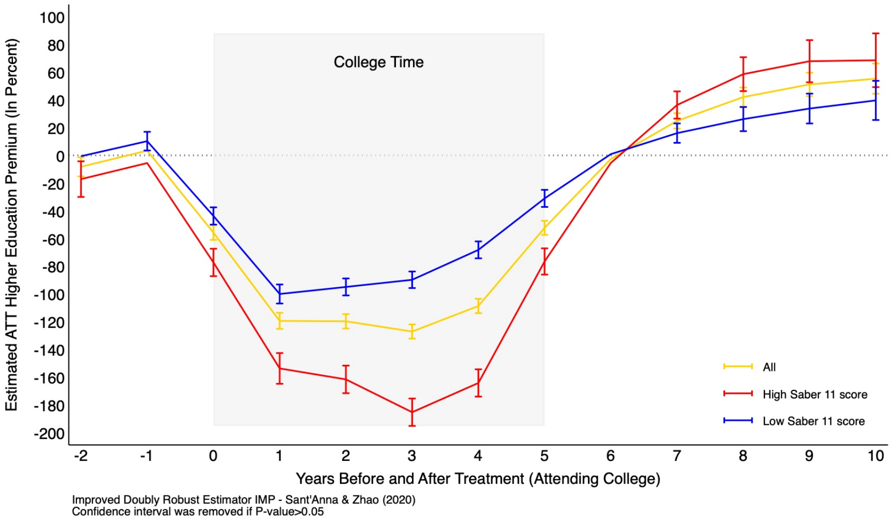
Notes: The figure shows the estimated ATT coefficients obtained through Equation 3.2 that calculates the returns to education. The treated group consists of individuals who attended college and is further divided into two subgroups: Incompletes in the first year or Incompletes after 1st year. The control group comprises secondary school students who have not yet enrolled in higher education (dotted horizontal line in Y=0). The X-axis represents the years before and after attending college, while the whiskers depict the 95 percent confidence intervals. The shadowed area indicates the expected duration for completing a bachelor's degree, which is five years after commencing as freshmen, whereas an associate program typically lasts three years. The premiums presented in the figure are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 3.16) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) IMP estimator (Equation 3.13). For further details on the variables used as controls, readers can refer to Table 3.1.

Figure A3.4: Evolution of ATT by Gender



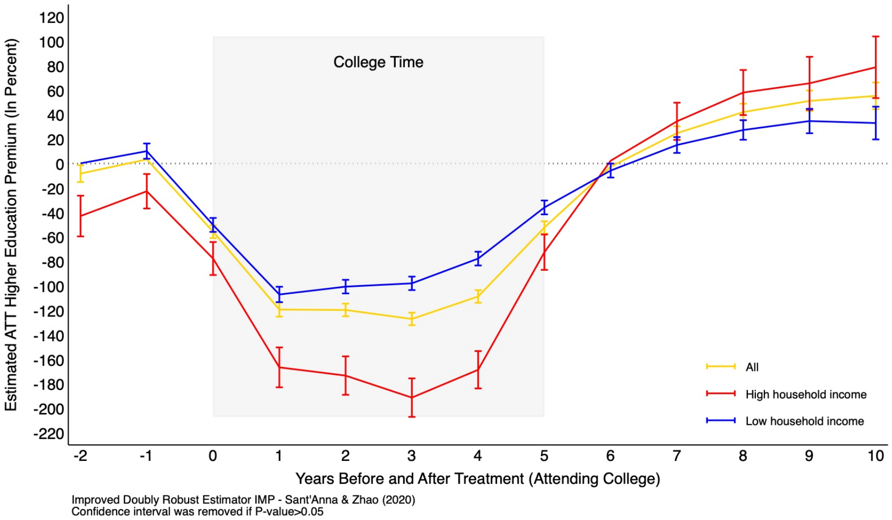
Notes: The figure shows the estimated ATT coefficients obtained through Equation 3.2 that calculates the returns to education. The treated group consists of individuals who attended college and is further divided into two subgroups: Males and Females. The control group comprises secondary school students who have not yet enrolled in higher education (dotted horizontal line in Y=0). The X-axis represents the years before and after attending college, while the whiskers depict the 95 percent confidence intervals. The shadowed area indicates the expected duration for completing a bachelor's degree, which is five years after commencing as freshmen, whereas an associate program typically lasts three years. The premiums presented in the figure are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 3.16) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) IMP estimator (Equation 3.13). For further details on the variables used as controls, readers can refer to Table 3.1.

Figure A3.5: Evolution of ATT by Secondary Test Score



Notes: The figure shows the estimated ATT coefficients obtained through Equation 3.2 that calculates the returns to education. The treated group consists of individuals who attended college and is further divided into two subgroups: High and Low Saber 11 scores. The control group comprises secondary school students who have not yet enrolled in higher education (dotted horizontal line in Y=0). The X-axis represents the years before and after attending college, while the whiskers depict the 95 percent confidence intervals. The shadowed area indicates the expected duration for completing a bachelor's degree, which is five years after commencing as freshmen, whereas an associate program typically lasts three years. The premiums presented in the figure are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 3.16) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) IMP estimator (Equation 3.13). For further details on the variables used as controls, readers can refer to Table 3.1.

Figure A3.6: Evolution of ATT by Household Income Level



Notes: The figure shows the estimated ATT coefficients obtained through Equation 3.2 that calculates the returns to education. The treated group consists of individuals who attended college and is further divided into two subgroups: High and Low household income secondary graduates. The control group comprises secondary school students who have not yet enrolled in higher education (dotted horizontal line in Y=0). The X-axis represents the years before and after attending college, while the whiskers depict the 95 percent confidence intervals. The shadowed area indicates the expected duration for completing a bachelor's degree, which is five years after commencing as freshmen, whereas an associate program typically lasts three years. The premiums presented in the figure are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 3.16) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) IMP estimator (Equation 3.13). For further details on the variables used as controls, readers can refer to Table 3.1.

Abstract Chapter 4

EN

This paper examines the effects of mineral extraction on human capital formation in Colombia, a country rich in natural resources but struggling with low college attendance, high youth unemployment, and high informality in the labor market. Leveraging the allocation of natural resources as a quasi-experimental setting, we link administrative data for 5.5 million secondary school graduates from 2002 to 2014 with information on legal mines in 2014 based on the distances from their respective schools. Employing Instrumental Variables and a Differences-in-Differences approach, we identify treated individuals as graduates from secondary schools located closest to operational mines at the time of their graduation. The findings indicate that active mines positively influence school cohort sizes, student academic performance, and enrollment in higher education. However, they also negatively impact entry into the formal labor market, particularly for roles associated with extractive industries. Substantial heterogeneity exists in the outcomes associated with the various extracted products, leading to the identification of distinct categories: "good mines" and "bad mines."

DE

In diesem Beitrag werden die Auswirkungen des Abbaus von Bodenschätzen auf die Bildung von Humankapital in Kolumbien untersucht, einem Land, das reich an natürlichen Ressourcen ist, aber mit einer niedrigen Schulbesuchsquote, hoher Jugendarbeitslosigkeit und hoher Informalität auf dem Arbeitsmarkt zu kämpfen hat. Wir nutzen die Zuteilung natürlicher Ressourcen als quasi-experimentellen Rahmen und verknüpfen administrative Daten für 5,5 Millionen Sekundarschulabsolventen von 2002 bis 2014 mit Informationen über legale Minen im Jahr 2014 auf der Grundlage der Entfernungen von ihren jeweiligen Schulen. Mithilfe von Instrumentalvariablen und einees Differenzen-in-Differenzen-Ansatzes identifizieren wir die behandelten Personen als Absolventen von Sekundarschulen, die zum

Zeitpunkt ihres Abschlusses am nächsten zu aktiven Minen lagen. Die Ergebnisse deuten darauf hin, dass aktive Minen einen positiven Einfluss auf die Größe der Schulkohorte, die akademischen Leistungen der Schüler und die Einschreibung in höhere Bildungswege haben. Sie wirken sich jedoch auch negativ auf den Eintritt in den formellen Arbeitsmarkt aus, insbesondere auf Positionen in der mineralgewinnenden Industrie. Die Ergebnisse im Zusammenhang mit den verschiedenen abgebauten Produkten sind sehr heterogen, was zur Identifizierung verschiedener Kategorien führt: "gute Bergwerke" und "schlechte Bergwerke".

Chapter 4

The Impact of Mining on Educational and Labor Market Outcomes in Colombia

4.1 Introduction

This chapter contributes to the literature by separating the regional effects within political boundaries from the effects caused by the mine and extending results to other materials. To date, the literature has made progress in measuring the impacts of mines on academic performance and other labor market outputs but has not been able to separate the effect of the region. It also complements Bonilla Mejía (2020) on exploiting labor market mechanisms and the distances between mines and schools. Since mines leave royalties and generate policies evidenced in the academic results, the transmission mechanism is mainly the region and not necessarily by the exposure of the school to mine in operation. we address this by controlling for political boundaries and the influence of the mine on the school, depending on the distance between the school and the mine and the number of mines around the school. In this way, we allow a mine located in a specific region to have effects on a school located in another, separating the regional effects from the direct influence of the mine on the student's academic results.

Latin America and the Caribbean (LAC) is a world region with abundant natural resources and complex institutional, political, financial, economic, and social characteristics that have impacted its economic development. In Colombia, the exploitation of oil and minerals is a significant part of the main economic activities: more than 33.2% of its exports are from oil and 21.6% from minerals, according to the National Department of Statistics (DANE).

The literature regarding the effect of natural resource exploitation from mining activities on economic development indicators is broad. This explains why governments concentrate their investments in this activity instead of promoting other sectors (Gylfason, 2001). Several papers have suggested strategies or found interesting results on the consequences of exploiting natural resources for social dynamics (Bonilla Mejía, 2020; van der Ploeg, 2011), economic growth (Barro, 2001; Martínez Ortiz and Aguilar, 2012) , poverty (Litschig and Morrison, 2013; Loayza and Rigolini, 2016; Pegg, 2006), government efficiency in the provision of public goods (Angrist and Kugler, 2008; Caselli and Michaels, 2013; Loayza et al., 2014; Martínez, 2023), and quality of education (Agüero et al., 2016; Álvarez and Vergara, 2022; Hanushek and Woessmann, 2010). However, it is essential to discuss the relevance of the geographical location where the activity is carried out and the impact generated by its exposure to the people (Torvik, 2002). In the literature, the results regarding educational outcomes are more related to pollution-related school absences due to cognitive skill impairments during "the school stage" (Almond et al., 2009; Currie et al., 2009; Park et al., 2002).

Using administrative data for Colombia, instrumental variables (IV) and differences in differences approaches, this chapter provides evidence of the link between mining and some educational and labor market outputs in Colombia. we exploit the mines' locations as a natural experiment and the operation of the closest mine to each school in Colombia as the treatment for the nation's secondary graduates from 2002 to 2012. we use the universe of secondary school graduates in Colombia from 2002 to 2012 merged with information from the Ministry of Education and Ministry of Health to mark those who went to college and/or the formal labor market, respectively. All legal mines and schools were geo-localized, creating a matrix of distances to assign the closest mine's information to each school. The students who graduated from a secondary school whose nearest mine was operating during their graduation year are the treated group. Assigning the treatment by using the closest mine to each school allows us to avoid the correlation between the municipalities' information and their outcomes. It is vital to separate the effects of the mines and the regional budget; the closest mine can be crossing the border of the country's domestic political division.

In line with Angrist and Kugler (2008)'s results, we find that the size of cohorts increases approximately 6% (LATE) and 7.2% (ATT) when the nearest mine is in operation. The Saber 11 test score increases if the nearest mine is in operation by 3.87 points (an improvement of 8.2% compared to the control mean), reaching up to 17 points 9 years after the opening of the mine. we also found positive results for the probability of enrolling in higher education; if the nearest mine is in operation, the likelihood of enrolling in higher education increases by 4.5% (LATE) and 12.2% (ATT). These results are impressive because the same effect that causes an increase in the size of the cohorts does not affect the quality of the education since the Saber 11 score does not decrease. It also boosts the probability of pursuing higher education. Finally, the results show that if the mine closest to the schools is in operation, the probability of having a formal job drops between 0.2% and 8.6%.

The rest of the document is organized as follows: Section 4.2 presents a literature review, Section 4.3 describes the data, Section 4.4 outlines the modeling strategy, Section 4.5 reports the results, and Section 4.6 offers the conclusions.

4.2 Literature Review

In this section, I describe the literature related to the chapter. First, I discuss how the broad literature analyzes and links different macroeconomic topics, natural resources, and educational outcomes. Next, I present some specific papers related to the Colombian case and discuss the current context.

Since mining is one of the largest contributors to State revenues, and public policy tends to be in line with the incentives for foreign direct investment (Gylfason, 2001) , mining regulation should be a priority in the government's agenda. With effective regulation, governments can also control the collateral side effects of mining activity. However, mining activities' effects in LACs are dual: although there is evidence of an improvement in macroeconomic indicators (Loayza et al., 2014; Maldonado, 2018), the mining industry has increased social inequality.

On the one hand, mining activity encourages per capita consumption and employment, with a significant spillover in education. The mine's labor force will require better-educated immigrants, such as operators, technicians, engineers, and ecologists, who will move to the area and will require improvements in the region's education and health system quality for their families (Loayza and Rigolini, 2016).

On the other hand, the returns generated by natural resources are approximately 60% of local governments' revenue in developing countries. However, the perception is that their distribution is inefficient in providing public services such as health and education. Regarding education, a deterioration of these services can help explain the decrease in cognitive abilities and increase in school absenteeism (Gilliland et al., 2001; Lavy and Roth, 2014); regarding health, the community may be affected by exposure to pollution generated by extractive activity (Martínez, 2023; Martínez et al., 2017; Romero and Saavedra, 2016).

The related literature suggests some strategies that explain and interpret in different contexts the consequences of the exploitation of natural resources for social dynamics and economic growth (Angrist and Kugler, 2008; Loayza et al., 2014; van der Ploeg, 2011). For instance, different results or behaviors have been identified in terms of economic policy (Caselli, 2006), including modest reductions in poverty coupled with a positive impact on literacy rates (Litschig and Morrison, 2013; Pegg, 2006). Moreover, the government's efficiency in the provision of public goods has been scrutinized (Maldonado, 2018; Maldonado and Ardanaz, 2023). Along these lines, it is worth highlighting Martínez et al. (2017), who compare the effects of public spending from two types of returns–oil revenues (transfers) and the fiscal effort of local governments and finds equivalences in the spending on education as a result of both sources of income.

In this context, Agüero et al. (2016) found that the distribution of transfers from mining activity in a "boom" context directly impacts the conditions of public services provided to the population, especially in the quality of education, as observed via mathematics scores, as Hanushek and Woessmann (2010) predict. Additionally, Álvarez and Vergara (2022) complemented these findings by introducing the change in wages in the producing regions as an important transmission mechanism for human capital formation, measured in years of schooling, and the high demand for work during the "boom" of natural resource exports. According to economic growth theory, there are two ways in which human capital impacts growth: on the one hand, through the number of years of education (years of schooling) and on the other, via the effect of the quality of education on economic growth - productivity - (Barro, 2001). On this subject, Angrist and Kugler (2008) find that an increase in the demand for coca, opium, and diamond production increases school attendance and child labor in the same rural areas where such goods are concentrated during "boom" times in Colombia. This result is linked to civil conflicts and the violence associated with growing illegal crops, mainly coca (Currie et al., 2014; Loayza et al., 2014; Ross, 2015). This link to civil conflicts is also one reason that education results are not good, particularly in rural areas (Dube and Vargas, 2013).

However, the impact of mines' locations on human capital formation and entry into the labor market remains ambiguous, with investment changes only partially accounting for the observed outcomes. Geographical location and exposure to mining activities play a crucial role, resulting in both gains and losses for the affected population (Torvik, 2002). Notably, the quality of and access to health systems in these areas undergo significant changes, leading to increased pollution and toxic emissions (Currie et al., 2014; Romero and Saavedra, 2016). These environmental and health concerns contribute to heightened school absences, attributed to compromised health and impaired cognitive skills during the schooling phase (Almond et al., 2009; Currie et al., 2009; Park et al., 2002).

Finally, one key paper is Bonilla Mejía (2020), which explores how gold mining (legal and illegal) impacts human capital accumulation through two main measures, an invariable interaction over time between the intensity of gold deposits near schools and the recording of international gold prices, and two alternatives for assessing the differential effects of illegal mining (active mining titles and mining deforestation). Its objective is to elucidate the labor market, especially violence and corruption, as mechanisms through which mining may affect human capital accumulation (enrollment, dropout rate, and results of standardized tests or household surveys). Using differences in differences and instrumental variables models, Bonilla Mejía (2020) finds that mining increases enrollment and progression (promotion rates between levels and reduces dropouts) at lower school levels. These effects tend to fade at the upper secondary level. Illegal mining shows larger but consistent effects on extensive outcomes (multiplied up to 3 times when including instrumental variables) for gold mines located between 10 and 50 km from the school. The effects are higher for the elementary school level located closer than 30 km and for secondary schools located between 30 and 50 km from the mines. In general, gold mining decreases student performance at school, especially in the early stages; the impacts are three times higher when including (illegal) mining deforestation.

This research complements Bonilla Mejía (2020) on exploiting labor market mechanisms and the distances between mines and schools. I incorporate the number of mines surrounding the school to test the strength of the effects. I differed in the effects studied on the size of the cohort, the Saber 11 test score, and the probabilities of enrolling in higher education and the labor market. I extended the controls applied to the models, including family, socioeconomic, and institutional characteristics, and other minerals extracted in the mines. I also include all high schools, their schedules, sector, and type. I have a relevant value-added, the measurement of exposure. This measurement is crucial, as in the literature, it is not very easy to separate the effect of the mine (in this case) and the impacts of the municipality's royalties. I include the effects of temporal variation in the definition of treatment and exposure. Furthermore, by doing so, I also eliminate biases from external issues such as economic benefits from operating the mine in a specific municipality, since I observe the effect of the proximity of the mine on nearby schools regardless of the municipality.

4.3 Data

This section describes the data and outline our assumptions. The first part describes the databases for individuals, schools, and mines. The second part describes how these databases were combined.

In Colombia, the Ministry of Education (MEN) administers the Saber 11 exam, a prerequisite for higher education enrollment, to all secondary education students through the Colombian Institute for the Promotion of higher education (ICFES). The ICFES database (also known as the Saber 11 database) contains comprehensive information, including student demographics, exam scores, and various economic, individual, family, and academic variables. While utilizing the Saber 11 database from 2002 to 2016, certain modifications were made. Notably, the absence of household income data for certain periods necessitated imputation based on the mode of household income within the same school for other periods, prioritizing the higher value when multiple options existed. Moreover, the standardized test score scale changed over time, preventing direct comparisons across different years. To address this issue, each student's percentile on the Saber 11 exam was calculated and used as the new standardized score variable. This approach aligns with the Ministry of Education's methodology for standardizing the Saber 11 score in its own information systems. Additionally, key variables, such as the year of birth, gender, standardized Saber 11 test score, household income, school ID code for integration with the school census, and an ethnic group indicator, were extracted from the ICFES database for analysis. The size of the cohort was estimated by counting the students who presented the test linked to a school code in a period of time (see descriptive stats in Table 4.1).

The information gathered on secondary schools originates from the Ministry of Education. The dataset encompasses essential details, such as the school's address, geolocation, urban or rural classification, and descriptive information concerning the sector (public or private) and type (academic, technical, or military). Notably, due to varying schedules, multiple schools can share the same campus in Colombia's secondary education system. Consequently, certain schools may share the same geographical location while differing in their respective sectors. For instance, a public-sector building might serve as a secondary school in the morning and be leased to a private school for afternoon or evening sessions. In this scenario, the schools would share the location while differing in their operational sectors. Schools not operated by the government are categorized as private, including those managed by private entities under contract with the government. In terms of the urban/rural classification, the definition of "rural" encompasses any location not explicitly designated as urban. Thus, schools categorized as having mixed urban–rural or rural–urban statuses are coded as rural. From the secondary school census, the following attributes are utilized: school location (as specified above), school sector (as described earlier), school type (e.g., single-gender or coeducational), school shift (i.e., morning, afternoon, or evening classes), and school degree type (academic or technical) (see descriptive statistics in Table 4.1, a map showing the locations of the school is available in Figure 4.1).

In addition to the ICFES database and the school census, the Colombian Ministry of Education (MEN) manages the System for the Prevention and Analysis of Dropout in higher education Institutions (SPADIES) database. This comprehensive database comprises the academic information of all students enrolled in higher education institutions (HEIs) since 1998. The SPADIES database includes details such as the HEIs in which students are enrolled, first and last periods of enrollment in higher education, status in the system (active, dropout, graduated), program of study and area of concentration, type of degree (bachelor's or associate programs), and method of learning (classroom or online). This chapter uses a merged dataset from the ICFES and SPADIES databases from 2002 to 2017, employing a dummy variable that takes the value of 1 if the student enrolled in higher education (see descriptive statistics in Table 4.1).

The Planilla Integrada de Liquidación de Aportes (PILA) database, overseen by the Ministry of Labor, contains Social Security payment records for all individuals in the formal labor sector. The database includes details such as the number of days worked annually, employment type (e.g., full-time, part-time, self-employed), and employer type (e.g., public company, private company, nonprofit, nongovernmental organization). These data are extracted from the monthly report of contributions made by all formal Colombian workers to pension and health funds (see descriptive statistics in Table 4.1).

In this chapter, I solely utilize this dataset to evaluate the likelihood of an individual entering the formal labor sector subsequent to secondary school. Specifically, I investigate whether proximity to a mine influences the decision to pursue higher education versus entering the formal labor market directly after completing secondary school. A dummy variable is created from the merge between the ICFES database and the PILA database, taking the value of 1 if the student enrolled in the labor market.

For data on all legal mines in Colombia, I use the Colombian Mining Census, a database collected by *Tierra Minada*, a nonprofit organization that holds the information for the permits and requests to operate mines in Colombia[1]. The Colombian Mining Census contains information such as mine size, natural resource extracted, geographic location, and dates for

[1] Data are available at `https://sites.google.com/site/tierraminada/`

the start and close of operations. All data for mines that have been in operation at any point between 2002 and 2014 are used.

This database contains 9,545 active mines (valid mining titles) for the period of 2002 to 2014. The database includes an address for each mine, but some of the addresses referred to the offices that managed the mine or to the mining complex's entrance, both of which may have been far from the actual mine. To ensure an accurate location for each mine, I programmed algorithms to analyze Google Earth pixel data to detect each mine's most precise location.[2] The final product was a database with the longitude and latitude for each mine reported in the census. The descriptive information for this database can be found in Table A4.1 and a map with their location is available in Figure 4.2.

Finally, to use in the IV approach, we used the data from the Base Metals Price Index (PMETA) from the International Monetary Fund (2023). It includes the prices of aluminum, cobalt, iron ore, lead, molybdenum, nickel, tin, uranium, and zinc. The IMF estimates the PMETA at least twice per year as part of spring and fall assumptions. The price index's base year is 2016=100 (see descriptive statistics in Table 4.1).

4.3.1 The Administrative Data Matching Process and Final Database

Various merging approaches for the administrative databases were employed, contingent on distinct identification variables. First, we merged the ICFES (Saber 11) and SPADIES databases using the same merging technique that the MEN uses[3]. The merge of Saber 11 and PILA was executed by the Ministry of Health and Social Protection employing the national

[2] Activisual, a software development company, provided support programming the code, cross-checking on-field some results from the algorithm, and contacting some mines with incomplete contact information.

[3] The algorithm takes two key variables, namely, the full name and the date of birth, from the databases. First, the algorithm removes the spaces, converts all alphabetic characters to uppercase, and then decomposes the strings into all possible combinations of the characters. For instance, the name "Tom" is transformed into TOM, MOT, OTM, OMT, TMO, MTO. Next, the algorithm compares each discomposed key variable for every observation in each database to all possible observation matches between the databases. If the comparison reaches a certain "trigger" level, the algorithm identifies the observation as a match. The level of match is the percentage of similarity between the discomposed variables. The algorithm is cautious, meaning that if there is more than one potential matching option, it will not execute the matching. In this chapter, the trigger value used is 98%, the same as the value used by the Ministry of Education in the SPADIES-ICFES match.

identification number of Colombia. The merge between Saber 11 and the Census of Schools was achieved using the ICFES school code. MEN provided the school's location. Activisual obtained the location of the mine, and it is computed the orthodromic distance between secondary schools and mines in kilometers. Then, the information about the mine closest to each student is assigned to that student's record.

The final database, created at the individual level, originates from the ICFES database, encompassing information on 7,517,983 students who took the Saber 11 exam between 2002 and 2016. To match the Colombian Mining Census data, entries after 2013 were excluded, resulting in 6,172,756 students in the database. Among them, 400,819 students without a linked secondary school during the Saber 11 test, 60,460 with missing values in the Saber 11 score, and 491 students associated with mines lacking a defined extractable material were dropped. Variables from the ICFES database, including cohort sizes and Saber 11 exam scores, were retained. The SPADIES database was used to obtain higher education enrollment status. Social Security records were used to capture students' labor status in the formal sector after graduating from secondary school.

Additionally, information regarding the nearest mine to each student's record was incorporated, associating students with the mine closest to their respective secondary schools. To ensure result transparency and stability, the approach used the nearest actively operating mine, resolving ambiguities arising from active and nonactive mines within specified radius.

The final database comprised 5,710,986 students with data for the variables of interest (size of cohort, score in Saber 11, enrollment in higher education, and enrollment in the labor market). Data also include individual information such as year of birth, gender, family household income, parents' education levels, ethnic minority status, school type, school term duration, school coordinates, and mine size. Nonmerge students or those with missing data for the controls or variables of interest are distributed homogeneously across time. Finally, PMETA and other commodities prices were merged by year. The final database is a repeated cross-section at the individual level.

4.4 Modeling Strategy

In this section, the model specification is presented. In the first part, I present an ordinary least squares approach (OLS) to provide a guide about the sign of the factors that can affect any of the five outcomes: (1) size of secondary graduation cohort (2) score on the Saber 11 exam, (3) probability of enrollment in higher education, and (4) probability of entry into the formal labor market. In the second part, an instrumental variables (IV) approach is developed to obtain causal estimators of the mine's operation on each of the four outcomes mentioned above. Finally, in the third part, we follow the Callaway and Sant'Anna (2021) methodology to aggregate the results of the difference and difference (DiD) approach.

4.4.1 OLS Approach

The model is based on models used by Balza et al. (2021) and Bonilla Mejía (2020). I estimate the impact of mining activity on educational and labor market outcomes using information from the mines that are located near secondary schools. I control for the size of the mine and the extracted product, extending the analysis of Bonilla Mejía (2020) beyond gold mines and complementing the analysis of Balza et al. (2021) in reaching the full extractive sector.

Using school location and detailed information from the Colombian Mining Census, I incorporated not only the number of mines around each school, but also the moment when these mines started to operate. In short, the model compares a student's outcomes before and after the operation time of the closest mine to the school. To do so, I created my variable of interest $OM = Closest mine is operating$ as a dummy variable that takes the value of 1 when the closest mine to each school starts its operation $OM = 1$ and the value of zero $OM = 0$ when the nearest mine is not operating. Mines can be out of business or not yet working. The treatment is assigned to each student through his or her secondary school and year of graduation; this means that the treatment group consists of those students who took the Saber 11 exam during the mine's period of active operation. The possibility of identifying how one mine simultaneously impacts different schools located in different municipalities helps me avoid selection problems due to the municipality in the results. However, a control for departmental fixed effects is included. Equation 4.1 shows the regression approach model:

$$Y_{i,s,t} = \alpha_i + \delta_t + \rho OM_{i,s,t} + X'_{i,s,t}\beta + \varepsilon_{i,s,t} \qquad (4.1)$$

In this equation, $Y_{i,s,t}$ represents one of the four outputs that are analyzed: (1) size of secondary graduation cohort (2) score on the Saber 11 exam, (3) probability of enrollment in higher education, and (4) probability of entry into the formal labor market. Some tables will show a small variation of output (4) extending it only to those who enrolled in works related to the formal labor market in the mining sector. The variable of interest is "OM" a dummy variable that equals 1 if individual "i" graduated from a secondary school “s” whose nearest mine was operating in the year "t" and 0 otherwise. The parameter ρ denotes the rate of change in output $Y_{i,s,t}$ due to the closest mine's operation. The control α_i incorporates all the time-invariant characteristics of each individual, including gender, Saber 11 score (used as a proxy for academic ability), household income, parents' education, ethnicity, and year of birth. The parameter δ_t captures time-varying drivers. The vector X includes observable predictors for the outputs that are linked to the student by his/her school Distance to the closest mine, Distance square, Time of operation of the closest mine, Number of mines in certain ratios, Size of the closest mine, Sector of the school, if school is coed, if school is in urban area, if school conducts to academic (regular degree), the calendar of the school (starting the academic year in January), the location of the school.

It is important to acknowledge that this model, while aligned with previous research and pertinent to our inquiries, faces limitations in establishing causality. Notably, the nonrandom assignment of education levels poses a significant challenge in studying the connection between education and earnings. Individuals make deliberate choices concerning their educational paths, considering opportunity costs (as emphasized by Wood (2009)). To mitigate potential econometric challenges such as sample selection and endogeneity, we employ an instrumental variables approach to estimate the Local Average Treatment Effect (LATE) of the mine in operation. Additionally, it is crucial to recognize that effects may vary across different cohorts, as mines commence operations at various locations and times. To address this concern, we utilize Callaway and Sant'Anna (2021)'s framework to estimate the Average Treatment Effect on the Treated (ATT). This strategy enables us to investigate variations in the effects of education on earnings among diverse subgroups.

4.4.2 Instrumental Variables Approach

The use of instrumental variables is an appropriate methodology when addressing potential endogeneity concerns. In this study, we employ a two-stage least squares (2SLS) estimator to systematically address these issues. To do so, we utilize the price index for metals from the IMF (PETA) interacted with the number of mines within a 1 km radius of schools in the year preceding the start of their operation plus one, similar to the approach used by Balza et al. (2021); Black et al. (2005); Bonilla Mejía (2020); Dube and Vargas (2013) and Michaels (2011). This interaction serves as an instrumental variable, representing a proxy for the supply of mines per school, to quantify the probability of the mines commencing operations.

Therefore, our approach involves estimating a first step to predict the probability of start operation, employing our instrument as an independent variable. The vectors of control variables α_i ,δ_t and $X_{i,s,t}$ remain consistent with Equation 4.1. The first step equation is formally specified as follows:

$$OM_{i,s,t} = \alpha_i + \delta_t + \mu PETA_t \times Mines_{i,s,t-1} + X'_{i,s,t}\beta + \varepsilon_{i,s,t} \qquad (4.2)$$

The instrument used in this study is exogenous, as it is derived from a set of prices in the international market, where Colombia has no control over these prices and acts as a price taker. Moreover, mines are not established with the direct consideration of schools.

Therefore, the instrument can incentivize mine operation (relevance assumption), as an increase in prices would make mining more attractive. Simultaneously, a congested region near schools can discourage mine establishment in certain locations. However, the instrument itself cannot directly impact any of the outputs (exchangeability assumption), as it does not share common causes with the outcomes. International prices and rents do not directly affect students, households, or schools (exclusion restriction).

In Equation 4.3, we utilize the estimated probability of a mine commencing operations, obtained from Equation 4.2, as an instrument for "OM". Given that our instrument satisfies the relevance assumption, exchangeability assumption, and exclusion restriction, as explained earlier, the exogenous variation provided by the instrument in the IV approach yields a precise LATE. Hence, the results in Equation 4.3 can be interpreted as the causal effect of an operational mine on the analyzed output.

$$Y_{i,s,t} = \alpha_i + \delta_t + \rho\widehat{OM_{i,s,t}} + X^{'}_{i,s,t}\beta + \varepsilon_{i,s,t} \qquad (4.3)$$

4.4.3 Heterogeneous Difference in Differences (DiD) Approach

The conventional DiD approach typically employs a 2X2 model involving two time periods and two groups. In the initial period (t=0), both groups exhibited similar characteristics and lacked exposure to the treatment. In the subsequent period (t=1), some individuals receive the treatment, forming a "treated" group (OM=D=1), while others remain "controls" (OM=D=0) without the treatment. This fundamental model aligns with the interpretation presented in Equation 1, where t=0 corresponds to the year preceding the commencement of the nearest mine's operations, and t=1 represents the subsequent year when the mine begins to operate. Equation 4.4 outlines the foundational framework for a DiD analysis based on Equation 4.1.

$$Y_{i,s,t} = \alpha_i + \delta_t + \rho OM_{i,s} \times t + X^{'}_{i,s,t}\beta + \varepsilon_{i,s,t} \qquad (4.4)$$

In this framework, each individual has two potential outcomes: one with treatment and one without treatment. However, our observations are restricted to the outcomes corresponding to each group (treated or not treated) in t=1. In theory, these outcomes should diverge due to the presence or absence of the treatment and can be expressed as:

$$Y_{ist}(OM) = OM_{is} \times Y_{ist}(1) + (1 - OM_{is}) \times Y_{ist}(0) \qquad (4.5)$$

Under the assumption that the treated group would follow a predetermined trajectory in the absence of treatment, any deviation from this path can be attributed to the causal impact of the treatment on this group. This deviation, denoted as the ATT, is described in Equation 4.6.

$$ATT = \underbrace{E(Y_{is,1}(1)|OM_{is} = 1)}_{A=Observed\,outcome\,for\,treated} - \underbrace{E(Y_{is,1}(0)|OM_{is} = 1)}_{B=Unobserved\,outcome\,for\,treated} \quad (4.6)$$

In Equation 4.6, we have information about the value of part A, as it represents the observed outcome for the treated group in t=1 after the treatment. However, in regard to part B (as defined in Equation 4.6), the path that the treated group would have followed in the absence of treatment is unknown. To make this estimation, we rely on the assumption that this path would be parallel to the trajectory followed by the control group. This assumption is referred to as the Parallel Trend Assumption (PTA). In simpler terms, we assume that the unobserved path taken by the treated group (B) in the scenario where they did not receive treatment is the same as the observed path in the control group (Equation 4.7).

$$E(Y_{is,1}(0) - Y_{is,0}|OM_{is} = 1) = E(Y_{is,1} - Y_{is,0}|OM_{is} = 0) \quad (4.7)$$

Finally, using Equation 4.7 in Equation 4.6, we can construct a feasible estimator for the ATT that will be given by:

$$\widehat{ATT} = [E(Y_{is,1}(0) - Y_{is,0}|OM_{is} = 1)] - [E(Y_{is,1} - Y_{is,0}|OM_{is} = 0)]$$
$$= E\left(Y_{is,1}|OM_{is} = 1\right) - \hat{E}\left(Y_{is,1}(0)|OM_{is} = 1\right) \quad (4.8)$$

While the PTA can be challenging to satisfy in practice due to potential dissimilarities between the treated and control groups, Callaway and Sant'Anna (2021) propose a generalized approach incorporating additional groups and fixed effects in the specification. DiD designs often involve more than two periods or treated groups, further complicating the PTA assumption. To mitigate this, Sant'Anna and Zhao (2020) suggest using the PTA for groups with identical pretreatment characteristics (α and X), minimizing bias due to group differences. Let "W" represent the set of individuals (α) and mine-linked characteristics (X). Here, $\theta\left(W_{is}\right)$ is the ΔY_{is} if there was no treatment based on pretreatment characteristics. With this updated assumption, the new DiD estimator is $\widehat{ATT_*}$ (Equation 4.9).

$$E(Y_{is,1}(0) - Y_{is,0}|OM_{is} = 1, W_{is}) = E(Y_{is,1} - Y_{is,0}|OM_{is} = 0, W_{is}) = \theta(W)$$
$$\widehat{ATT_*} = E(Y_{is,1}|OM_{is} = 1) - \Big[E(Y_{is,0}|OM_{is} = 1) + \widehat{E}(\theta(W_{is})|OM_{is} = 1)\Big] \qquad (4.9)$$

Various approaches have been proposed in the literature to estimate the component $\widehat{E}(\theta(W_{is})|OM_{is} = 1)$ from Equation 4.9. In this chapter, as it is a cross-sectional database with many different individuals, we adopt the Outcome Regression approach (OR) estimator proposed by Sant'Anna and Zhao (2020). The OR estimator employs a two-step procedure. In the first step using data from the control group, we model $E(\theta_{is}|W_{is}) = \theta(W_{is})$ as a function of W, so $E(\theta_{is}|W_{is} = w_{is}) = \theta(w_{is}) \; \forall \; i|OM_{is} = 0$. Then, $E(\theta_{is}|OM_{is} = 1)$ is estimated by substituting θ_{is} with the predicted outcome $\hat{\theta}\left(w_{is}\right)$. Thus the OR estimator for the ATT becomes:

$$\widehat{ATT_{OR}} = E(\Delta Y_{is}|OM_{is} = 1) - E(\widehat{\theta}(w_{is})|OM_{is} = 1) \qquad (4.10)$$

To handle the varying timing and groups affected by mine operations, we adopt the framework proposed by Callaway and Sant'Anna (2021), building upon the work of Sant'Anna and Zhao (2020). This approach introduces a designated group "g" to represent the cohort of mine operations, accommodating temporal variations marked by "t." Equation 4.11 illustrates how this framework allows us to track the evolution of the proposed ATT over time within a specific group. In our estimation, we implement the methodology developed by Rios-Avila et al. (2021), which is based on Callaway and Sant'Anna (2021). This approach dissects the combinations of groups and times into multiple 2X2 models and then aggregates them based on "g." By following this approach, we can identify ATTs for each treated group "G" at every time point "t" as $ATT(g,t)$.

$$ATT(g,t) = E\left(Y_{is,t} - Y_{is,g-1} | G_{is} = g\right) - \left(E\left(Y_{is,t}(0) - Y_{is,g-1} | G_{is} = g\right)\right) \qquad (4.11)$$

Following this process, we calculate an ATT and corresponding weights for each group within each period. This enables us to consolidate the ATT over time, similar to an event analysis as detailed in the results section, and by group to analyze the impacts within each group and make comparisons. As previously mentioned, the groups may vary in their timing. Thus, in this framework, the population, initially divided into two groups (treatment and control), is now sorted into three sets: treated, not yet treated, and control. The final step is to estimate the expected change in outcomes in the absence of treatment. For this purpose, we apply the conditional PTA assumption to the "not yet treated" group. The PTA assumption is defined as:

$$E\left(Y_{is,t}(0) - Y_{is,g-1} | G_{is} = g\right) = E\left(Y_{is,t} - Y_{is,g-1} | G_{is} = 0\right) \qquad (4.12)$$

Therefore, Equation 4.13 describes the final ATT to be estimated using the PTA for not yet treated.

$$ATT(g,t) = E\left(Y_{is,t} - Y_{is,g-1} | G_{is} = g\right) - \left(E\left(Y_{is,t} - Y_{is,g-1} | G_{is} = 0\right)\right) \quad (4.13)$$

This framework allows us to estimate the causal impact of each period in which mines start to operate in the census and examine how this impact changes over time. We implement both the Simple and Event aggregation methods from Rios-Avila et al. (2021). These aggregation methods are defined as:

$$ATT_{Simple} = \frac{\sum\limits_{t \geq g} w_{g,t} ATT\left(g,t\right)}{\sum\limits_{t \geq g} w_{g,t}} \quad (4.14)$$

$$ATT_{Event} = \frac{\sum\limits_{t+e=g} w_{g,t} ATT\left(g,t\right)}{\sum\limits_{t+e=g} w_{g,t}} \quad (4.15)$$

4.5 Results

In this section, we first analyze the outcomes derived from the Panel Ordinary Least Squares (OLS) model (Equation 4.1, estimated using the Correia (2017)´s command "reghdfe"). Then, we proceed to present the findings from the IV approach, specifically focusing on the LATE estimation as outlined in Equation 4.2 (estimated as a probit in Stata) and Equation 4.3 (estimated using Correia (2018)'s "ivreghdfe"). To conclude, we analyze the results of the heterogeneous DiD results (ATT estimation) in accordance with Equation 4.14 (estimated using Rios-Avila et al. (2021)'s "CSDID").

4.5.1 OLS Results

The duration of the nearest mine's operation has a positive and significant impact on secondary school cohort size and higher education enrollment. Conversely, it has a negative and significant effect on Saber 11 test scores. Interestingly, it does not significantly affect the probability of labor market enrollment (Table 4.2).

Moreover, schools located farther from the nearest mine tend to have smaller cohorts, lower Saber 11 test scores, and a lower probability of college enrollment. However, they do exhibit a slightly higher probability of enrolling in the labor market. Longer mine operation positively influences all four output variables of interest. Notably, having more mines within a 1 km radius negatively affects cohort size, Saber 11 test scores, and college enrollment but positively impacts the probability of labor market enrollment (Table A4.2).

Examining the materials extracted, CO-AS extraction has a negative influence on cohort size, Saber 11 test scores, and labor market enrollment. Conversely, gold extraction has a positive impact on cohort size, test scores, and enrollment in higher education (Table 4.3).

4.5.2 Instrumental Variables Results

The initial phase in obtaining the NL2SLS estimator involves using instruments and regular controls from Equation 4.1 to estimate the probability of the nearest mine's operational status in Equation 4.2. Various instruments, including the price index PMETA, as well as well as the prices of aluminum, gold, iron ore, and zinc, were tested, and all yielded robust instruments. In Table A4.3, Columns 1 to 5 present the results for Equation 4.2 using these instruments. PMETA was selected as it comprises a sample of metals that offers greater accuracy given the diversity of mines in the country. In the subsequent step, employing PMETA as an instrument for OM, the Equation 4.3 was estimated. The coefficients obtained represent the LATE and reveal the causal impact of the mine's operation on various outcomes (Table 4.3).

The primary findings indicate that the coefficient for the Saber 11 test score is not statistically significant, while the probability of enrollment in the labor market is negative and significant. In contrast, the size of the cohort (6%) and the probability of enrolling in college (4.5%) are positive and significant, and their effects are stronger than those found in the OLS section.

Specifically, Co-As and metals are associated with a 16.3% and 15.7% reduction in cohort size, respectively, while gold, construction, and other mines positively impact cohort sizes by 14.4%, 5%, and 26.3%, respectively. Regarding Saber 11 test scores, there is an average increase of 3.8 points when the nearest mine extracts gold. In contrast, there are reductions of 3.6 points and 2.13 points when the nearest mine extracts metals or other materials, respectively.

Despite the decline in test scores, the probability of enrolling in college increases for students whose closest mine extracts construction products. Conversely, students closer to a mine that extracts metals experience a decrease in the probability of enrolling in college. There is no significant effect for Co-As, gold, and Other in terms of the probability of college enrollment.

Finally, the results show that the probability of securing formal employment decreases by 2 percentage points if the closest mine to the school is operational. The negative impact is generally small, but there is significant heterogeneity across most mine types. If the closer mine extracts Co-As or other products, the probability of enrolling in the labor market decreases by 1% and 1.6%, respectively. However, if the secondary school is located near a mine that extracts construction products, gold, or metals, the probability of enrolling in the labor market increases by 1.1%, 1%, and 4.6%, respectively.

4.5.3 Heterogeneous Difference in Differences Results

In this subsection, we aim to analyze the ATT for the closest mine in operation, employing the heterogeneous DiD results from Equation 4.14.

To support the results, the framework rely on the PTA, which requires the pretreatment average to be nonsignificant. In cases where the pre-treatment differs from zero, it is neededa significant change in trend (inverting the sign with significant values) after the treatment, along with other tests to support the results. In this case, the results for the Saber 11 test score and the enrollment in the labor market hold the PTA test, allowing to rely on the reported ATT.

Regarding the size of the cohort, the ATT reports an increase of approximately 7.2%, slightly higher than the 6% reported by the IV approach and the 3.5% from the OLS. The Saber 11 test score shows an increase of 3.8 points (impact of 8.2% compared with the control mean), which is positive and significant, differing from the result from LATE and with the opposite sign from the report from the OLS (Table 4.2).

The probability of enrolling in college increases by 12.2% according to the ATT if the closer mine starts to operate. This coefficient is positive and significant, aligning with the positive results from the IV approach and the OLS. The ATT approach reports a reduction of 8.6% in the probability of enrolling in the labor market. This result is negative, similar to the LATE approach result but also stronger than the LATE coefficient. I also examined whether the impact of the closer mine was specific to the mining sector, and I found similar results in the ATT and LATE, reporting a decrease in the probability of enrolling in the labor market in the mining sector of 1 percentage point (Table 4.2).

Finally, as the Saber 11 test score and Enrollment in the Labor market satisfy the PTA assumption, the aggregation using Equation 4.15 enables us to investigate the impact of the closest mine in operation, resembling an event study. Figure 4.3 illustrates that although the increase in the Saber 11 test score averages approximately 3.8 points, it actually peaks at 17.9 points in the ninth year after the mine starts operating. The average for the posttreatment period is 7.28 points. Figure 4.4 presents the time event for the probability of enrollment in the labor market. In this case, there is a decline of 25.7% in the ninth year after the mine commences operation, and the full posttreatment period records a decrease in the probability of labor market enrollment of 13%.

4.6 Conclusions

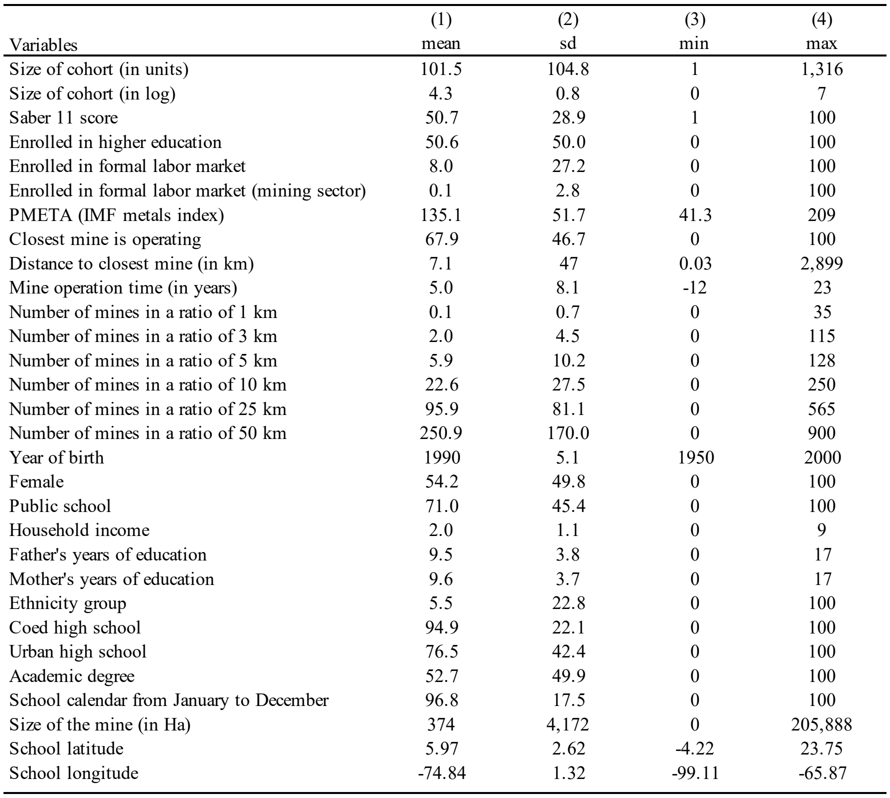
The key findings of this study highlight significant disparities across different types of mined products and their impacts on various outcomes. Consistent with Angrist and Kugler (2008), the observed increase in student cohort size of approximately 6% is notable. The positive influence of the nearest operational mine on cohort size is evident across all three analyzed approaches. However, schools in proximity to gold or Co-As and metals extraction mines show a notable decrease in cohort size. Possible reasons for this decline include the establishment of new schools, student migration due to contamination concerns, or the emergence of informal businesses drawing students away from academics. Wood (2009)'s framework suggests that students with lower academic performance may discontinue their education due to high opportunity costs, especially if the mine encourages informal employment.

While the study revealed no major changes in academic performance, as evidenced by higher Saber 11 test scores, it showed that legal gold mines tend to improve the Saber 11 test score, contrary to the findings of Bonilla Mejía (2020) for both legal and illegal gold mines, whereas other metal extractions have negative impacts. It is important to highlight that the effect of an operational mine nearby is positive, resulting in an increased cohort size annually without compromising academic performance (as measured by the Saber 11 test score) or even enhancing the quality of education (as measured by the increased probability of college enrollment). This effect occurs even as the demand for education outpaces the fixed supply, which is particularly relevant as establishing a new school requires time. Remarkably, the system has effectively managed potential issues related to overcrowding in cohorts, leading to a rise in the participation of secondary graduates in college, and subsequently reducing the likelihood of immediate entry into the formal labor market. These consistent trends suggest that students are actively choosing to prioritize their continued education over immediate employment, underscoring the positive impact of the mine on educational aspirations.

Additionally, the distance from the mine significantly affects cohort size, Saber 11 test scores, and the probability of college enrollment, with a positive effect on labor market participation. Although the size of the mine has a significant but minimal influence, the type of extracted product plays a crucial role. Notably, gold mining has been shown to increase student cohort sizes and Saber 11 scores without affecting the likelihood of college attendance, a result similar to what is found in schools closer to mines extracting other products. These findings contrast with the effects observed in schools closer to mines extracting construction materials, where an increase in cohort size affects the Saber 11 score but not the probability of college enrollment or labor market participation.

Ultimately, this research serves as a vital tool for policymakers grappling with the intricate balance between the economic benefits of mining operations and the imperative of sustainable resource management. It underscores the need for effective regulations and enforcement measures in the extractive industries, safeguarding both the environment and the long-term well-being of Colombian citizens. Proper regulation and enforcement mechanisms are vital to ensure that mining operations remain sustainable, avoid illegal practices, and contribute positively to local communities, preserving the life path of the nation's young students.

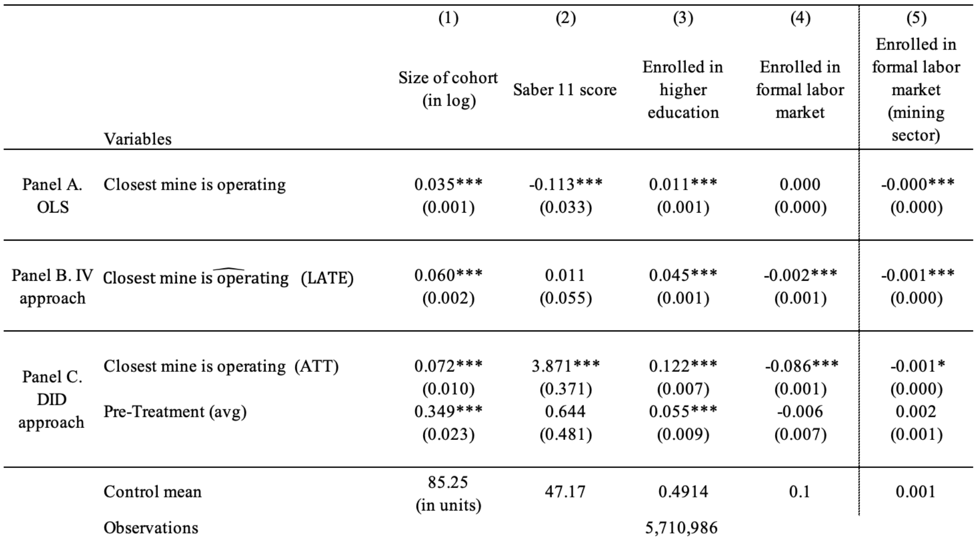
Table 4.1: Main Results -Full Output-



Variables	(1) mean	(2) sd	(3) min	(4) max
Size of cohort (in units)	101.5	104.8	1	1,316
Size of cohort (in log)	4.3	0.8	0	7
Saber 11 score	50.7	28.9	1	100
Enrolled in higher education	50.6	50.0	0	100
Enrolled in formal labor market	8.0	27.2	0	100
Enrolled in formal labor market (mining sector)	0.1	2.8	0	100
PMETA (IMF metals index)	135.1	51.7	41.3	209
Closest mine is operating	67.9	46.7	0	100
Distance to closest mine (in km)	7.1	47	0.03	2,899
Mine operation time (in years)	5.0	8.1	-12	23
Number of mines in a ratio of 1 km	0.1	0.7	0	35
Number of mines in a ratio of 3 km	2.0	4.5	0	115
Number of mines in a ratio of 5 km	5.9	10.2	0	128
Number of mines in a ratio of 10 km	22.6	27.5	0	250
Number of mines in a ratio of 25 km	95.9	81.1	0	565
Number of mines in a ratio of 50 km	250.9	170.0	0	900
Year of birth	1990	5.1	1950	2000
Female	54.2	49.8	0	100
Public school	71.0	45.4	0	100
Household income	2.0	1.1	0	9
Father's years of education	9.5	3.8	0	17
Mother's years of education	9.6	3.7	0	17
Ethnicity group	5.5	22.8	0	100
Coed high school	94.9	22.1	0	100
Urban high school	76.5	42.4	0	100
Academic degree	52.7	49.9	0	100
School calendar from January to December	96.8	17.5	0	100
Size of the mine (in Ha)	374	4,172	0	205,888
School latitude	5.97	2.62	-4.22	23.75
School longitude	-74.84	1.32	-99.11	-65.87

Note: Table shows the mean, standard deviation, minimum, and maximum for the main characteristics of all secondary school graduates who took Saber 11 test from 2002 to 2012. Dummies in percent.

Table 4.2: Main Results



	Variables	(1) Size of cohort (in log)	(2) Saber 11 score	(3) Enrolled in higher education	(4) Enrolled in formal labor market	(5) Enrolled in formal labor market (mining sector)
Panel A. OLS	Closest mine is operating	0.035*** (0.001)	-0.113*** (0.033)	0.011*** (0.001)	0.000 (0.000)	-0.000*** (0.000)
Panel B. IV approach	Closest mine is $\widehat{\text{operating}}$ (LATE)	0.060*** (0.002)	0.011 (0.055)	0.045*** (0.001)	-0.002*** (0.001)	-0.001*** (0.000)
Panel C. DID approach	Closest mine is operating (ATT)	0.072*** (0.010)	3.871*** (0.371)	0.122*** (0.007)	-0.086*** (0.001)	-0.001* (0.000)
	Pre-Treatment (avg)	0.349*** (0.023)	0.644 (0.481)	0.055*** (0.009)	-0.006 (0.007)	0.002 (0.001)
	Control mean	85.25 (in units)	47.17	0.4914	0.1	0.001
	Observations			5,710,986		

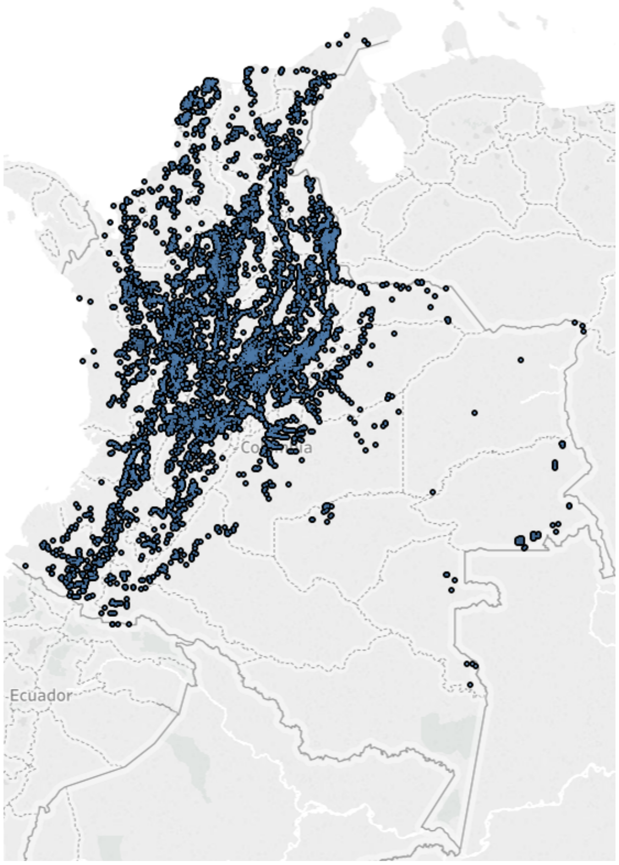
Note: Table shows the coefficients of interest for size of cohort, Saber 11 test score, probability of enrollment in higher education, probability of enrollment in the labor market and in the mining sector of the labor market. Panel A estimated following specification in Equation 4.1, Panel B following specification in Equation 4.3 (first step of IV approach can be found in the Appendix). Panel C estimated with specification in Equation 4.13 for ATT and in Equation 4.15 for pre period for pretreatment check. Robust standard errors in parentheses. ***p<0.01, **p<0.05, *p<0.1.

Table 4.3: Results with Extracted Product

	Full sample	By extracted product				
		Co-As	Construction	Gold	Metals	Other
Variables	(1)	(2)	(3)	(4)	(5)	(6)
		Panel A. OLS				
Size of cohort (in log)	0.035***	-0.065***	0.040***	0.064***	0.003	0.185***
	(0.001)	(0.005)	(0.001)	(0.004)	(0.014)	(0.004)
Saber 11 score	-0.174***	-0.974***	-0.088**	0.608***	-0.315	-0.060
	(0.033)	(0.205)	(0.037)	(0.140)	(0.456)	(0.151)
Enrolled in higher education	0.011***	0.004	0.011***	0.006**	-0.080***	0.007**
	(0.001)	(0.004)	(0.001)	(0.003)	(0.009)	(0.003)
Enrolled in formal labor market	0.000	-0.006***	0.002***	0.002	0.016***	-0.015***
	(0.000)	(0.002)	(0.000)	(0.001)	(0.005)	(0.001)
		Panel B IV approach				
Size of cohort (in log)	0.060***	-0.163***	0.050***	0.144***	-0.157***	0.263***
	(0.002)	(0.007)	(0.002)	(0.007)	(0.023)	(0.007)
Saber 11 score	0.011	0.671**	-0.309***	3.833***	-3.594***	-2.133***
	(0.055)	(0.306)	(0.061)	(0.268)	(0.721)	(0.252)
Enrolled in higher education	0.045***	0.004	0.036***	0.007	-0.115***	0.007
	(0.001)	(0.006)	(0.001)	(0.005)	(0.014)	(0.005)
Enrolled in formal labor market	-0.002***	-0.010***	0.011***	0.010***	0.046***	-0.016***
	(0.001)	(0.003)	(0.001)	(0.003)	(0.008)	(0.002)
Observations	5,710,986	190,625	4,843,896	321,301	42,258	312,906

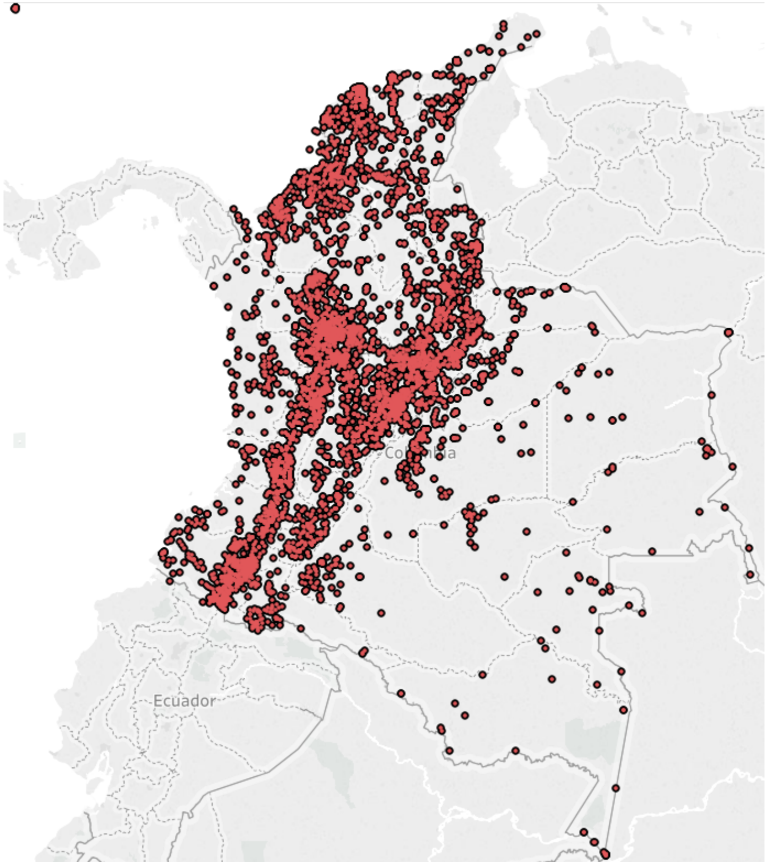
Note: Table shows the coefficients of interest for size of cohort, Saber 11 test score, probability of enrollment in higher education, probability of enrollment in the labor market and in the mining sector of the labor market. Panel A estimated following specification in Equation 4.1 (full results can be found in the Appendix), Panel B following specification in Equation 4.3 (first step of IV approach and full regression results can be found in the Appendix). Co-As is coal and asbestos- Robust standard errors in parentheses. ***p<0.01, **p<0.05, *p<0.1.

Figure 4.1: Location of Legal Mines in Colombia



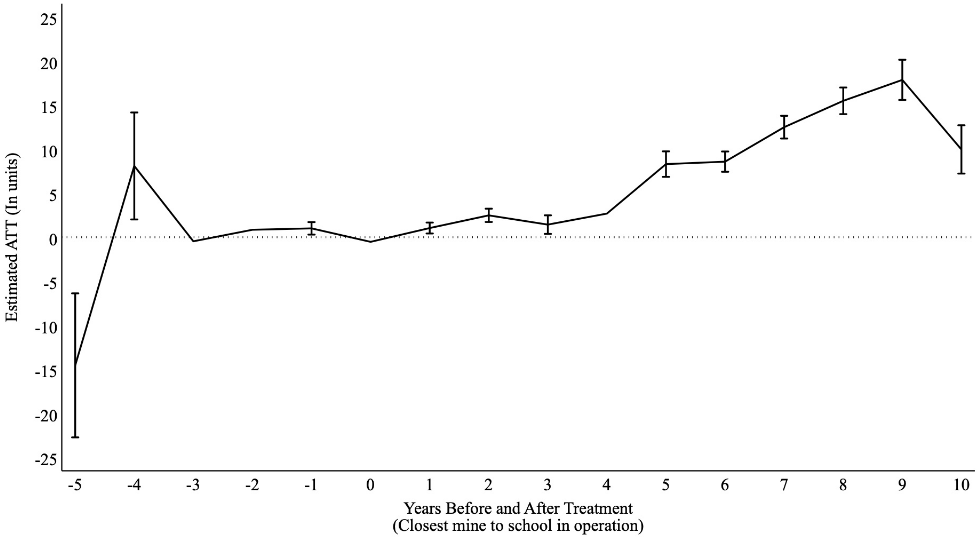
Note: The map shows the location of the mining titles according to the geolocation made by Activisual. The Colombian mining census data for 2014 from Tierra Minera was downloaded in 2017.

Figure 4.2: Location of secondary schools in Colombia



Note: The map shows the schools' location according to the Ministry of Education's geolocation and the cross-check made by Activisual.

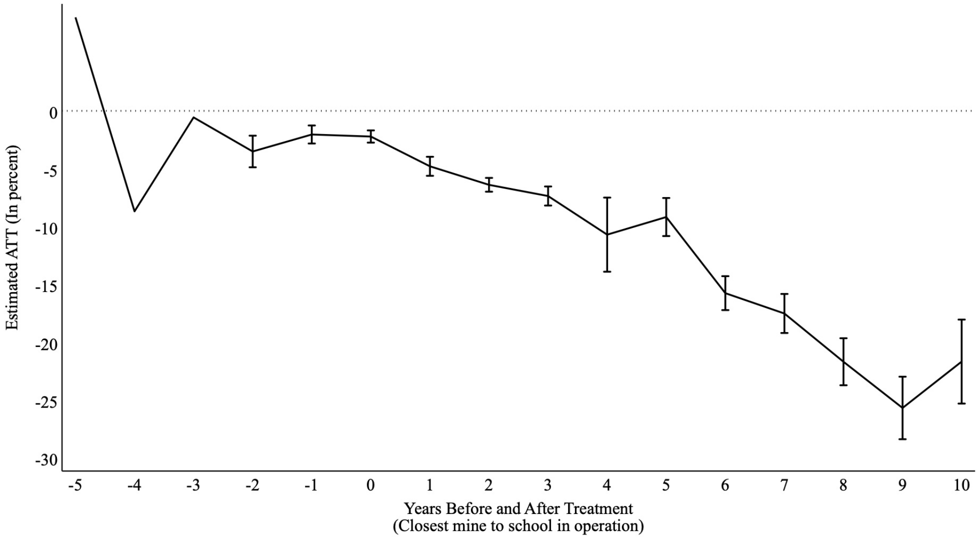
Figure 4.3: ATT – Saber 11 test score



Output Regression Estimator OR - Sant'Anna & Zhao (2020)
Confidence interval was removed if P-value>0.05

Notes: The figure shows the estimated ATT coefficients for the Saber 11 test score. The treated group consists of individuals who presented the Saber 11 test while they were enrolled in a secondary school whose closest mine was not in operation at the moment of the exam (dotted horizontal line in Y=0). The X-axis represents the years before and after the closest mine starts operation, while the whiskers depict the 95 percent confidence intervals. Confidence interval was removed if Pvalue>0.05. The coefficients are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 4.15) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) OR estimator (Equation 4.11). For further details on the variables used as controls, readers can refer to Table 4.1.

Figure 4.4: ATT – Saber 11 test score



Output Regression Estimator OR - Sant'Anna & Zhao (2020)
Confidence interval was removed if P-value>0.05

Notes: The figure shows the estimated ATT coefficients for the probability of enrollment in the labor market. The treated group consists of individuals who presented the Saber 11 test while they were enrolled in a secondary school whose closest mine was not in operation at the moment of the exam (dotted horizontal line in Y=0). The X-axis represents the years before and after the closest mine starts operation, while the whiskers depict the 95 percent confidence intervals. Confidence interval was removed if Pvalue>0.05. The coefficients are obtained utilizing the framework proposed by Callaway and Sant'Anna (2021) (Equation 4.15) through the Rios-Avila et al. (2021) methodology for the Sant'Anna and Zhao (2020) OR estimator (Equation 4.11). For further details on the variables used as controls, readers can refer to Table 4.1.

References

Agüero, Jorge M, Carlos Felipe Balcázar, Stanislao Maldonado, and Hugo Ñopo, "Natural Resources, Redistribution and Human Capital Formation," *Documentos de trabajo Universidad del Rosario*, 2016, (1-46).

Almond, Douglas, Lena Edlund, and Palme Marten, "Chernobyl's Subclinical Legacy: Prenatal Exposure to Radioactive Fallout and School Outcomes in Sweden," *Quarterly Journal of Economics*, 2009, *124* (4), 1729–1772.

Álvarez, Roberto and Damián Vergara, "Natural resources and educational attainment: Evidence from Chile," *Resources Policy*, jun 2022, *76*, 102573.

Angrist, Joshua D. and Adriana Kugler, "Rural Windfall or a New Resource Curse? Coca, Income, and Civil Conflict in Colombia.," *Review of Economics and Statistics*, 2008, pp. 1–53.

Balza, Lenin H, Camilo De los Rios, Raul Jimenez, and Osmel Manzano, "The Local Human Capital Cost of Oil Exploitation," 2021.

Barro, Robert J., "Human capital: Growth, history, and policy - A session to honor Stanley Engerman: Human capital and growth," *American Economic Review*, 2001, *91* (2), 12–17.

Black, Dan A., Terra G. McKinnish, and Seth Sanders, "The Economic Impact Of The Coal Boom And Bust," *The Economic Journal*, apr 2005, *115* (503), 449–476.

Bonilla Mejía, Leonardo, "Mining and human capital accumulation: Evidence from the Colombian gold rush," *Journal of Development Economics*, jun 2020, *145*, 102471.

Callaway, Brantly and Pedro H.C. Sant'Anna, "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

Caselli, Francesco, "Power Struggles and the Natural Resource Curse," *LSE working paper*, 2006, (December 2005), 1–2.

_ **and Guy Michaels**, "Do oil windfalls improve living standards? Evidence from brazil," *American Economic Journal: Applied Economics*, 2013, *5* (1), 208–238.

Correia, Sergio, "reghdfe: Stata module for linear and instrumental-variable/gmm regression absorbing multiple levels of fixed effects," 2017.

_ , "ivreghdfe: Stata module for extended instrumental variable regressions with multiple levels of fixed effects," 2018.

Currie, Janet, Eric Hanushek, E. Megan Kahn, Matthew Neidell, and Steven G. Rivkin, "Does Pollution Increase School Absences?," *Review of Economics and Statistics*, 2009, *91* (4), 682–694.

_ **, Joshua Graff Zivin, Jamie Mullins, and Matthew Neidell**, "What Do We Know About Short- and Long-Term Effects of Early-Life Exposure to Pollution?," *Annual Review of Resource Economics*, 2014, *6* (1), 217–247.

Dube, Oeindrila and Juan F. Vargas, "Commodity price shocks and civil conflict: Evidence from Colombia," *Review of Economic Studies*, 2013, *80* (4), 1384–1421.

Gilliland, Frank, Kiros Berhane, Edward Rappaport, Duncan Thomas, Edward Avol, James Gauderman, Stephanie London, Helene Margolis, Rob McConnell, K. Talat Islam, and John Peters, "The effects of ambient air pollution on school absenteeism due to respiratory illnesses.," *Epidemiology (Cambridge, Mass.)*, 2001, *12* (1), 43–54.

Gylfason, Thorvaldur, "Natural resources, education, and economic development," *European Economic Review*, 2001, *45* (4-6), 847–859.

Hanushek, Eric and Ludger Woessmann, "Education and Economic Growth Early Studies of Schooling Quantity and Economic Growth," *International Encyclopedia of Education*, 2010, *2*, 245–252.

International Monetary Fund, "IMF Primary Commodity Prices," 2023.

Lavy, Victor and Sefi Roth, "The Impact of Short Term Exposure to Ambient Air Pollution on Cognitive Performance and Human Capital Formation," *National Bureau of Economic Research*, 2014, pp. 1–40.

Litschig, Stepan and Kevin M. Morrison, "The impact of intergovernmental transfers on education outcomes and poverty reduction," *American Economic Journal: Applied Economics*, 2013, *5* (4), 206–240.

Loayza, Norman, Alfredo Mier y Teran, and Jamele Rigolini, "Poverty, Inequality, and the Local Natural Resource Curse," 2013, (February).

_ and Jamele Rigolini, "The Local Impact of Mining on Poverty and Inequality : Evidence from the Commodity Boom in Peru The Local Impact of Mining on Poverty and Inequality : Evidence from the Commodity Boom in Peru," *World Development*, 2016, *84* (33), 219–234.

_ , _ , and Oscar Calvo-González, "More than you can handle: Decentralization and spending ability of peruvian municipalities," *Economics and Politics*, 2014, *26* (1), 56–78.

Maldonado, Stanislao, "The Non-Monotonic Political Effects of Resource Booms," 2018.

_ and Martin Ardanaz, "Natural resource windfalls and efficiency in local government expenditure: Evidence from Peru," *Economics and Politics*, mar 2023, *35* (1), 28–64.

Martínez, Luis Roberto, "Natural Resource Rents, Local Taxes, and Government Performance: Evidence from Colombia," *Review of Economics and Statistics*, apr 2023, pp. 1–28.

Martínez Ortiz, Astrid and Tatiana Aguilar, "Impacto socioeconómico de la Minería en Colombia," Technical Report 2012.

Martínez, Zoraya, María González, Jessica Paternina, and Mónica Cantero, "Crop soils pollution by heavy metals, the Alacran mining area, Córdoba-Colombia," *Revista Temas Agrarios*, 2017, *22* (2), 20–32.

Michaels, Guy, "The Long Term Consequences of Resource-Based Specialisation*," *The Economic Journal*, mar 2011, *121* (551), 31–57.

Park, Hyesook, Boeun Lee, Eun-Hee Ha, Jong-Tae Lee, Ho Kim, and Yun-Chul Hong, "Association of Air Pollution With School Absenteeism Due to Illness," *Archives of Pediatrics and Adolescent Medicine*, dec 2002, *156* (12), 1235–1239.

Pegg, Scott, "Mining and poverty reduction: Transforming rhetoric into reality," *Journal of Cleaner Production*, 2006, *14* (3-4), 376–387.

Rios-Avila, Fernando, Brantly Callaway, and Pedro H.C. Sant'Anna, "csdid: Difference-in-Differences with Multiple Time Periods in Stata," 2021.

Romero, Mauricio and Santiago Saavedra, "The Effects of Gold Mining on Newborns' Health," 2016.

Ross, Michael, "What Have We Learned about the Resource Curse?," *Annual Review of Political Science*, 2015, *18* (1), 239–259.

Sant'Anna, Pedro H.C. and Jun Zhao, "Doubly robust difference-in-differences estimators," *Journal of Econometrics*, nov 2020, *219* (1), 101–122.

Torvik, Ragnar, "Natural resources, rent seeking and welfare," *Journal of Development Economics*, 2002, *67* (2), 455–470.

van der Ploeg, Frederick, "Natural Resources: Curse or Blessing?," *Journal of Economic Literature*, 2011, *49* (2), 366–420.

Wood, Tom, "The Sheepskin Effect," Technical Report July 2009.

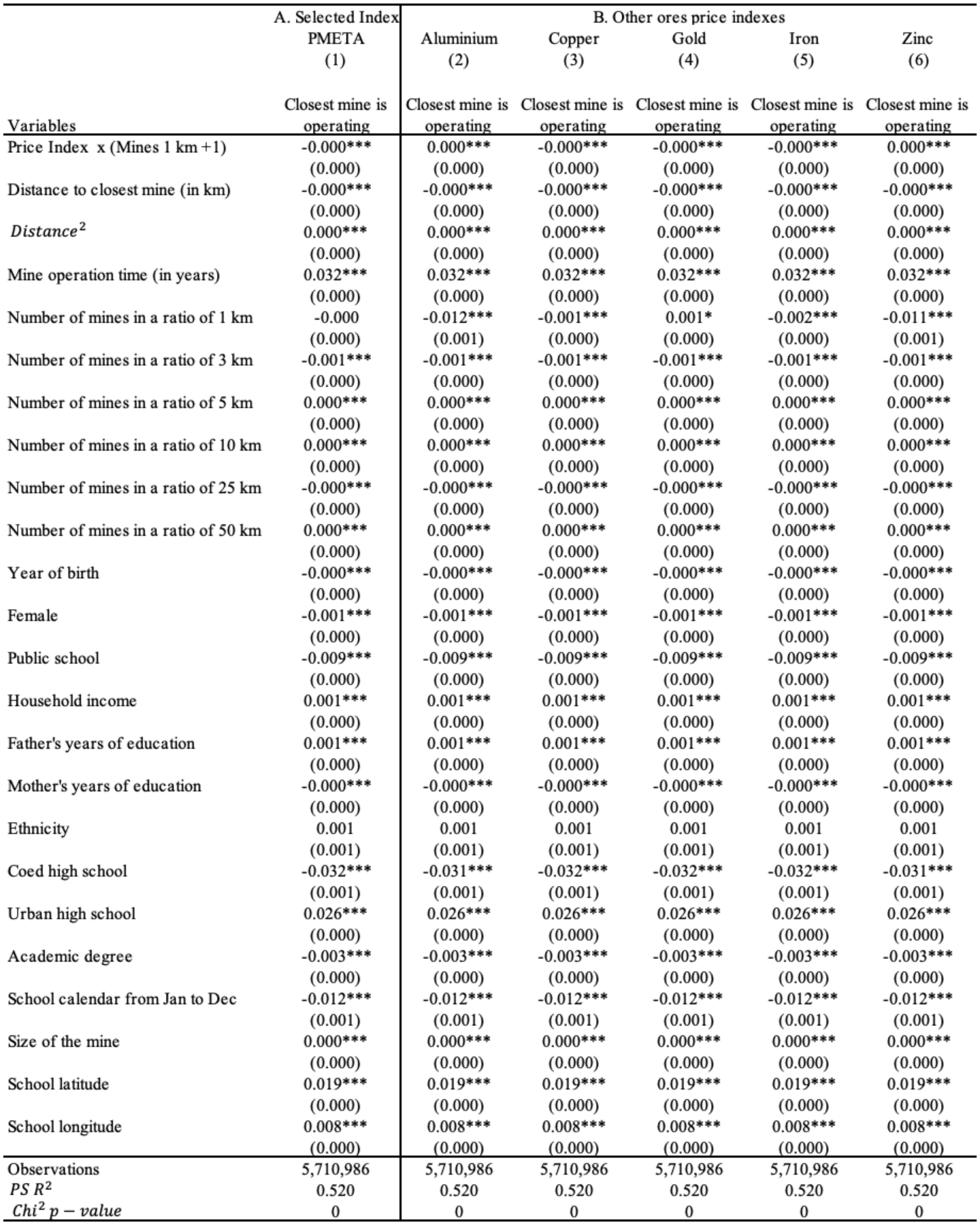
4.7 Appendix

Table A4.1: Status of Mines According with Census in 2014

Department		Status		
	In progress	Reactivated	Discontinued	Total
Antioquia	1,580	8	1	1,589
Arauca	44	0	0	44
Atlantico	98	2	0	100
Bogota	34	0	0	34
Bolivar	444	0	0	444
Boyaca	1,534	3	1	1,538
Caldas	396	1	1	398
Caqueta	60	0	0	60
Casanare	147	2	0	149
Cauca	216	0	0	216
Cesar	382	0	2	384
Choco	162	0	16	178
Cordoba	103	0	1	104
Cundinamarca	993	4	0	997
Guainia	32	2	0	34
La Guajira	53	0	0	53
Guaviare	14	0	0	14
Huila	211	0	0	211
Magdalena	73	1	0	74
Meta	226	0	0	226
Narino	206	0	0	206
N.Santander	709	0	0	709
Putumayo	52	0	0	52
Quindío	68	0	0	68
Risaralda	69	0	0	69
Santander	673	1	0	674
Sucre	67	0	0	67
Tolima	595	0	2	597
Valle Del Cauca	304	0	0	304
Vaupes	9	0	0	9
Vichada	6	0	0	6
Total	9,545	24	24	9,593

Source: Agencia Nacional de Mineria (National Agency for Mining) -ANM-; Tierra Minera.

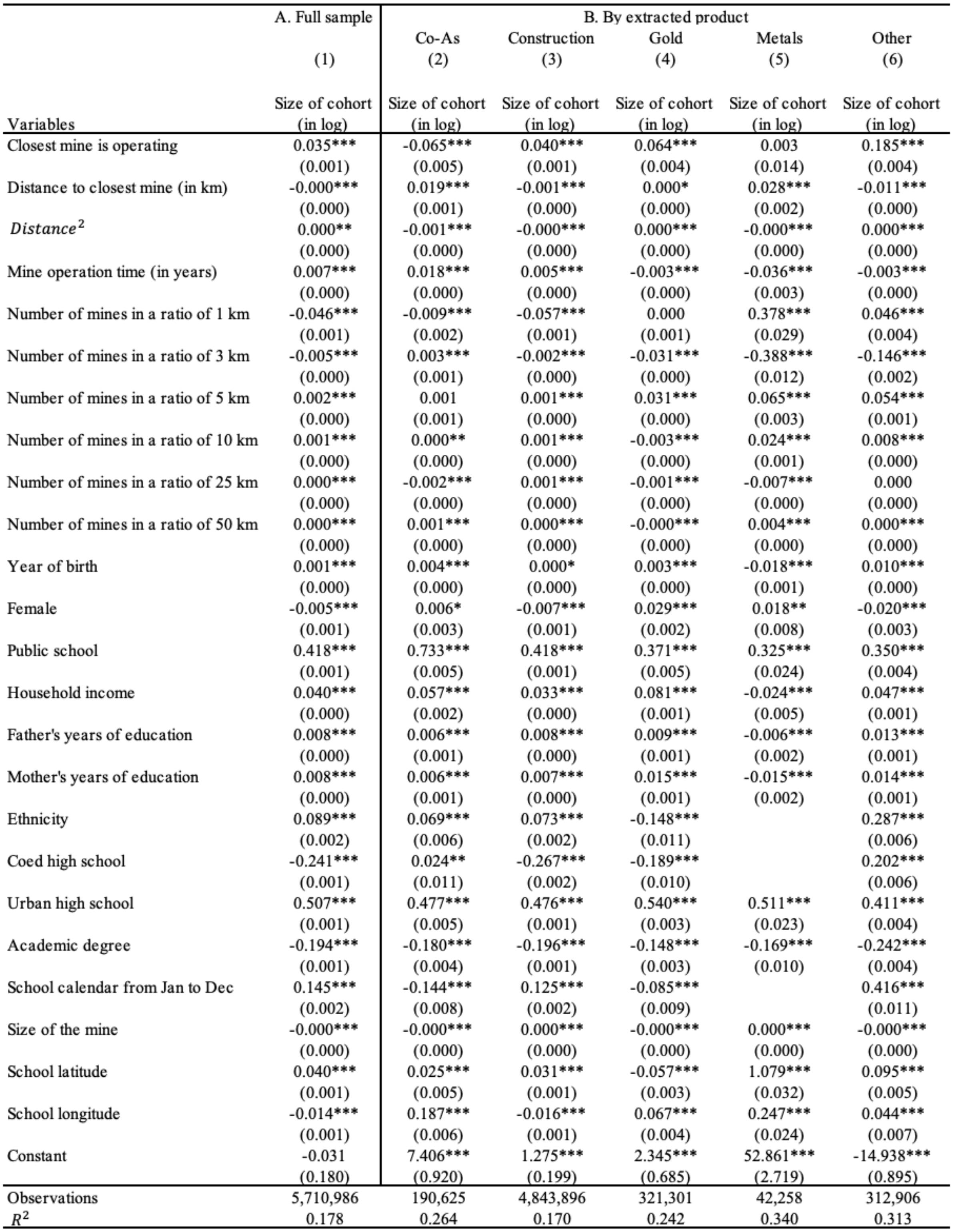
Table A4.2: First Step. Marginal Effects



	A. Selected Index	B. Other ores price indexes				
	PMETA	Aluminium	Copper	Gold	Iron	Zinc
	(1)	(2)	(3)	(4)	(5)	(6)
Variables	Closest mine is operating	Closest mine is operating	Closest mine is operating	Closest mine is operating	Closest mine is operating	Closest mine is operating
Price Index x (Mines 1 km +1)	-0.000***	0.000***	-0.000***	-0.000***	-0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Distance to closest mine (in km)	-0.000***	-0.000***	-0.000***	-0.000***	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$Distance^2$	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mine operation time (in years)	0.032***	0.032***	0.032***	0.032***	0.032***	0.032***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Number of mines in a ratio of 1 km	-0.000	-0.012***	-0.001***	0.001*	-0.002***	-0.011***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.001)
Number of mines in a ratio of 3 km	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Number of mines in a ratio of 5 km	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Number of mines in a ratio of 10 km	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Number of mines in a ratio of 25 km	-0.000***	-0.000***	-0.000***	-0.000***	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Number of mines in a ratio of 50 km	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Year of birth	-0.000***	-0.000***	-0.000***	-0.000***	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Public school	-0.009***	-0.009***	-0.009***	-0.009***	-0.009***	-0.009***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Household income	0.001***	0.001***	0.001***	0.001***	0.001***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Father's years of education	0.001***	0.001***	0.001***	0.001***	0.001***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mother's years of education	-0.000***	-0.000***	-0.000***	-0.000***	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Ethnicity	0.001	0.001	0.001	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Coed high school	-0.032***	-0.031***	-0.032***	-0.032***	-0.032***	-0.031***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Urban high school	0.026***	0.026***	0.026***	0.026***	0.026***	0.026***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Academic degree	-0.003***	-0.003***	-0.003***	-0.003***	-0.003***	-0.003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
School calendar from Jan to Dec	-0.012***	-0.012***	-0.012***	-0.012***	-0.012***	-0.012***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Size of the mine	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
School latitude	0.019***	0.019***	0.019***	0.019***	0.019***	0.019***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
School longitude	0.008***	0.008***	0.008***	0.008***	0.008***	0.008***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	5,710,986	5,710,986	5,710,986	5,710,986	5,710,986	5,710,986
$PS\ R^2$	0.520	0.520	0.520	0.520	0.520	0.520
$Chi^2\ p-value$	0	0	0	0	0	0

Note: The table displays the coefficients of interest for the first step in the IV approach (Equation 4.2). Panel A presents the results for PMETA, while Panel B divides the results by other types of commodities. The estimations follow the specifications outlined in Equation 4.2 with a probit model, with the regression incorporating time and departmental controls (not shown). Marginal effects for the output are shown. Robust standard errors are reported in parentheses. ***p<0.01, **p<0.05, *p<0.1.

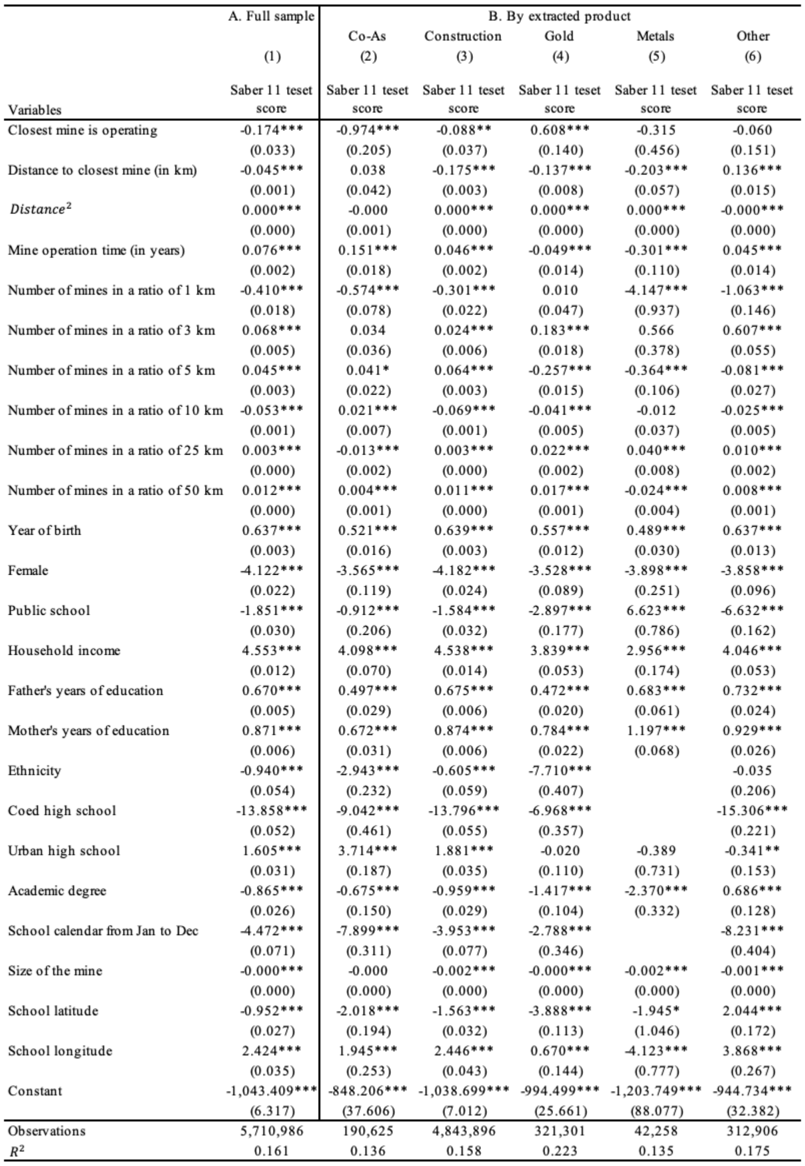
Table A4.3: Size of cohort (OLS)



	A. Full sample	B. By extracted product				
		Co-As	Construction	Gold	Metals	Other
	(1)	(2)	(3)	(4)	(5)	(6)
Variables	Size of cohort (in log)	Size of cohort (in log)	Size of cohort (in log)	Size of cohort (in log)	Size of cohort (in log)	Size of cohort (in log)
Closest mine is operating	0.035***	-0.065***	0.040***	0.064***	0.003	0.185***
	(0.001)	(0.005)	(0.001)	(0.004)	(0.014)	(0.004)
Distance to closest mine (in km)	-0.000***	0.019***	-0.001***	0.000*	0.028***	-0.011***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.002)	(0.000)
$Distance^2$	0.000**	-0.001***	-0.000***	0.000***	-0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mine operation time (in years)	0.007***	0.018***	0.005***	-0.003***	-0.036***	-0.003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.003)	(0.000)
Number of mines in a ratio of 1 km	-0.046***	-0.009***	-0.057***	0.000	0.378***	0.046***
	(0.001)	(0.002)	(0.001)	(0.001)	(0.029)	(0.004)
Number of mines in a ratio of 3 km	-0.005***	0.003***	-0.002***	-0.031***	-0.388***	-0.146***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.012)	(0.002)
Number of mines in a ratio of 5 km	0.002***	0.001	0.001***	0.031***	0.065***	0.054***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.003)	(0.001)
Number of mines in a ratio of 10 km	0.001***	0.000**	0.001***	-0.003***	0.024***	0.008***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Number of mines in a ratio of 25 km	0.000***	-0.002***	0.001***	-0.001***	-0.007***	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Number of mines in a ratio of 50 km	0.000***	0.001***	0.000***	-0.000***	0.004***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Year of birth	0.001***	0.004***	0.000*	0.003***	-0.018***	0.010***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Female	-0.005***	0.006*	-0.007***	0.029***	0.018**	-0.020***
	(0.001)	(0.003)	(0.001)	(0.002)	(0.008)	(0.003)
Public school	0.418***	0.733***	0.418***	0.371***	0.325***	0.350***
	(0.001)	(0.005)	(0.001)	(0.005)	(0.024)	(0.004)
Household income	0.040***	0.057***	0.033***	0.081***	-0.024***	0.047***
	(0.000)	(0.002)	(0.000)	(0.001)	(0.005)	(0.001)
Father's years of education	0.008***	0.006***	0.008***	0.009***	-0.006***	0.013***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	(0.001)
Mother's years of education	0.008***	0.006***	0.007***	0.015***	-0.015***	0.014***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	(0.001)
Ethnicity	0.089***	0.069***	0.073***	-0.148***		0.287***
	(0.002)	(0.006)	(0.002)	(0.011)		(0.006)
Coed high school	-0.241***	0.024**	-0.267***	-0.189***		0.202***
	(0.001)	(0.011)	(0.002)	(0.010)		(0.006)
Urban high school	0.507***	0.477***	0.476***	0.540***	0.511***	0.411***
	(0.001)	(0.005)	(0.001)	(0.003)	(0.023)	(0.004)
Academic degree	-0.194***	-0.180***	-0.196***	-0.148***	-0.169***	-0.242***
	(0.001)	(0.004)	(0.001)	(0.003)	(0.010)	(0.004)
School calendar from Jan to Dec	0.145***	-0.144***	0.125***	-0.085***		0.416***
	(0.002)	(0.008)	(0.002)	(0.009)		(0.011)
Size of the mine	-0.000***	-0.000***	0.000***	-0.000***	0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
School latitude	0.040***	0.025***	0.031***	-0.057***	1.079***	0.095***
	(0.001)	(0.005)	(0.001)	(0.003)	(0.032)	(0.005)
School longitude	-0.014***	0.187***	-0.016***	0.067***	0.247***	0.044***
	(0.001)	(0.006)	(0.001)	(0.004)	(0.024)	(0.007)
Constant	-0.031	7.406***	1.275***	2.345***	52.861***	-14.938***
	(0.180)	(0.920)	(0.199)	(0.685)	(2.719)	(0.895)
Observations	5,710,986	190,625	4,843,896	321,301	42,258	312,906
R^2	0.178	0.264	0.170	0.242	0.340	0.313

Note: The table displays the coefficients of interest for the size of the cohort. Panel A presents the results for the complete dataset, while Panel B divides the results by the type of extracted product. The estimations follow the specifications outlined in Equation 4.1, with the regression incorporating time and departmental controls (not shown). Co-As represents coal and asbestos. Robust standard errors are reported in parentheses. ***p<0.01, **p<0.05, *p<0.1.

Table A4.4: Saber 11 test score (OLS)



	A. Full sample	B. By extracted product				
		Co-As	Construction	Gold	Metals	Other
	(1)	(2)	(3)	(4)	(5)	(6)
Variables	Saber 11 teset score	Saber 11 teset score	Saber 11 teset score	Saber 11 teset score	Saber 11 teset score	Saber 11 teset score
Closest mine is operating	-0.174***	-0.974***	-0.088**	0.608***	-0.315	-0.060
	(0.033)	(0.205)	(0.037)	(0.140)	(0.456)	(0.151)
Distance to closest mine (in km)	-0.045***	0.038	-0.175***	-0.137***	-0.203***	0.136***
	(0.001)	(0.042)	(0.003)	(0.008)	(0.057)	(0.015)
Distance^2	0.000***	-0.000	0.000***	0.000***	0.000***	-0.000***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
Mine operation time (in years)	0.076***	0.151***	0.046***	-0.049***	-0.301***	0.045***
	(0.002)	(0.018)	(0.002)	(0.014)	(0.110)	(0.014)
Number of mines in a ratio of 1 km	-0.410***	-0.574***	-0.301***	0.010	-4.147***	-1.063***
	(0.018)	(0.078)	(0.022)	(0.047)	(0.937)	(0.146)
Number of mines in a ratio of 3 km	0.068***	0.034	0.024***	0.183***	0.566	0.607***
	(0.005)	(0.036)	(0.006)	(0.018)	(0.378)	(0.055)
Number of mines in a ratio of 5 km	0.045***	0.041*	0.064***	-0.257***	-0.364***	-0.081***
	(0.003)	(0.022)	(0.003)	(0.015)	(0.106)	(0.027)
Number of mines in a ratio of 10 km	-0.053***	0.021***	-0.069***	-0.041***	-0.012	-0.025***
	(0.001)	(0.007)	(0.001)	(0.005)	(0.037)	(0.005)
Number of mines in a ratio of 25 km	0.003***	-0.013***	0.003***	0.022***	0.040***	0.010***
	(0.000)	(0.002)	(0.000)	(0.002)	(0.008)	(0.002)
Number of mines in a ratio of 50 km	0.012***	0.004***	0.011***	0.017***	-0.024***	0.008***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.004)	(0.001)
Year of birth	0.637***	0.521***	0.639***	0.557***	0.489***	0.637***
	(0.003)	(0.016)	(0.003)	(0.012)	(0.030)	(0.013)
Female	-4.122***	-3.565***	-4.182***	-3.528***	-3.898***	-3.858***
	(0.022)	(0.119)	(0.024)	(0.089)	(0.251)	(0.096)
Public school	-1.851***	-0.912***	-1.584***	-2.897***	6.623***	-6.632***
	(0.030)	(0.206)	(0.032)	(0.177)	(0.786)	(0.162)
Household income	4.553***	4.098***	4.538***	3.839***	2.956***	4.046***
	(0.012)	(0.070)	(0.014)	(0.053)	(0.174)	(0.053)
Father's years of education	0.670***	0.497***	0.675***	0.472***	0.683***	0.732***
	(0.005)	(0.029)	(0.006)	(0.020)	(0.061)	(0.024)
Mother's years of education	0.871***	0.672***	0.874***	0.784***	1.197***	0.929***
	(0.006)	(0.031)	(0.006)	(0.022)	(0.068)	(0.026)
Ethnicity	-0.940***	-2.943***	-0.605***	-7.710***		-0.035
	(0.054)	(0.232)	(0.059)	(0.407)		(0.206)
Coed high school	-13.858***	-9.042***	-13.796***	-6.968***		-15.306***
	(0.052)	(0.461)	(0.055)	(0.357)		(0.221)
Urban high school	1.605***	3.714***	1.881***	-0.020	-0.389	-0.341**
	(0.031)	(0.187)	(0.035)	(0.110)	(0.731)	(0.153)
Academic degree	-0.865***	-0.675***	-0.959***	-1.417***	-2.370***	0.686***
	(0.026)	(0.150)	(0.029)	(0.104)	(0.332)	(0.128)
School calendar from Jan to Dec	-4.472***	-7.899***	-3.953***	-2.788***		-8.231***
	(0.071)	(0.311)	(0.077)	(0.346)		(0.404)
Size of the mine	-0.000***	-0.000	-0.002***	-0.000***	-0.002***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
School latitude	-0.952***	-2.018***	-1.563***	-3.888***	-1.945*	2.044***
	(0.027)	(0.194)	(0.032)	(0.113)	(1.046)	(0.172)
School longitude	2.424***	1.945***	2.446***	0.670***	-4.123***	3.868***
	(0.035)	(0.253)	(0.043)	(0.144)	(0.777)	(0.267)
Constant	-1,043.409***	-848.206***	-1,038.699***	-994.499***	-1,203.749***	-944.734***
	(6.317)	(37.606)	(7.012)	(25.661)	(88.077)	(32.382)
Observations	5,710,986	190,625	4,843,896	321,301	42,258	312,906
R^2	0.161	0.136	0.158	0.223	0.135	0.175

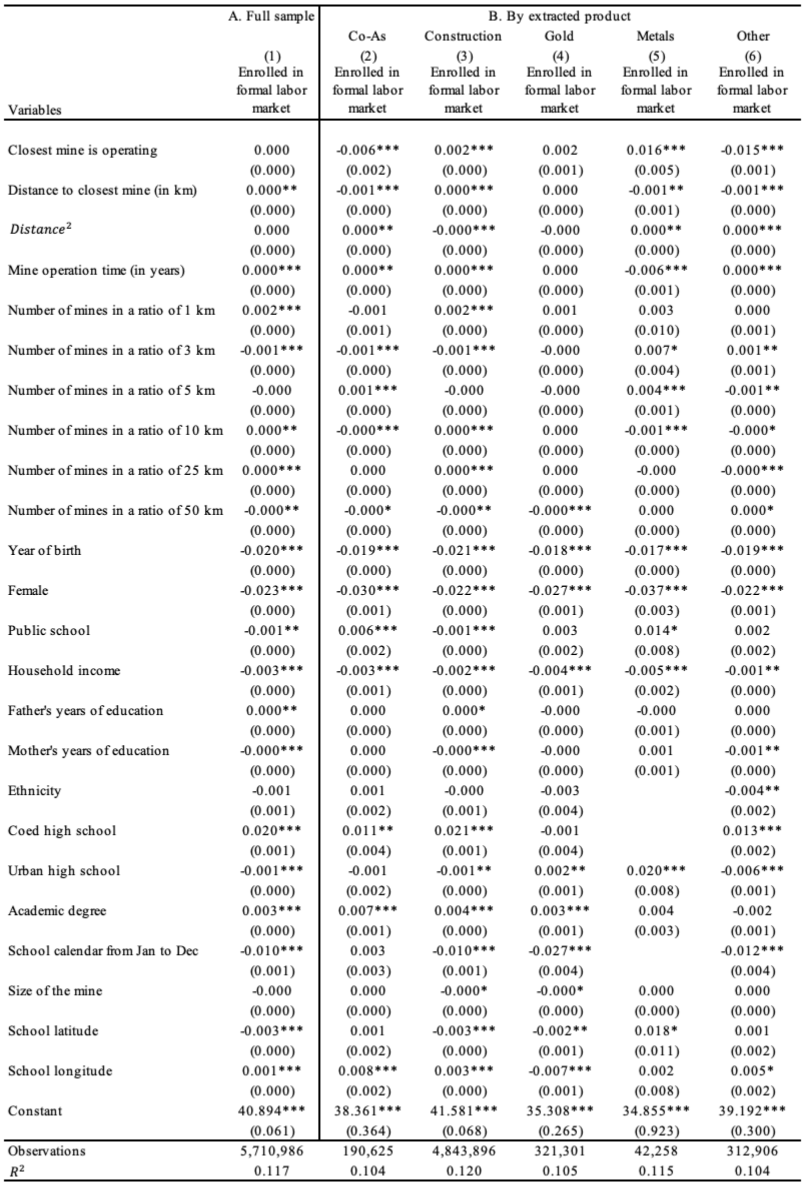
Note: The table displays the coefficients of interest for the Saber 11 test score. Panel A presents the results for the complete dataset, while Panel B divides the results by the type of extracted product. The estimations follow the specifications outlined in Equation 4.1, with the regression incorporating time and departmental controls (not shown). Co-As represents coal and asbestos. Robust standard errors are reported in parentheses. ***p<0.01, **p<0.05, *p<0.1.

Table A4.5: Enrollment in Higher Education (OLS)

	A. Full sample	B. By extracted product				
		Co-As	Construction	Gold	Metals	Other
Variables	(1) Enrolled in higher education	(2) Enrolled in higher education	(3) Enrolled in higher education	(4) Enrolled in higher education	(5) Enrolled in higher education	(6) Enrolled in higher education
Closest mine is operating	0.011***	0.004	0.011***	0.006**	-0.080***	0.007**
	(0.001)	(0.004)	(0.001)	(0.003)	(0.009)	(0.003)
Distance to closest mine (in km)	-0.000***	0.002**	-0.000	-0.001***	0.000	-0.001***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)
$Distance^2$	0.000***	-0.000***	-0.000***	0.000**	0.000	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mine operation time (in years)	0.000***	-0.000	0.000	-0.001***	0.012***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)	(0.000)
Number of mines in a ratio of 1 km	-0.004***	0.001	-0.003***	-0.002*	0.046***	-0.002
	(0.000)	(0.001)	(0.000)	(0.001)	(0.018)	(0.003)
Number of mines in a ratio of 3 km	-0.000*	-0.001	-0.000***	0.000	-0.035***	-0.005***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.007)	(0.001)
Number of mines in a ratio of 5 km	0.001***	0.000	0.001***	-0.000	0.001	0.004***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)	(0.000)
Number of mines in a ratio of 10 km	-0.000***	0.000	-0.000***	-0.000	0.003***	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Number of mines in a ratio of 25 km	0.000	-0.000***	0.000	-0.000	-0.000	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Number of mines in a ratio of 50 km	0.000***	0.000	0.000***	0.000***	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Year of birth	0.026***	0.025***	0.026***	0.026***	0.025***	0.028***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Female	-0.015***	-0.004*	-0.015***	-0.008***	-0.007	-0.015***
	(0.000)	(0.002)	(0.000)	(0.002)	(0.005)	(0.002)
Public school	-0.055***	-0.059***	-0.054***	-0.060***	0.015	-0.088***
	(0.001)	(0.004)	(0.001)	(0.003)	(0.015)	(0.003)
Household income	0.039***	0.048***	0.038***	0.036***	0.041***	0.034***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.003)	(0.001)
Father's years of education	0.003***	0.002***	0.003***	0.003***	0.002	0.002***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)
Mother's years of education	0.005***	0.005***	0.005***	0.005***	0.006***	0.005***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)
Ethnicity	-0.011***	-0.034***	-0.010***	-0.079***		0.002
	(0.001)	(0.004)	(0.001)	(0.008)		(0.004)
Coed high school	-0.099***	-0.129***	-0.099***	-0.023***		-0.095***
	(0.001)	(0.009)	(0.001)	(0.007)		(0.004)
Urban high school	0.032***	0.034***	0.035***	0.023***	-0.014	0.015***
	(0.001)	(0.003)	(0.001)	(0.002)	(0.014)	(0.003)
Academic degree	-0.008***	0.006**	-0.009***	-0.018***	-0.028***	-0.013***
	(0.000)	(0.003)	(0.001)	(0.002)	(0.006)	(0.002)
School calendar from Jan to Dec	-0.102***	-0.112***	-0.100***	-0.090***		-0.145***
	(0.001)	(0.006)	(0.001)	(0.007)		(0.007)
Size of the mine	-0.000**	-0.000***	-0.000***	0.000***	0.000*	-0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
School latitude	0.004***	-0.004	0.003***	-0.013***	-0.111***	0.009***
	(0.000)	(0.004)	(0.001)	(0.002)	(0.020)	(0.003)
School longitude	0.004***	0.010**	0.006***	-0.020***	0.004	0.025***
	(0.001)	(0.005)	(0.001)	(0.003)	(0.015)	(0.005)
Constant	-51.166***	-48.859***	-50.746***	-51.636***	-49.968***	-52.724***
	(0.114)	(0.696)	(0.127)	(0.488)	(1.646)	(0.587)
Observations	5,710,986	190,625	4,843,896	321,301	42,258	312,906
R^2	0.079	0.081	0.079	0.084	0.093	0.085

Note: The table displays the coefficients of interest for the probability of enrollment in higher education. Panel A presents the results for the complete dataset, while Panel B divides the results by the type of extracted product. The estimations follow the specifications outlined in Equation 4.1, with the regression incorporating time and departmental controls (not shown). Co-As represents coal and asbestos. Robust standard errors are reported in parentheses. ***p<0.01, **p<0.05, *p<0.1.

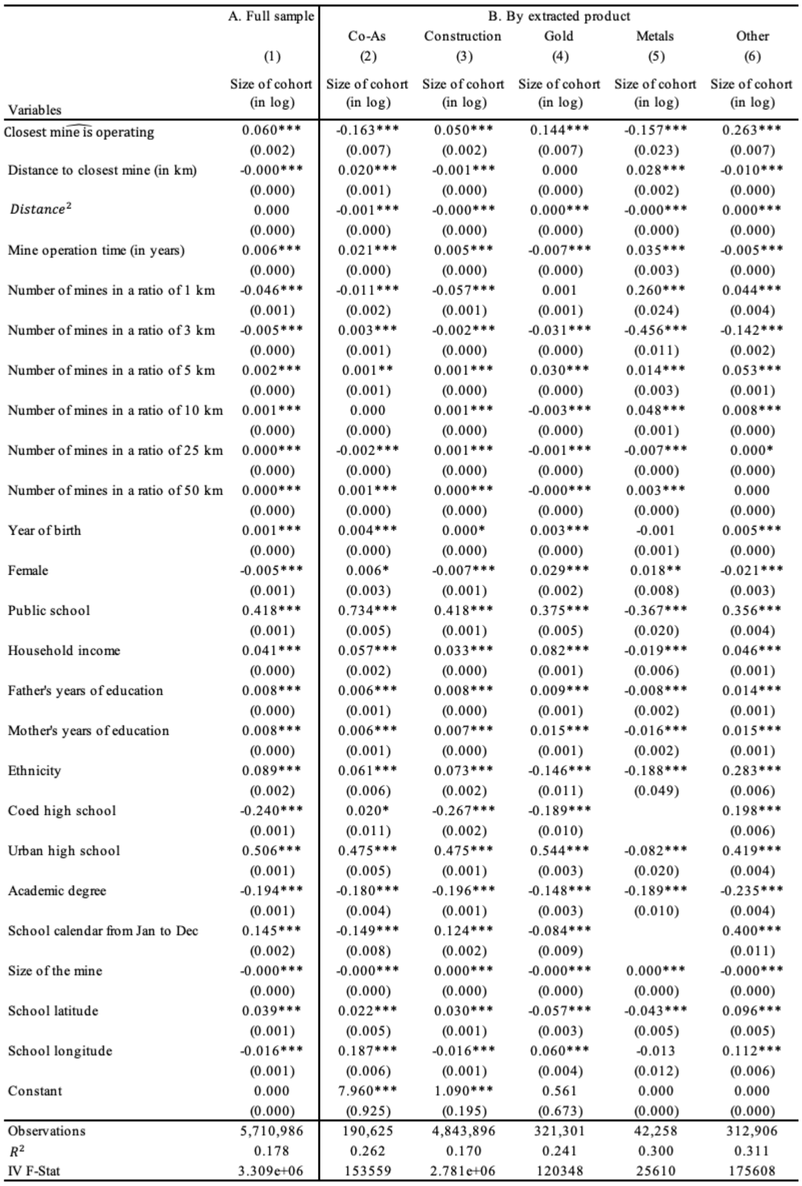
Table A4.6: Enrollment in labor market (OLS)



	A. Full sample	B. By extracted product				
		Co-As	Construction	Gold	Metals	Other
Variables	(1) Enrolled in formal labor market	(2) Enrolled in formal labor market	(3) Enrolled in formal labor market	(4) Enrolled in formal labor market	(5) Enrolled in formal labor market	(6) Enrolled in formal labor market
Closest mine is operating	0.000	-0.006***	0.002***	0.002	0.016***	-0.015***
	(0.000)	(0.002)	(0.000)	(0.001)	(0.005)	(0.001)
Distance to closest mine (in km)	0.000**	-0.001***	0.000***	0.000	-0.001**	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
$Distance^2$	0.000	0.000**	-0.000***	-0.000	0.000**	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mine operation time (in years)	0.000***	0.000**	0.000***	0.000	-0.006***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Number of mines in a ratio of 1 km	0.002***	-0.001	0.002***	0.001	0.003	0.000
	(0.000)	(0.001)	(0.000)	(0.000)	(0.010)	(0.001)
Number of mines in a ratio of 3 km	-0.001***	-0.001***	-0.001***	-0.000	0.007*	0.001**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.004)	(0.001)
Number of mines in a ratio of 5 km	-0.000	0.001***	-0.000	-0.000	0.004***	-0.001**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Number of mines in a ratio of 10 km	0.000**	-0.000***	0.000***	0.000	-0.001***	-0.000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Number of mines in a ratio of 25 km	0.000***	0.000	0.000***	0.000	-0.000	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Number of mines in a ratio of 50 km	-0.000**	-0.000*	-0.000**	-0.000***	0.000	0.000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Year of birth	-0.020***	-0.019***	-0.021***	-0.018***	-0.017***	-0.019***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female	-0.023***	-0.030***	-0.022***	-0.027***	-0.037***	-0.022***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.003)	(0.001)
Public school	-0.001**	0.006***	-0.001***	0.003	0.014*	0.002
	(0.000)	(0.002)	(0.000)	(0.002)	(0.008)	(0.002)
Household income	-0.003***	-0.003***	-0.002***	-0.004***	-0.005***	-0.001**
	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	(0.000)
Father's years of education	0.000**	0.000	0.000*	-0.000	-0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Mother's years of education	-0.000***	0.000	-0.000***	-0.000	0.001	-0.001**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Ethnicity	-0.001	0.001	-0.000	-0.003		-0.004**
	(0.001)	(0.002)	(0.001)	(0.004)		(0.002)
Coed high school	0.020***	0.011**	0.021***	-0.001		0.013***
	(0.001)	(0.004)	(0.001)	(0.004)		(0.002)
Urban high school	-0.001***	-0.001	-0.001**	0.002**	0.020***	-0.006***
	(0.000)	(0.002)	(0.000)	(0.001)	(0.008)	(0.001)
Academic degree	0.003***	0.007***	0.004***	0.003***	0.004	-0.002
	(0.000)	(0.001)	(0.000)	(0.001)	(0.003)	(0.001)
School calendar from Jan to Dec	-0.010***	0.003	-0.010***	-0.027***		-0.012***
	(0.001)	(0.003)	(0.001)	(0.004)		(0.004)
Size of the mine	-0.000	0.000	-0.000*	-0.000*	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
School latitude	-0.003***	0.001	-0.003***	-0.002**	0.018*	0.001
	(0.000)	(0.002)	(0.000)	(0.001)	(0.011)	(0.002)
School longitude	0.001***	0.008***	0.003***	-0.007***	0.002	0.005*
	(0.000)	(0.002)	(0.000)	(0.001)	(0.008)	(0.002)
Constant	40.894***	38.361***	41.581***	35.308***	34.855***	39.192***
	(0.061)	(0.364)	(0.068)	(0.265)	(0.923)	(0.300)
Observations	5,710,986	190,625	4,843,896	321,301	42,258	312,906
R^2	0.117	0.104	0.120	0.105	0.115	0.104

Note: The table displays the coefficients of interest for the probability of enrollment in the labor market. Panel A presents the results for the complete dataset, while Panel B divides the results by the type of extracted product. The estimations follow the specifications outlined in Equation 4.1, with the regression incorporating time and departmental controls (not shown). Co-As represents coal and asbestos. Robust standard errors are reported in parentheses. ***p<0.01, **p<0.05, *p<0.1.

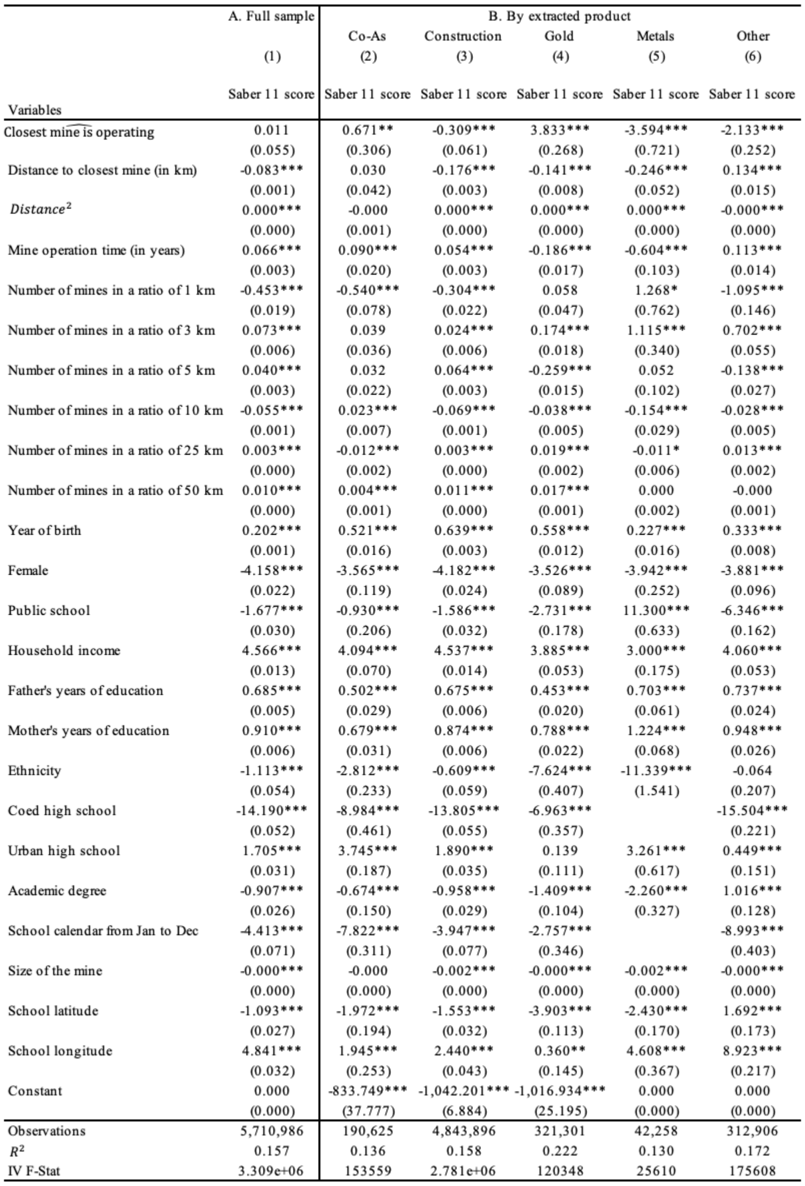
Table A4.7: Size of cohort (IV)



	A. Full sample	B. By extracted product				
		Co-As	Construction	Gold	Metals	Other
	(1)	(2)	(3)	(4)	(5)	(6)
Variables	Size of cohort (in log)	Size of cohort (in log)	Size of cohort (in log)	Size of cohort (in log)	Size of cohort (in log)	Size of cohort (in log)
Closest $\widehat{\text{mine is}}$ operating	0.060***	-0.163***	0.050***	0.144***	-0.157***	0.263***
	(0.002)	(0.007)	(0.002)	(0.007)	(0.023)	(0.007)
Distance to closest mine (in km)	-0.000***	0.020***	-0.001***	0.000	0.028***	-0.010***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.002)	(0.000)
$Distance^2$	0.000	-0.001***	-0.000***	0.000***	-0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mine operation time (in years)	0.006***	0.021***	0.005***	-0.007***	0.035***	-0.005***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.003)	(0.000)
Number of mines in a ratio of 1 km	-0.046***	-0.011***	-0.057***	0.001	0.260***	0.044***
	(0.001)	(0.002)	(0.001)	(0.001)	(0.024)	(0.004)
Number of mines in a ratio of 3 km	-0.005***	0.003***	-0.002***	-0.031***	-0.456***	-0.142***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.011)	(0.002)
Number of mines in a ratio of 5 km	0.002***	0.001**	0.001***	0.030***	0.014***	0.053***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.003)	(0.001)
Number of mines in a ratio of 10 km	0.001***	0.000	0.001***	-0.003***	0.048***	0.008***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Number of mines in a ratio of 25 km	0.000***	-0.002***	0.001***	-0.001***	-0.007***	0.000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Number of mines in a ratio of 50 km	0.000***	0.001***	0.000***	-0.000***	0.003***	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Year of birth	0.001***	0.004***	0.000*	0.003***	-0.001	0.005***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Female	-0.005***	0.006*	-0.007***	0.029***	0.018**	-0.021***
	(0.001)	(0.003)	(0.001)	(0.002)	(0.008)	(0.003)
Public school	0.418***	0.734***	0.418***	0.375***	-0.367***	0.356***
	(0.001)	(0.005)	(0.001)	(0.005)	(0.020)	(0.004)
Household income	0.041***	0.057***	0.033***	0.082***	-0.019***	0.046***
	(0.000)	(0.002)	(0.000)	(0.001)	(0.006)	(0.001)
Father's years of education	0.008***	0.006***	0.008***	0.009***	-0.008***	0.014***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	(0.001)
Mother's years of education	0.008***	0.006***	0.007***	0.015***	-0.016***	0.015***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	(0.001)
Ethnicity	0.089***	0.061***	0.073***	-0.146***	-0.188***	0.283***
	(0.002)	(0.006)	(0.002)	(0.011)	(0.049)	(0.006)
Coed high school	-0.240***	0.020*	-0.267***	-0.189***		0.198***
	(0.001)	(0.011)	(0.002)	(0.010)		(0.006)
Urban high school	0.506***	0.475***	0.475***	0.544***	-0.082***	0.419***
	(0.001)	(0.005)	(0.001)	(0.003)	(0.020)	(0.004)
Academic degree	-0.194***	-0.180***	-0.196***	-0.148***	-0.189***	-0.235***
	(0.001)	(0.004)	(0.001)	(0.003)	(0.010)	(0.004)
School calendar from Jan to Dec	0.145***	-0.149***	0.124***	-0.084***		0.400***
	(0.002)	(0.008)	(0.002)	(0.009)		(0.011)
Size of the mine	-0.000***	-0.000***	0.000***	-0.000***	0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
School latitude	0.039***	0.022***	0.030***	-0.057***	-0.043***	0.096***
	(0.001)	(0.005)	(0.001)	(0.003)	(0.005)	(0.005)
School longitude	-0.016***	0.187***	-0.016***	0.060***	-0.013	0.112***
	(0.001)	(0.006)	(0.001)	(0.004)	(0.012)	(0.006)
Constant	0.000	7.960***	1.090***	0.561	0.000	0.000
	(0.000)	(0.925)	(0.195)	(0.673)	(0.000)	(0.000)
Observations	5,710,986	190,625	4,843,896	321,301	42,258	312,906
R^2	0.178	0.262	0.170	0.241	0.300	0.311
IV F-Stat	3.309e+06	153559	2.781e+06	120348	25610	175608

Note: The table displays the coefficients of interest for the Size of cohort. Panel A presents the results for the complete dataset, while Panel B divides the results by the type of extracted product. The estimations follow the specifications outlined in Equation 4.3, with the regression incorporating time and departmental controls (not shown). First step for the IV approach in Table A4.2. Co-As represents coal and asbestos. Robust standard errors are reported in parentheses. ***p<0.01, **p<0.05, *p<0.1.

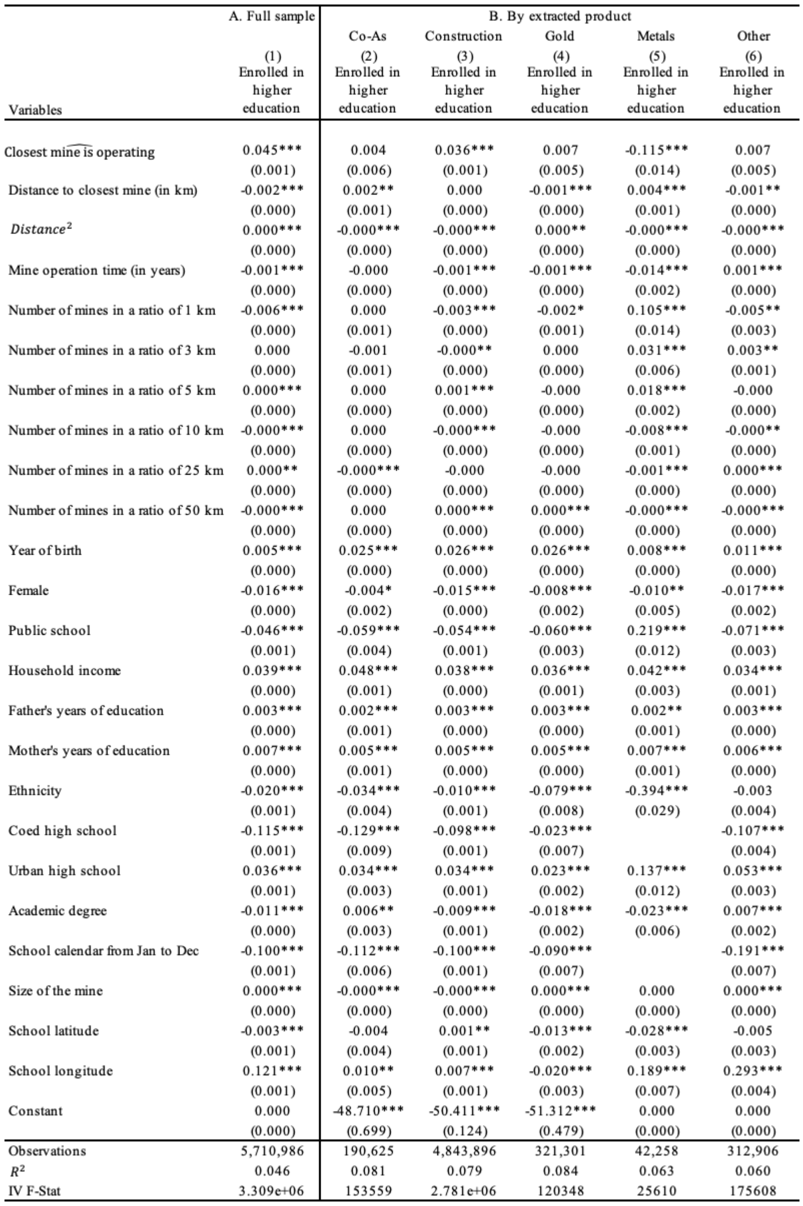
Table A4.8: Saber 11 test score (IV)



	A. Full sample	B. By extracted product				
		Co-As	Construction	Gold	Metals	Other
	(1)	(2)	(3)	(4)	(5)	(6)
Variables	Saber 11 score	Saber 11 score	Saber 11 score	Saber 11 score	Saber 11 score	Saber 11 score
Closest $\widehat{\text{mine is}}$ operating	0.011	0.671**	-0.309***	3.833***	-3.594***	-2.133***
	(0.055)	(0.306)	(0.061)	(0.268)	(0.721)	(0.252)
Distance to closest mine (in km)	-0.083***	0.030	-0.176***	-0.141***	-0.246***	0.134***
	(0.001)	(0.042)	(0.003)	(0.008)	(0.052)	(0.015)
$Distance^2$	0.000***	-0.000	0.000***	0.000***	0.000***	-0.000***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
Mine operation time (in years)	0.066***	0.090***	0.054***	-0.186***	-0.604***	0.113***
	(0.003)	(0.020)	(0.003)	(0.017)	(0.103)	(0.014)
Number of mines in a ratio of 1 km	-0.453***	-0.540***	-0.304***	0.058	1.268*	-1.095***
	(0.019)	(0.078)	(0.022)	(0.047)	(0.762)	(0.146)
Number of mines in a ratio of 3 km	0.073***	0.039	0.024***	0.174***	1.115***	0.702***
	(0.006)	(0.036)	(0.006)	(0.018)	(0.340)	(0.055)
Number of mines in a ratio of 5 km	0.040***	0.032	0.064***	-0.259***	0.052	-0.138***
	(0.003)	(0.022)	(0.003)	(0.015)	(0.102)	(0.027)
Number of mines in a ratio of 10 km	-0.055***	0.023***	-0.069***	-0.038***	-0.154***	-0.028***
	(0.001)	(0.007)	(0.001)	(0.005)	(0.029)	(0.005)
Number of mines in a ratio of 25 km	0.003***	-0.012***	0.003***	0.019***	-0.011*	0.013***
	(0.000)	(0.002)	(0.000)	(0.002)	(0.006)	(0.002)
Number of mines in a ratio of 50 km	0.010***	0.004***	0.011***	0.017***	0.000	-0.000
	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	(0.001)
Year of birth	0.202***	0.521***	0.639***	0.558***	0.227***	0.333***
	(0.001)	(0.016)	(0.003)	(0.012)	(0.016)	(0.008)
Female	-4.158***	-3.565***	-4.182***	-3.526***	-3.942***	-3.881***
	(0.022)	(0.119)	(0.024)	(0.089)	(0.252)	(0.096)
Public school	-1.677***	-0.930***	-1.586***	-2.731***	11.300***	-6.346***
	(0.030)	(0.206)	(0.032)	(0.178)	(0.633)	(0.162)
Household income	4.566***	4.094***	4.537***	3.885***	3.000***	4.060***
	(0.013)	(0.070)	(0.014)	(0.053)	(0.175)	(0.053)
Father's years of education	0.685***	0.502***	0.675***	0.453***	0.703***	0.737***
	(0.005)	(0.029)	(0.006)	(0.020)	(0.061)	(0.024)
Mother's years of education	0.910***	0.679***	0.874***	0.788***	1.224***	0.948***
	(0.006)	(0.031)	(0.006)	(0.022)	(0.068)	(0.026)
Ethnicity	-1.113***	-2.812***	-0.609***	-7.624***	-11.339***	-0.064
	(0.054)	(0.233)	(0.059)	(0.407)	(1.541)	(0.207)
Coed high school	-14.190***	-8.984***	-13.805***	-6.963***		-15.504***
	(0.052)	(0.461)	(0.055)	(0.357)		(0.221)
Urban high school	1.705***	3.745***	1.890***	0.139	3.261***	0.449***
	(0.031)	(0.187)	(0.035)	(0.111)	(0.617)	(0.151)
Academic degree	-0.907***	-0.674***	-0.958***	-1.409***	-2.260***	1.016***
	(0.026)	(0.150)	(0.029)	(0.104)	(0.327)	(0.128)
School calendar from Jan to Dec	-4.413***	-7.822***	-3.947***	-2.757***		-8.993***
	(0.071)	(0.311)	(0.077)	(0.346)		(0.403)
Size of the mine	-0.000***	-0.000	-0.002***	-0.000***	-0.002***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
School latitude	-1.093***	-1.972***	-1.553***	-3.903***	-2.430***	1.692***
	(0.027)	(0.194)	(0.032)	(0.113)	(0.170)	(0.173)
School longitude	4.841***	1.945***	2.440***	0.360**	4.608***	8.923***
	(0.032)	(0.253)	(0.043)	(0.145)	(0.367)	(0.217)
Constant	0.000	-833.749***	-1,042.201***	-1,016.934***	0.000	0.000
	(0.000)	(37.777)	(6.884)	(25.195)	(0.000)	(0.000)
Observations	5,710,986	190,625	4,843,896	321,301	42,258	312,906
R^2	0.157	0.136	0.158	0.222	0.130	0.172
IV F-Stat	3.309e+06	153559	2.781e+06	120348	25610	175608

Note: The table displays the coefficients of interest for the Saber 11 test score. Panel A presents the results for the complete dataset, while Panel B divides the results by the type of extracted product. The estimations follow the specifications outlined in Equation 4.3, with the regression incorporating time and departmental controls (not shown). First step for the IV approach in Table A4.2. Co-As represents coal and asbestos. Robust standard errors are reported in parentheses. ***p<0.01, **p<0.05, *p<0.1.

Table A4.9: Enrollment in Higher Education (IV)



	A. Full sample	B. By extracted product				
		Co-As	Construction	Gold	Metals	Other
	(1)	(2)	(3)	(4)	(5)	(6)
Variables	Enrolled in higher education	Enrolled in higher education	Enrolled in higher education	Enrolled in higher education	Enrolled in higher education	Enrolled in higher education
Closest $\widehat{\text{mine is}}$ operating	0.045***	0.004	0.036***	0.007	-0.115***	0.007
	(0.001)	(0.006)	(0.001)	(0.005)	(0.014)	(0.005)
Distance to closest mine (in km)	-0.002***	0.002**	0.000	-0.001***	0.004***	-0.001**
	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)
$Distance^2$	0.000***	-0.000***	-0.000***	0.000**	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mine operation time (in years)	-0.001***	-0.000	-0.001***	-0.001***	-0.014***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)	(0.000)
Number of mines in a ratio of 1 km	-0.006***	0.000	-0.003***	-0.002*	0.105***	-0.005**
	(0.000)	(0.001)	(0.000)	(0.001)	(0.014)	(0.003)
Number of mines in a ratio of 3 km	0.000	-0.001	-0.000**	0.000	0.031***	0.003**
	(0.000)	(0.001)	(0.000)	(0.000)	(0.006)	(0.001)
Number of mines in a ratio of 5 km	0.000***	0.000	0.001***	-0.000	0.018***	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)	(0.000)
Number of mines in a ratio of 10 km	-0.000***	0.000	-0.000***	-0.000	-0.008***	-0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Number of mines in a ratio of 25 km	0.000**	-0.000***	-0.000	-0.000	-0.001***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Number of mines in a ratio of 50 km	-0.000***	0.000	0.000***	0.000***	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Year of birth	0.005***	0.025***	0.026***	0.026***	0.008***	0.011***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female	-0.016***	-0.004*	-0.015***	-0.008***	-0.010**	-0.017***
	(0.000)	(0.002)	(0.000)	(0.002)	(0.005)	(0.002)
Public school	-0.046***	-0.059***	-0.054***	-0.060***	0.219***	-0.071***
	(0.001)	(0.004)	(0.001)	(0.003)	(0.012)	(0.003)
Household income	0.039***	0.048***	0.038***	0.036***	0.042***	0.034***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.003)	(0.001)
Father's years of education	0.003***	0.002***	0.003***	0.003***	0.002**	0.003***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)
Mother's years of education	0.007***	0.005***	0.005***	0.005***	0.007***	0.006***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)
Ethnicity	-0.020***	-0.034***	-0.010***	-0.079***	-0.394***	-0.003
	(0.001)	(0.004)	(0.001)	(0.008)	(0.029)	(0.004)
Coed high school	-0.115***	-0.129***	-0.098***	-0.023***		-0.107***
	(0.001)	(0.009)	(0.001)	(0.007)		(0.004)
Urban high school	0.036***	0.034***	0.034***	0.023***	0.137***	0.053***
	(0.001)	(0.003)	(0.001)	(0.002)	(0.012)	(0.003)
Academic degree	-0.011***	0.006**	-0.009***	-0.018***	-0.023***	0.007***
	(0.000)	(0.003)	(0.001)	(0.002)	(0.006)	(0.002)
School calendar from Jan to Dec	-0.100***	-0.112***	-0.100***	-0.090***		-0.191***
	(0.001)	(0.006)	(0.001)	(0.007)		(0.007)
Size of the mine	0.000***	-0.000***	-0.000***	0.000***	0.000	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
School latitude	-0.003***	-0.004	0.001**	-0.013***	-0.028***	-0.005
	(0.001)	(0.004)	(0.001)	(0.002)	(0.003)	(0.003)
School longitude	0.121***	0.010**	0.007***	-0.020***	0.189***	0.293***
	(0.001)	(0.005)	(0.001)	(0.003)	(0.007)	(0.004)
Constant	0.000	-48.710***	-50.411***	-51.312***	0.000	0.000
	(0.000)	(0.699)	(0.124)	(0.479)	(0.000)	(0.000)
Observations	5,710,986	190,625	4,843,896	321,301	42,258	312,906
R^2	0.046	0.081	0.079	0.084	0.063	0.060
IV F-Stat	3.309e+06	153559	2.781e+06	120348	25610	175608

Note: The table displays the coefficients of interest for the probability of enrollment in higher education. Panel A presents the results for the complete dataset, while Panel B divides the results by the type of extracted product. The estimations follow the specifications outlined in Equation 4.3, with the regression incorporating time and departmental controls (not shown). First step for the IV approach in Table A4.2. Co-As represents coal and asbestos. Robust standard errors are reported in parentheses. ***$p<0.01$, **$p<0.05$, *$p<0.1$.

Table A4.10: Enrollment in labor market (IV)

	A. Full sample	B. By extracted product				
		Co-As	Construction	Gold	Metals	Other
Variables	(1) Enrolled in formal labor market	(2) Enrolled in formal labor market	(3) Enrolled in formal labor market	(4) Enrolled in formal labor market	(5) Enrolled in formal labor market	(6) Enrolled in formal labor market
Closest $\widehat{\text{mine is}}$ operating	-0.002***	-0.010***	0.011***	0.010***	0.046***	-0.016***
	(0.001)	(0.003)	(0.001)	(0.003)	(0.008)	(0.002)
Distance to closest mine (in km)	0.001***	-0.001***	0.000***	0.000	-0.005***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
$Distance^2$	-0.000***	0.000**	-0.000***	-0.000	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mine operation time (in years)	0.000***	0.001***	-0.000***	-0.000	0.010***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Number of mines in a ratio of 1 km	0.003***	-0.001	0.002***	0.001*	-0.031***	0.002*
	(0.000)	(0.001)	(0.000)	(0.000)	(0.008)	(0.001)
Number of mines in a ratio of 3 km	-0.001***	-0.001***	-0.001***	-0.000	-0.037***	-0.004***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.004)	(0.001)
Number of mines in a ratio of 5 km	0.000***	0.001***	-0.000	-0.000	-0.006***	0.003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Number of mines in a ratio of 10 km	0.000***	-0.000***	0.000***	0.000	0.005***	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Number of mines in a ratio of 25 km	0.000	0.000	0.000***	0.000	0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Number of mines in a ratio of 50 km	0.000***	-0.000	-0.000**	-0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Year of birth	-0.003***	-0.019***	-0.021***	-0.018***	-0.005***	-0.007***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female	-0.021***	-0.030***	-0.022***	-0.027***	-0.035***	-0.020***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.003)	(0.001)
Public school	-0.007***	0.006***	-0.001***	0.003*	-0.100***	-0.012***
	(0.000)	(0.002)	(0.000)	(0.002)	(0.007)	(0.002)
Household income	-0.003***	-0.003***	-0.002***	-0.004***	-0.007***	-0.001**
	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	(0.001)
Father's years of education	-0.000***	0.000	0.000	-0.000	-0.001	-0.000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Mother's years of education	-0.002***	0.000	-0.000***	-0.000	0.000	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Ethnicity	0.006***	0.001	0.000	-0.003	0.254***	-0.000
	(0.001)	(0.002)	(0.001)	(0.004)	(0.017)	(0.002)
Coed high school	0.033***	0.011**	0.021***	-0.001		0.022***
	(0.001)	(0.004)	(0.001)	(0.004)		(0.002)
Urban high school	-0.005***	-0.001	-0.001***	0.003**	-0.059***	-0.035***
	(0.000)	(0.002)	(0.000)	(0.001)	(0.007)	(0.001)
Academic degree	0.005***	0.007***	0.004***	0.003***	0.003	-0.016***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.004)	(0.001)
School calendar from Jan to Dec	-0.012***	0.003	-0.010***	-0.027***		0.022***
	(0.001)	(0.003)	(0.001)	(0.004)		(0.004)
Size of the mine	-0.000***	0.000	-0.000	-0.000	0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
School latitude	0.003***	0.000	-0.003***	-0.003**	0.016***	0.011***
	(0.000)	(0.002)	(0.000)	(0.001)	(0.002)	(0.002)
School longitude	-0.093***	0.008***	0.003***	-0.008***	-0.112***	-0.194***
	(0.000)	(0.002)	(0.000)	(0.001)	(0.004)	(0.002)
Constant	0.000	38.306***	41.555***	35.243***	0.000	0.000
	(0.000)	(0.366)	(0.066)	(0.260)	(0.000)	(0.000)
Observations	5,710,986	190,625	4,843,896	321,301	42,258	312,906
R^2	0.047	0.104	0.120	0.105	0.069	0.052
IV F-Stat	3.309e+06	153559	2.781e+06	120348	25610	175608

Note: The table displays the coefficients of interest for the probability of enrollment in the labor market. Panel A presents the results for the complete dataset, while Panel B divides the results by the type of extracted product. The estimations follow the specifications outlined in Equation 4.3, with the regression incorporating time and departmental controls (not shown). First step for the IV approach in Table A4.2. Co-As represents coal and asbestos. Robust standard errors are reported in parentheses. ***p<0.01, **p<0.05, *p<0.1.

Chapter 5

Conclusions

This thesis examines the impact of specific events during secondary and higher education on academic performance, education system indicators, and salaries. The analyses in all three essays are based on sophisticated databases that use the individual as the unit of analysis. These databases are unique worldwide and enable a detailed examination of these phenomena.

The first essay focuses on the impact of SPADIES, a program implemented to reduce dropouts in higher education across Colombia. There are no similar studies in the literature, and the few documents in this area are limited in scope. This study, in contrast, uses a comprehensive dataset covering the entire country for more than ten years to demonstrate the program's effectiveness in reducing the dropout rate and improving graduation rates. The collaborative efforts of the Ministry of Education, academia, and higher education institutions were crucial for the program's success. The evidence presented in this chapter indicates that the program has significant implications for future student income.

Chapter 3 demonstrates the importance of college graduation for an individual's future earnings. The improvements resulting from the SPADIES program were particularly pronounced among the most vulnerable populations. The program helped more students stay in school and enabled many students who had previously dropped out to return and graduate.

Over the span of ten years, SPADIES played a pivotal role in averting the dropout of around 14,000 students, supporting the graduation of 12,000 individuals, and ensuring the timely completion of studies for 8,000 students. These outcomes are especially significant when considering that the typical size of a Colombian higher education institution is 8,000 students. Furthermore, the cost incurred per college graduate student is noteworthy. SPADIES cost the Colombian Government only USD$ 374 in 2022 prices per student saved from dropout and graduated, a figure slightly surpassing 2 minimum monthly salaries (USD$ 235 in 2022). The primary limitation of this study pertains to the limited time horizon of the available data. Although the dataset covers almost 20 years, the requirement of complete cohort information to accounting for attrition or graduation and the waiting period to determine dropout status restricts the analysis to a maximum of 10 years. Future research would benefit from access to more recent data to evaluate the program's long-term effects. Moreover, incorporating dynamic individual information, such as employment status while studying, household income, and the residential location at the beginning of each academic year could provide additional insights. Furthermore, the Saber Pro test score, which serves as the college exit exam, represents an interesting variable to include in future analyses. While the first essay demonstrated the benefits of using SPADIES to improve efficiency, further research is needed to clarify its effects on educational quality. The second essay presents a longitudinal analysis of the impact of higher education on students who completed secondary education in Colombia from 2002 to 2015. The study provides detailed information about the secondary schools, and higher education institutions (HEIs) attended by students, their academic performance, and salaries received. The essay aims to investigate two main areas: (i) the returns to higher education, by comparing the outcomes of students who pursued higher education with those of their peers who did not; and (ii) the returns to obtaining a degree, by examining the sheepskin effect and comparing the earnings of students who received a degree with those who did not but had completed the majority of their academic program.

The contribution of this essay to the literature is significant, as it provides a comprehensive analysis of a country's total population over an extended time horizon, with a level of detail not previously seen in the literature. Previous studies have focused only on students who have graduated from college, while this dissertation follows students from the moment they finish secondary school. Additionally, the essay provides an excellent case study for the sheepskin effect. The Colombian experience allows for comparing individuals without a degree to those with the same level of education who have obtained one. This study's findings demonstrate the importance of timely graduation, as it significantly improves students' income. The analysis also reveals that the sheepskin effect and returns to education are very similar. The income of absentees and candidates (dropouts in both cases) is comparable to that of students who never attended higher education. Furthermore, the study notes a slow improvement in the gender wage gap and social mobility since implementing the SPADIES program.

However, the time horizon is a significant limitation of the essay, as the analysis only includes data up to 2013. Incorporating salary information through 2018 or 2020 would extend the longitudinal analysis, providing a more robust understanding of the relationship between education and income. Additionally, including data on doctoral degree earnings, Saber Pro test scores, and the field of work of the student (when in the labor market) would also strengthen the analysis, enabling the measurement of college student quality while controlling for the field of work. These additions would complement previous research, such as that by MacLeod et al. (2017), and extend the time horizon to include college students who did not graduate.

The third essay of this dissertation delves into an investigation of the influence of mining activities on educational and labor market outcomes, specifically examining academic performance and the pathways to higher education or the formal labor market. The notable findings of this study underscore substantial variations across different types of mined products and their effects on diverse outcomes. The presence of an operational mine in close proximity manifests a positive impact, resulting in an annual increase in cohort size without compromising academic performance, as measured by the Saber 11 test score. Moreover, it contributes to enhancing the quality of education, as indicated by the heightened probability of college enrollment. This effect persists even as the demand for education surpasses the fixed supply, a matter of particular significance given the time-consuming nature of establishing new educational institutions. Remarkably, the educational system has adeptly navigated potential challenges associated with cohort overcrowding. This has resulted in a heightened participation of secondary graduates in college, concurrently diminishing the likelihood of immediate entry into the formal labor market. These consistent trends suggest a deliberate choice by students to prioritize continued education over immediate employment, emphasizing the positive impact of mining on educational aspirations.

The findings of this study have important implications for education system planning, policymakers, practitioners, administrators, and students and their families. The study provides valuable insights into the effectiveness of anti-dropout strategies and programs in higher education. It can guide the government's decisions on where to build schools or where families should reside. By demonstrating the value of a college degree, this study can motivate students to pursue higher education and encourage countries and universities worldwide to implement programs that reduce barriers to graduation. Finally, this study's results can motivate policymakers to strive for greater improvements in the higher education system.

References

Becker, Gary, “Investment in Human Capital: A Theoretical Analysis”, Journal of Political Economy, 1962, 70 (S5), 9.

Epple, Dennis, Richard Romano, and Holger Sieg, “Admission, tuition, and financial aid policies in the market for higher education”, 2006.

Ferreyra, María Marta, Ciro Avitabile, Javier Botero Álvarez, Francisco Haimovich Paz, and Sergio Urzúa, “Momento decisivo: La educación superior en América Latina y el Caribe”, 2017.

Herrera-Prada, Luis Omar, and Bernardo Kugler, “¿Qué tanto nos sirve el programa Ser Pilo Paga?” Razón Pública, 2017. https://razonpublica.com/que-tanto-nos-sirve-el-programa-ser-pilo-paga/

Lucio, Ricardo and Mariana Serrano, “La Educación Superior: Tendencias y Políticas Estatales“,1992.

MacLeod, William Bentley and Miguel Urquiola, “Reputation and school competition”, 2015.

MacLeod, William Bentley, Evan Riehl, Juan Saavedra, and Miguel Urquiola, “The big sort: College reputation and labor market outcomes,” American Economic Journal: Applied Economics, 2017, 9 (3), 223–261.

Ministerio de Educación Nacional, “La Revolución Educativa 2002 - 2010. Informe de gestión”, Technical Report, Bogotá, 2010.

Morley, Samuel,”El problema de la distribución del ingreso en América Latina”. CEPAL, 2000.

Orozco Silva, Luis Enrique, “La Política de Cobertura: eje de la revolución educativa, 2002-2008”, Bogotá: Ediciones Uniandes, 2010.

Schultz, Theodore,”Investing in People: The Economics of Population Quality.” In Univ of California (Ed.), University of California Press, 1981.

Thurow, Lester,“Inversión en capital humano.” Trillas, 1978.

Weller, Jurgen,”Reformas económicas, crecimiento y empleo: los mercados de trabajo en América Latina y el Caribe”. CEPAL, 2000.

Declaration

1. I, hereby, declare that this dissertation has no not been presented to any other examining body either in its present or a similar form.
   Furthermore, I also affirm that I have not applied for a Ph.D. at any other higher school of education.

   Göttingen, 1.12.2023

   (Signature)

   LUIS OMAR HERRERA PRADA

   (Name in block capitals)

2. I, hereby, solemnly declare that this dissertation was undertaken independently and without any unauthorized aid.

   Göttingen, 1.12.2023

   (Signature)

   LUIS OMAR HERRERA PRADA

   (Name in block capitals)

3. I hereby declare that the digital version of this dissertation matches the printed version of this dissertation.

   Göttingen, 1.12.2023

   (Signature)

   LUIS OMAR HERRERA PRADA

   (Name in block capitals)