AARHUS UNIVERSITY

MASTER'S THESIS

ESTIMATING CAUSAL EFFECTS OF PRESCHOOL ON SCHOOL DROPOUT USING NON-EXPERIMENTAL DATA FROM ANDHRA PRADESH

FRANCISCO CARBALLO SANTIAGO

[201500080]

Supervised by: Nabanita Datta Gupta

This thesis is submitted for:

Master's Degree Program in Economics and Management

Deadline: August 15, 2017

SCHOOL OF BUSINESS AND SOCIAL SCIENCES

AARHUS UNIVERSITY

Abstract

There were 112 million children out of school all over the world in 2012. Aware of the benefits of education on future outcomes, it is of utmost importance to reduce the number of children out of school in order to reduce inequalities. While remedial education, conditional cash transfers and demand- and supply-side inputs have been the central point of educational policies in developing countries, preschooling has been barely studied. Using propensity score matching on non-experimental data from the Young Lives project, I estimate the effect of preschooling in the state of Andhra Pradesh, India. The richness and features of the data allow me to account for potential confounders. The estimation suggests that preschool affects dropouts significantly, reducing them by a 7%. Therefore, school attainment can be improved by developing a strong preschool system and providing quality preschool years to children.

As for the purpose of recognition of the data used in this thesis:

"The data used in this publication come from Young Lives, a 15-year survey investigating the changing nature of childhood poverty in Ethiopia, India (Andhra Pradesh), Peru and Vietnam, based at the University of Oxford (www.younglives.org.uk). Young Lives is core funded by the UK Department for International Development. The views expressed here are those of the author(s). They are not necessarily those of the Young Lives project, the University of Oxford, DFID or other funders."

Contents

1 Introduction

From the beginning of the 21st century, there have been increasing efforts on the achievement of universal education among children all over the world. This goal has been achieved in the majority of the developed world, even before the 21st century. There have also been large progresses in some of the developing regions, like South and West Asian countries where net enrollment ratio increased from 78 percent to 94 percent between 1999 and 2012. However, there are still regions on Earth struggling to bring universal education to their youth, particularly in sub-Saharan Africa (UNESCO, 2015b). According to UNESCO, nearly 112 million children in primary and secondary school-ages were out of school in 2012. Another striking fact is that one in six children in low and middle income countries will not complete primary education by 2015[1].

Many studies have explored the returns to education all over the world, and most of them have found significant benefits from education, both in the short and long term. Education is linked to increases in life-income, low criminality levels or reduction of unwanted pregnancies, and all together, contributes to reduce poverty[2]. Hence, it seems that providing access to education to everyone in the world, including those 100 million children who will not complete primary education, has considerable benefits for everyone.

In order to keep progressing on the universal education goal, a lot of initiatives and policies have taken place, involving the creation of new schools, provision of more teachers, textbooks, uniforms, as well as reforms in education systems, public provision of education or government subsidies (e.g. vouchers). As it will be shown in the coming

[1]Estimation based on UNESCO (2015b), Education for All Report assessing achievements of EFA goals in 2015.

[2]See Card (1999) for an explanation of the theoretical model and literature review on returns to education.

sections, some are more effective than others while others are not effective at all.

This thesis is centered around the idea that preparatory years before elementary school might ease and smooth the children's transition from home to school, as well as helping them to go through school period with greater capability, and therefore, reducing or eliminating some of the possible dropout reasons like school boredom, uselessness or excessively difficult curricula. If this idea is correct, increasing Early Childhood Care and Education in general, and preschool attendance and quality in particular, will take us a step further towards the universal education goal. At the same time, it will provide governments an additional instrument to influence and increase school participation and completion. As it will be shown later, preschool not only has to do with school completion, but might be related to other later-life outcomes like criminality, delinquency or early pregnancy. Using an extensive longitudinal dataset from the Young Lives Project[3] in India, particularly the state of Andhra Pradesh, I explore the effects of preschool attendance on dropouts and on primary and secondary school completion. To do so, propensity score matching is used. This method is convenient when two different populations present different characteristics so that a simple mean comparison would not suffice to detect actual casual effects. Using propensity score matching, observations are matched or weighted according to their probability of treatment given their characteristics or covariates.

The thesis is structured as follows. After this introduction, Section 2 aims at describe Early Childhood Development in India as well as the motivation for the thesis. I present the background characteristics of the country, together with facts and particularities of the Indian institutions and education in general. Additionally, it explains the concept of Early Childhood Care and Education (ECCE) and the progress done on

[3] www.younglives.org.uk

the main education goals, set by the Education for All initiative and the Millennium Development Goals. Section 3 reviews recent literature on preschool, both in developed countries, where more studies are available, and in developing countries, where this topic has been less studied. It shows how large-scale preschool programs like Head Start have contributed to children's late-life outcomes like reduction in delinquency or income increments in the US. But also how smaller preschool programs in developing countries have successfully increased school attainment and learning outcomes. Section 4 presents the data used in this thesis and the Young Lives project itself, as well as the sampling method and data collection approach. The section concludes with the presentation of some descriptive statistics of the data, intended to give a more precise understanding of the sample.

Section 5 introduces the empirical strategy followed in the thesis. Firstly, outcomes and treatment are presented. In the following, I introduce the identification of the parameters of interest, as well as the methods used to estimate them, particularly propensity score matching. This section also covers the different matching methods used, as well as some important aspects of the propensity score. The last part explains the covariates choice and elaborates on different choices of them. In Section 6, I present the results of the propensity score matching estimation. I use dropout and completion rates as dependent variables, and show, under certain restrictive assumptions, that preschool have an arguably significant causal effect on dropout reduction. Additionally I present the results on different subgroups of the population that might have important implications for future policies.

Robustness tests are presented in Section 7 with the purpose of confirming consistency and validity. In Section 8, I discuss the implications as well as limitations of the thesis, and possible critiques to the method and assumptions taken. Additionally, I set the bases for future research using the same data and the same approach, opening new possibilities for preschool education. Lastly, I conclude on the thesis and the results.

2 Background

This section will guide the reader through an overview and explanation of how India has performed in terms of education in the past decades. The main objective is to show that although extensive gains have been achieved there is still a long way to go and to improve. It will also provide enough information to understand the importance of education in developing countries and India in particular, a 1.2 billion population country.

Preschool is a large part of ECCE, which stands for Early Childhood Care and Education. There is not an official, general definition of what Early Childhood Care and Education means, supports or stands up for. Different institutions describe it similarly, but not identically. According to UNESCO, ECCE is described as:

"Early childhood is defined as the period from birth to eight years old. A time of remarkable brain growth, these years lay the basis for subsequent development. Early childhood care and education (ECCE) is more than a preparatory stage assisting the child's transition to formal schooling. It places emphasis on developing the whole child - attending to his or her social, emotional, cognitive and physical needs - to establish a solid and broad foundation for lifelong learning and well-being."

It is then intended to help children's growth, survival, nutrition, health, hygiene, cognitive, social, physical and emotional development. And it does so through programs that range from parental focus, to community-based child care and formal pre-primary education. It is usually divided into two age groups, children under 3 years old and children above 3 years old and until 6-8 years old, the usual school entry age. The latter groups is the main target of this thesis.

According to the UNESCO[4], a child in the developing world has 40% chance of

[4]UNESCO (2015b)

living in extreme poverty[5]. Moreover, 6.3 million children died in 2013 before the age of 5. Early childhood is a really important stage in children's lives, and it lays the foundation for future development, learning and stimulation. Heckman *et al.* (2012) explains the benefits from early investments compared to later investments, using US experiments like High/Scope Perry Preschool program as empirical examples. Cunha and Heckman (2007) explains that skill begets skill and estimates the technologies of skill formation. According to them, the skills of children develop based on three elements: the stock of skills already accumulated, the investment made by their parents and the stock of skills accumulated by their parents. Therefore, if future progress and child development depend on the skills already acquired, the earlier those skills come up and are fostered, the more they can develop. Hence, improvements in ECCE might bring new opportunities to all those children, helping them to compensate early adversities and inequalities. Additionally, it has been demonstrated that policies based on ECCE may be more cost-efficient than policies targeted at later life stages[6].

One of the main components of ECCE is early childhood education, which applies to children under 6 years old. The main goal is to prepare children for the later school standards and requirements. Although this thesis focuses on preschools, early childhood education also covers kindergartens, childcare houses, home visiting and parenting. Usually, these policies are directed at improve school readiness of children and have special focus on disadvantaged families. These households are the population at risk and placing them on equal footing as the rest of the families is one of the main objectives of education. Additionally, ECCE has been demonstrated to have greater returns than other policies later in life (Heckman *et al.*, 2012; Elango *et al.*, 2015). While

[5]Extreme poverty is defined as living on less than US $1 a day

[6]See Engle *et al.* (2011) for cost-benefit analysis on preschool intervention and Cunha *et al.* (2006) for explanations on returns to early investments and remedies later in life.

a government that is concerned with equity wants to compensate for differences in final outcomes, this approach might be more expensive and less efficient that a compensation early in life. If those differences are already placed when one is born, and they are likely to increase with time, then it is a superior approach to tackle them earlier. Currie (2001) discusses a vast number of early intervention programs that have taken place in the US.

Furthermore, at the end of the last century, advances in brain research proved that early life matters for future child development, so that what happens in the first years of life can have important effects on the rest of our lives. Therefore, we have to strive at making the best positive impact. The old thinking that genes determine our development and that our future is much ruled by those gens, with little variability, is no longer present and brain research has demonstrated that the brain is affected by its environment and we are able to influence it. For a discussion, see Shore (1997).

2.1 Education for All and Millennium Development Goals

Previously, it has been argued that early childhood is an important stage in children's lives. I will now present two main initiatives established at the beginning of the 21st century to pursue the achievement of universal education among others. On the one hand, in 1990, the international community adopted the World Declaration on Education for All at Jomtien, Thailand. It emphasized the key role of education to sustainable development, equality and future. During the World Education Forum in Dakar in 2000, the Education for All initiative set six main goals[7], to be achieved by 2015, ECCE being the first one. The initiative was supported by 164 countries. On the other hand, at the United Nations Millennium Summit in 2000, the Millennium Development Goals (MDGs) were approved by world leaders with the ultimate purpose of reducing poverty

[7] Appendix A lists and describes these goals.

and improving lives. It has three goals referred to education. This international focus it is an indication of the relevance and importance of the topic in question here.

2.2 Institutional Background

India is the second world's largest country by population with over 1.2 billion people. It is also the most populous democracy. Its presence in several studies is explained by its large population and effect of its economy in the world's totals. India has probably the largest school system in the world, with around 260 million children attending school. It is therefore of outmost importance that this system works stably and efficiently for a young population that needs skills and competences to contribute to the future growth of the country and, with such population numbers, certainly to the world.

The school system in India is similar to what we know form western countries. Primary education covers ages from 6 to 14 years old. These years consist of what is called Standard I to Standard VIII, that is usually referred to as elementary education. It is also often divided into lower (Std. I to Std. V) and upper primary (Std. VI to Std. VIII). Secondary education covers the age group of 14 to 17 years old (Std. IX to Std. XII) and it is also commonly divided into lower (Std. IX to Std. X) and upper secondary (Std. XI to Std. XII). Regarding preschool, it usually covers age profiles from 3 to 6 years of age, being some children aged 5 years old already attending primary school. Before preschool, some families send their children to childcare centers, usually in the form of informal home-care centers, while other families take care of children at home.

Although most of the Indian schools are administered by the government, there is an increasing trend of private schools enrollment. According to Pratham, one of the largest non-governmental organizations in the country, 67,7% of the schools were governmental schools in 2016, teaching until Std.V. This proportion increases to 71,8% in higher primary, Std.VI to Std.VIII. Between 2010 and 2016 there has been an increase in the percentage of children attending private schools of around 10%. This rise in private school enrollment is believed to be driven by the failure of public schools to provide quality education. In recent years, this rising trend seems to be stabilized, with no increase in private school enrollment between 2014 and 2016. With respect to pre-primary schools, these are made available through three channels: public (Integrated Child Development Services, ICDS), private and non-governmental organizations, although this last channel is almost non-existent in the sample used in this paper.

The Indian government has been taking important actions towards the development of its education system since decades ago. As a general view of these developments, I briefly go through their chronological history. In 1986, the National Policy of Education was adopted. For the rest of the 20th century, some other policies and programs where adopted and revised, together with several Five-Year national development plans highly focused on education. During this period, the world's largest nutritional program in school, the Mid-Day Meal Scheme, was launched to support and improve nutrition in schools. It was not until 2001 when the flagship program Sarva Shiksha Abhiyan was launched at a national level. This program was intended to universalize elementary education, specially in remote rural areas, and it was an extension of the already implemented District Primary Education Program (DPED) launched in 1994. In 2002, children's right to free and compulsory education between six and fourteen years was introduced in the Indian Constitution. In 2013, the Integrated Child Development Service (ICDS), launched in 1975, was revised and strengthened, as the principal public initiative for ECCE. It seeks to provide nutritional support, health and preschool education through *Anganwadi* centers to children under 6 years old. These are only the most important milestones and programs for the purpose of this thesis, but there were several other programs aimed at universal, remedial, vocational and adult education, as well as gender equality.

Since the implementation of these policies and programs, enrollment has been increasing, reaching almost universal education nowadays. According to Pratham and its Annual Status of Education Report (ASER), gross enrollment ratio (GER)[8] was 96.9% in 2016 for children between 6 and 14 years old. That is for elementary education. However, if one looks at later ages, 15% of children of 15-16 years old were out of school in 2016. With respect to pre-primary enrollment rates, in 2016, close to 40% of three-year-old and 25% of four-year-old children were out of preschool. These figures come from the ASER report of 2016 where, unfortunately, data on secondary education is limited and where only rural India is represented. The National Review of India of the Education for All Report in 2015[9] has richer data about enrollment rates, although not from 2016. According to that source, in 2013, gross enrollment ratios in secondary education (Std.IX-Std.X) and higher secondary education (Std.XI-Std.XII) were $76,6\%$ and $52,2\%$ respectively. It is worth emphasizing here that enrollment does not imply completion, and although gross enrollment rates are high, not all of those children complete elementary and secondary education. The most recent data available on dropouts from the EFA National Review of India is from 2008-2009. In these years, dropout rate in elementary education (Std.I-Std.VIII) was $42,3\%$, still a worrisome datum. Although lot of progress has been achieved, more needs to be done. There is a long way towards universal provision of early childhood care and education. Despite the fact that universal

[8]According to UNESCO, GER is the total enrollment within a country in a specific level of education regardless of age, expressed as percentage of the population in the official age group corresponding to this level of education. While Net Enrollment Ratio (NER) refers only to enrollment of children who actually belong to the group that officially corresponds to that particular school level.

[9]UNESCO (2015a)

enrollment in elementary education seems to be achieved, secondary education is still far from been universal. Furthermore, as it has been pointed out previously, it should be paid more attention to completion rates than to enrollment, since it is a fixed and determinant indicator of how many children actually achieve a certain education level. Despite of the progress achieved at reducing dropout rate, from 54% in 2000 to 42% in 2008, it still should be a key talking point for the coming education policies. Therefore, it seems that there is universal education for primary levels, but above them, the picture is rotten.

Another important talking point not mentioned so far is the quality of education. School enrollment or attainment does not bring positive outcomes by itself. It is no point in increasing enrollment, in keeping children at school, if they are not learning anything there. If that is the case, then they might be more useful helping their family on the farm or doing house chores. In recent years, there has been more talk on the matter of quality rather than quantity. A quite popular paper by Hanushek and Woessmann (2015) demonstrated that what it is really important for future income and growth is the quality of the education received rather than years of education. More in particular, the important factor is what children learn at school. Muralidharan (2013) warned that besides all the progress and improvement in enrollment ratios, learning outcomes of Indian children are still mediocre. According to Pratham's ASER report, as for 2016, 20% of the children enrolled in Std.III cannot read a Std.I level simple text. In arithmetic, 35.6% of the children enrolled in Std.III cannot do subtractions. However, in order to make meaningful comparisons, it is necessary to put this numbers in an international perspective. Muralidharan (2013) highlights the striking results from some international tests administered to Indian children. Particularly, when Indian children are asked questions extracted from the TIMSS (Trends in International Mathematics and Science Studies) test and PISA (Program for International Student Assessment), it is common to find them at the lower tail end of the score distribution. For example, the ASER report from 2012 noted that the mean score on TIMSS questions from Std.IV language test is less than half of the international mean in Indian public schools. This is indicative that, although children pass grades, they do not retain or have the knowledge that corresponds to that level, to the contents taught there. This has a straightforward implication to their likelihood of completing subsequent grades. If children feel they are not learning anything in school, that they are falling behind and have no opportunities to catch up to the level of the class, they might feel that they could be better off somewhere else than at school.

This summarized the school system and its performance in India and where its weak spots are. In the next section I will review the academic literature about preschool and policies aimed at improving schooling outcomes in order to give us an idea of what has been found in other studies and countries.

3 Literature Review

This section will go through the recent literature about early childhood care and education, with particular emphasis on preschool. The main objective is to show the results from different policies and programs as well as non-experimental studies implemented in developed and developing countries, their validity and the size of their effects. So that the reader has a broad view of what other studies have found in similar environments as the one presented here.

I will now review three of the most important and interesting programs that have taken place during past decades in the US. Two of them are model demonstration programs that use small samples of around 100-200 participants, while Head Start is a large scale public program[10].

Head Start is one of the largest ECCE Program ever implemented in the US. It was established in 1965 and offers educational, nutritional, health and social services to children of families with low-incomes. The goal of the program is to improve family relationships, children well-being and readiness to school. The Office of Head Start (OHS) provides funding and oversights the respective agencies that offers Head Start services, while the Head Start delegate agencies follow guidelines and implement programs as required by the Office. It has an estimated cost per child of around \$8.000 and serves almost 1.000.000 children and infants, according to the Federal Register of the US for 2016. Considering that the transition to primary school is a challenge for children, requiring children to adapt to new peers, a new classroom, an out-of-home environment and school rules, it is worth investing on it.

There is a vast literature on the effects of the Head Start program from different states

[10]See Currie (2001) and Elango *et al.* (2015) for reviews of those programs and their benefits.

within the US. Barnett (1995) reviewed 36 articles about early childhood programs, many of them part of Head Start. They focus on achievement tests, grade retention and high school graduation. Previous research on Head Start state that its effects on achievement test scores are small and seem to fade out over time (Copple *et al.*, 1987). Therefore, the main issue here is that preschool intervention cannot provide long-lasting benefits by itself, but instead it needs to be paired with later school programs or follow-ups in primary, so the gains from preschool do not disappear during school years. In order to answer the question of long-term effects of the Head Start programs, Garces *et al.* (2000) use a sample from PSID and families with siblings where one of them is not part of Head Start. Controlling for observable differences and family based fixed effects, they found that Head Start has a positive effect on the probability of college attendance in adulthood. They conclude that Head Start has social and economic benefits that persist into adulthood. Therefore, Head Start should not be considered a failed program due to the short-term IQ fade out, since there are other important determinants of late-life outcomes like college attendance or crime participation.

Another important intervention on the effects of preschool is the *Perry Preschool Project*, which was a high-quality preschool scientific experiment for children from disadvantaged backgrounds carried out from 1962 to 1967. The intervention consisted of half-day preschool every week day plus 90 minutes home-visit each week, for a period of two years. It was targeted at three- and four-year-old African-American children. They followed a random assignment strategy so that the differences across groups are explained by preschool attendance. Cost of the program was $11,300 per child per school-year (in 2007 dollars). Although sample size was small, attrition is low even after 35 years of its implementation, with the last follow up to adults age 50. The study has found that adults who attended the preschool program had higher earnings ($12.000

vs \$10.000 at the age 40), less criminality (36% vs 55% arrested five or more times) and were more likely to graduate from high school (65% vs 45%).

The last important intervention reviewed here is the *Carolina Abecedarian Project*. Children born between 1972 and 1977 were randomly assigned to an early educational intervention group or to the control group. Treatment consisted of high-quality educational intervention in childcare through the age of 5. There have been several follow-ups at ages 12, 15, 21, 30 and 35. Estimated average cost per child was \$13.900 (Masse and Barnett, 2002). At 21, children who receive preschool intervention had higher test scores and were twice as likely to be in school that non-participants. On the 30 years-old follow up, the treatment group had more probabilities of employment, more likely to have graduated from college and completed 1.2 more years of education.

These three examples of ECCE interventions in childhood development show that there are positive and long-lasting effects of preschool on adulthood. It is important to note that from these studies and others, the long term effect of preschool on IQ scores is insignificant. Even though early intervention can affect early IQ scores, this effect fades out with time, but there are other factors affecting later outcomes positively. These factors are social and non-cognitive skills, readiness to school, crime participation, health, and other late-life outcomes that do not need to be related to IQ, but are important for individuals and society.

Only developed country research has been reviewed so far. There are a lot of factors influencing the results shown above, and although families and children taking part in the interventions are disadvantaged ones; institutions, environment, legislation, and everything surrounding them in a developed country is significantly different from the ones in a developing country. We know now that it is possible to have positive effects of preschool in life outcomes of disadvantaged children in developed countries. Even though there are several ways to improve education attainment and learning outcomes, in this setting one has to strive to find the most cost-efficient policies so that developing countries can effectively implement them. Hence, I turn into some of the literature on the developing world, to see if the picture looks similar.

As opposed to the developed world, literature about ECCE and preschool in the developing world is scarce. And unlike primary school provision, early childhood actions have been poorly executed or nonexistent, being only South American and Caribbean countries able to truly implement them (UNESCO, 2006). But before reviewing the literature on preschool in developing countries, I will shortly go through other policies and interventions aimed at improving school outcomes, as preschool does, so that one is aware of the different alternatives in government hands in order to improve school enrollment and completion.

Kremer *et al.* (2013) collects several randomized controlled trial studies in developing countries[11] to show how different policies affect enrollment and test scores in different contexts. Summarizing them: *Class size* seems not to be important when school accountability is weak, as results from Duflo *et al.* (2015b) showed in a Kenyan program. However, other studies (Magnuson *et al.*, 2007) found positive effects. Reducing class sizes is a costly measure since it requires hiring more teachers and constructing new classrooms. There certainly are more cost-efficient options. Other inputs like *books* do not seem to have an effect either (Kremer *et al.* (2013) and Glewwe *et al.* (2009)). However, provision of textbooks does have an effect on the top students of the class, implying that students already faring well in school can benefit from these textbooks, while the average student find them useless. Despite being official government textbooks, they

[11]For a full review of ECCE randomized experiments and studies, see also Ch.10 Economics of Education.

are probably too difficult for the average student. This is another argument in favor of the ”teaching at the wrong level” theory. Related to these results, Banerjee *et al.* (2007) found test-scores gains from an extra young woman hired to teach students basic literacy and numeracy concepts in India. The program increased the learning outcomes of those students, demonstrating that teaching at the right level is an effective solution to improve learning success. The same study also found large effects in math scores the first year from computer-assisted learning program (CAL). Going back to educational inputs, provision of *uniforms* found reductions in girls dropout rates of $3,1\%$ in Kenya (Duflo *et al.*, 2015a). The reason lies in the fact that children in Kenya and other developing countries face such strong social pressure to wear uniforms that some will not attend school if they cannot afford one. Hence, these subsidies reduce the cost of education. Other remedy found in the literature is tracking classes, where students are tracked ”down” and ”up” according to their level, grouping students that have similar levels. This practice has found increases in test scores (Kremer *et al.*, 2013). Technology-based instruction, with initiatives like computer-based classes or individual laptops for students have ambiguous results in different settings. Probably because the use of this technology was not connected to the curriculum. *More schools* is associated with more time at school since it reduces costs of transport. Burde and Linden (2012) found that a program that created schools within villages in Afghanistan increased children's school participation and learning. This is one of the reasons of the rise in enrollment in India in the past decades[12]. Another effective solution in India has been the inclusion of *school meals* in public schools, where Afridi (2011) showed a 12% increase in girls' school participation. Another big problem in India is *teacher absenteeism*, which of course re-

[12]According to UNESCO (2015a), from 2000 to 2014, there has been an increase in the number of schools from 845.000 to 1.448.000.

lates to student performance, but focusing on policies and solutions for this problems is beyond the scope of this paper. Other policies like *conditional cash transfers, vouchers, lotteries and merit scholarships* do have an effect on schooling because they reduce the face cost of school to families that are under financial pressures. Haenfler *et al.* (2002) show gains from a voucher program in Colombia. Glewwe and Kassouf (2012) studied the impact of Bolsa Escola/Familia, a cash transfer program in Brazil, finding significant enrollment increase and dropout reduction. Lastly, Glewwe and Muralidharan (2015) review different papers on school governance and how transparent management and monitoring might have positive effects on learning outcomes.

All these programs and policies seem to have effects on short term outcomes, or on enrollment by facilitating access to school. If we look at recent trends of enrollment, policies intended at keeping children in school are successful. However, learning outcomes are the factors that make children lagging behind and not continue school. Teaching at the right level might be a good solution for that[13]. School curricula was set in times where only elite populations attended to school. As a result, only the best students were enrolled in school. Times have changed, and due to all the efforts to bring education to all children, enrollment rates are higher than ever, making education almost universal. Therefore, some have argued that curricula are obsolete, result of a different education approach, and needs to be adequate to current times. Perhaps, what it is needed in the education systems of developing countries it is not more remedies or fixes, but better education structures, coherent curricula, and smoother paths through school levels, that adapt to the children population that characterizes those countries. Introducing new levels of compulsory quality education, like preschool, could be a first

[13]Chapter 4 of Poor Economics (Banerjee and Duflo, 2012) go through this and other problems of schooling in developing countries, sometimes specifically in India.

step aimed at adjusting these education systems that seem to lack what is necessary for the future of their children.

Unfortunately, and opposed to the developed world, experimental studies, programs or policies aimed at improving preschools and preschool attendance are not that present in developing countries. Furthermore, preschool programs, policies and experimental studies are more difficult to implement. Programs like High/Scope Perry Preschool study are expensive, and require long follow-ups, usually implausible. In the following lines, studies about preschool in developing countries will be reviewed.

Berlinski *et al.* (2009) study the effect of pre-primary education on Uruguayan children's later school outcomes. Exploiting data from the Uruguayan household survey on siblings and using instrumental variables, they found small improvements from preschool attendance at early stage, that increased with time. At the age of 15, preschool attenders had 0.8 more years of education and were 27% more likely to be in school that their untreated siblings.

In Bangladesh, Aboud *et al.* (2008) compare children who attended preschools that were part of the "Succeed" program with a group of children who did not. The Succeed program aimed at improve learning capacities of children. The control group was selected before Succeed program was established. One year later, students graduated from Succeed preschool were recruited, so control and treatment group individuals belong to different age cohorts. With two different preschool models, a home-based and school-based one, the results from this evaluation show that children that attended "Succeed" preschool performed better than their counterparts in four out of five competences where they were tested. Because competence tests were undertaken in first grade, only one year after preschool graduation, this a extremely short-term indicator. Additionally, the empirical strategy followed does not meet enough quality standards, since it does not measure possible differences in the time periods of control and treatment group.

In East Africa, Mwaura *et al.* (2008) investigated the effect of preschool attendance on the cognitive development of children. They tested students before enrolling in preschool and 18 months later and showed that attenders performed better in different tests (verbal comprehension, basic number, pictures) than the control group, formed by children that stayed at home instead of attending preschool. As the above-mentioned study, this study is short-term aimed, so that these preschool gains might fade out later in children's lives. The relevance of this study is hindered by the fact that their main empirical strategy was a hierarchical regression on test scores, controlling for possible confounders including types of preschool. Later they compared coefficients on the type of preschool provision to assess to what extent it predicted total cognitive test scores. Therefore, it presumably fails to control for self-selection into treatment.

Taiwo and Tyolo (2002) studied the effects of attending preschool on academic performance of grade one pupils in Botswana. Academic outcomes tested were English language, maths and science. They found that children with preschool experience outperformed their counterpart in all the areas. However, this study does not meet quality standards as to be considered relevant enough, being t-test comparisons between control and treatment groups its main empirical strategy. Assignment was non-random, therefore this study is subjected to be biased.

Veramendi and Urzúa (2011) studied the effect of out-of-home care on children aged 2-5 years old in Chile. Using a specific questionnaire designed to study early childhood development, they concluded that childcare centers seem to boost children's test scores. They control for selection into treatment through multidimensional characteristics of parents and availability of child care providers and through the use of instrumental variables.

For more information about the literature on early childhood development in developing countries, Engle *et al.* (2011) and Nores and Barnett (2010) both offer rich and extensive literature reviews on intervention studies that have taken place in developing countries, with the purpose of evaluating early childhood development in low- and middle-income countries.

As the literature on preschool has shown, it is common to find significant gains from attending preschool, although this literature is scarce and mainly focused on IQ and test scores, which have been demonstrated to produce only temporary gains. Additionally, most of the studies that explore the causal effect of preschool in developing countries face problematic empirical estimation strategies, and none of them uses experimental designs. As for the Indian case, the effects of preschool on dropouts and completion rates have not been thoroughly discussed in the literature.

4 Data

This section aims at giving the reader a broad understanding about where the data that is being used in here stems from. Part of that will be about its unique characteristics, as well as its limits[14].

4.1 The Young Lives Project

The data used in this paper comes from a collaborative research project coordinated out of the Department of International Development at the University of Oxford. The project, called *Young Lives*, aims to study the drivers and impacts of child poverty in Ethiopia, Peru, India (state of Andhra Pradesh) and Vietnam. This study is following two groups of children through their youth, with a total of 12.000 children. The older cohort children are born in 1994-1995 and the younger cohort children are born in 2001-2002. The study was set to last for 15 years, so that by the end of the study (around 2017), the older and younger cohort children would be around 22 and 15 years old, respectively.

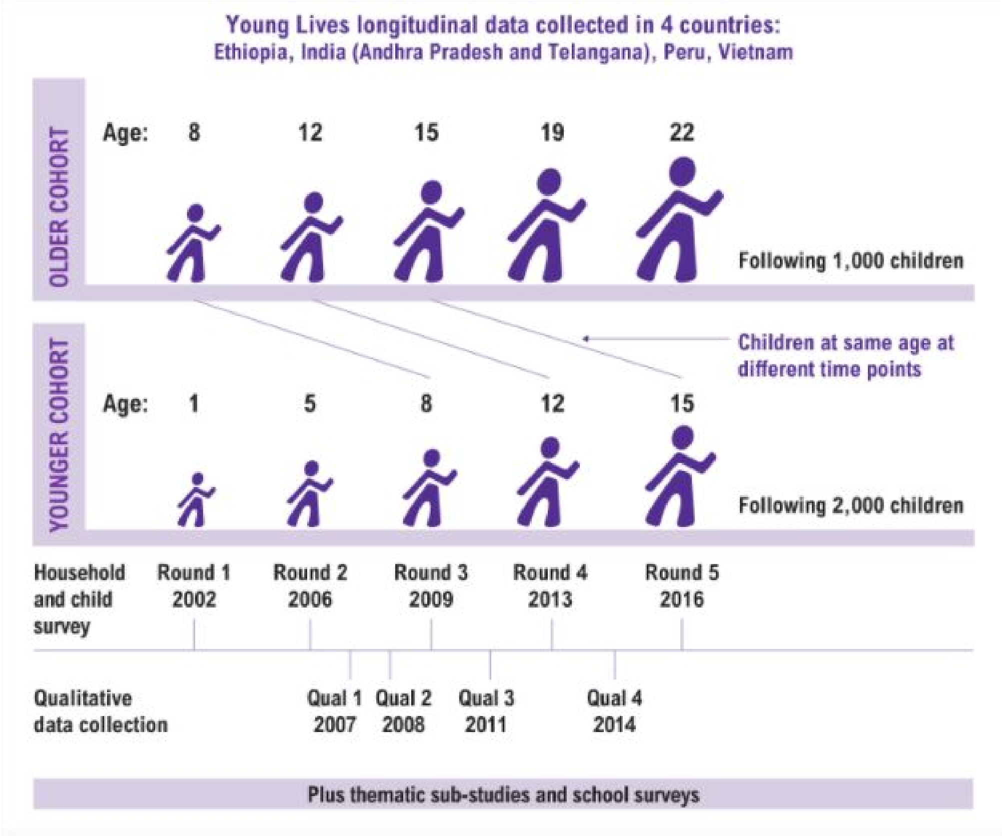
The survey consists of a set of questionnaires administered to children, parents/caregivers and key informants of their communities every three or four years. Small sub-sample qualitative in-depth research and school surveys were also conducted through the period of the study[15]. Figure 1 shows the timeline of the interviews graphically, as well as age and number of children per cohort. There have been five rounds of interviews so far, from 2002 until 2016. Unfortunately, only data from Rounds 1 to 4 is available and

[14]Data available at the UK Data Service. Series: Young Lives: an International Study of Childhood Poverty [https://discover.ukdataservice.ac.uk/series/?sn=2000060] GN 33379.

[15]These data provide additional important and useful information but I have decided to leave them out of the present study due to time limitations and small sample sizes.

data from Round 5 is expected to be publicly available in June 2018. Samples consist of 1.000 children for the older cohort and 2.000 children for the younger cohort.

Figure 1: Data collection timeline



Source: www.younglives.org.uk

There are different questionnaires for each type of recipient within each survey round. One questionnaire was directed at child's caregivers (the *Household Survey*), where topics like parental background and education, livelihood activities, assets, time use, food and non-food consumption and expenditure, recent economic changes, social capital, household members' health, and access to basic services were asked. It also includes questions regarding child's perceptions and attitudes towards different subjects, dynamics and aspirations. A second questionnaire was aimed at children above 8 years old, age when they were able to answer its questions. This *Child/Individual Survey* provides information on an individual level about children's perceptions of well-being, daily activities, attitudes towards school and work, how they feel they are treated by others, and their future aspirations. In later rounds, the child survey also asked children about their time use, mobility, and complete school histories. If the child was not 8 yet, similar questions were asked to caregivers and included in the *Household Survey*. Lastly, another questionnaire was administered to community informants and workers (*Community Questionnaire*), covering the social, economic and environmental context of each community. This questionnaire included questions about population, ethnicity, religion and language, economic activity and employment, infrastructure and education, health and child protection services. Table 1 gives an example of the contents of the different questionnaires in Round 1.

4.2 Sampling method and site information

This section describes the Young Lives' sampling methods and choice of sentinel sites in the state of Andhra Pradesh, in India[16]. Sampling has been carried out using a semi-purposive sampling strategy, where a site is expected to represent a certain type of population and to present typical trends affecting its people and locality. The state of Andhra Pradesh has three distinct agro-climatic regions: Coastal Andhra, Rayalaseema and Telangana, the latter eventually became a state in itself in 2014. The first step was to choose the districts from these regions, from which sentinel sites would be located. Those districts were chosen according to three development indicators (economic, hu-

[16]For a full detailed description of the sampling method in India: http://www.younglives.org.uk/node/7183

Table 1: Contents of child, household and community questionnaires in Round 1

	Household questionnaire	**Child questionnaire**	**Community questionnaire**
Both cohorts	Household composition Caregiver background Child health Household livelihoods Economic changes and events Socio-economic status Social status Child height and weight		Physical environment Social environment Infrastructure and amenities Economy Health and education Prices
Younger Cohort	Pregnancy, delivery and breastfeeding Child care Caregiver mental health	Children too young to answer direct questions	
Older Cohort	Child mental health Child education and daily activities	Perceptions of well-being Social capital School and work Health Literacy, numeracy and child development	

Source: Young Lives

man development and infrastructure). These districts are presented and described in Appendix B. One objective was that the distribution of districts across regions fully represented each region, and that poor and non-poor groups were selected. The districts selected covered around 28% of Andhra Pradesh population, and included around 318 of 1.119 mandals[17]. In a second step, mandals were selected to form the sentinel sites. These mandals were selected according to development indicators as well. However, the capitals of the urban districts were hand-picked as specific mandals due to a lack of enough urban sites. Additionally, another site was selected from the urban slums of Hyderabad. Hyderabad district was not chosen according to the same development indicators as other districts because is not part of the three regions, but it is an urban and

[17] A mandal is a sub-division of a district in some regions of India like Andhra Pradesh and Telangana. Its equivalent in other states are named differently.

metropolitan area of Andhra Pradesh. In total, there are 20 distinct mandals or sentinel sites. In the last step of the sampling process, some villages or communities were selected from the mandals. Mandals were divided in 4 contiguous areas and one village was randomly selected from each of the areas. They ensured that the 4 villages selected from each mandal had enough population to include 100 1-year old and 50 8-year old children. If not having enough children within four villages, additional villages were included. A detailed map of India, its regions and mandals can be found in Appendix C.

Before finishing this subsection, an observation to the quality of this dataset and its representative power of Indian population: Data from Young Lives (YL) has not been collected with the purpose of representing the actual Indian schooling situation, nor to understand what its problems are. The data has instead been collected with the purpose of exhibiting changes over time in a population that, due to the sampling method, might not be representative of the whole country. On a report by Young Lives about the similarities of YL data and Demographic and Health Survey (DHS)[18] data, they showed that on average, YL households are a bit wealthier, although DHS data was retrieved at earlier years (1998/1999). They conclude that, instead, Young Lives data is a good instrument to analyse causal relationships and examine child welfare. The fact that they state that YL households turn out to be a bit wealthier that the DHS data contradicts one of the widely-known features of the Young Lives data: its focus and inclination for poor and disadvantaged households. According to Young Lives, poor households are over-sampled, but non-poor families are also represented in the sample. That is the reason for choosing one poor and one non-poor district in each region. Therefore conclusions drawn based on Young Lives data should be taken with caution, bearing in mind that no

[18]The DHS Program collects nationally representative data on fertility, family planning, maternal and child health, gender, HIV/AIDS, malaria and nutrition in over 90 developing countries.

efforts have been made to make the sample representative of the whole country.

In randomized controlled trials, attrition and tracking are some of the most important problems that researchers often face. Longitudinal studies planned to follow individuals over a long period also face these problems. It is even more critical when the target group consists of children in a developing country. In a country like India, a 16 years old child is considered adult, able to marry, have descendants, and live in his own household. Therefore, fieldworkers have to go one step further on their tracking in order to locate the studied individuals over time, sometimes even to different communities. Despite all these difficulties, Young Lives researchers and fieldworkers did a very good job. As for Round 4 (2013), attrition rates were 2.6% for the younger cohort and 4.3% for the older cohort, mainly due to moving and marriage. Table 2 shows disaggregate figures about attrition.

Table 2: Attrition between Round 1 and Round 4

	Younger Cohort		**Older Cohort**	
Initial sample in Round 1 (2002)	2011		1008	
Died	45	2.2%	13	1.3%
Refused	14	0.7%	20	2.0%
Untraceable	36	1.8%	19	1.9%
Living abroad	1	0.0%	4	0.4%
Interviewed in Round 4 (2013)	1915	95.2%	952	94.4%
Attrition		2.6%		4.3%

Note: Deaths not included within attrition.
Source: Young lives

4.3 Data Selection

Previous sections have presented the data structure to the reader, together with sampling design and sites choice. In this section, however, I introduce the part of the data has been used in the present paper, as well as a discussion about it, its advantages and its disadvantages.

It is worth reminding here the purpose of the present thesis. I am interested in the dropout nature of Indian children throughout their school lives. The variable that gathers this information and its possible variations is dichotomous and fixed over time. Either a child finishes school or she drops out and usually she does not rejoin school afterwards[19]. Most of the rest of variables used in the estimation are also fixed or dichotomous, since they describe states or conditions. Therefore, I do not need to make use of the panel advantages of the data, as it would be enough with identifying the characteristics that preschoolers display and move forward in time to see whether they have completed school or have dropped out. For these reasons, I treat the data as cross-sectional.

As stated in the background section, one of the main problems in India was the low rate of secondary completion (or even enrollment). Thanks to the Millenium Development Goals and Education for All initiatives, primary school enrollment rates were high in India in the years of the study. The fact that the problem seems to emerge after primary education has brought up a difficult trade-off in the data choice for this thesis. The younger cohort has data from 2002, when children were around 1 year old. This data is more complete and, since the birth of the child was recent, includes richer questions, like breastfeeding or children's birth weight. Additionally, preschool attendance is the main topic of some subsections of the questionnaires. These characteristics, and a larger sample size, make this cohort appealing and suitable for this study. Unfortunately, in 2016, latest data available is from 2014, the fourth round of the survey, where the younger cohort children are right in the transition from primary to secondary educa-

[19]I have controlled for children that have rejoined school, taking them out of the dropout group.

tion. Since most of them have not even completed primary schools, dropouts are not yet identified. It is then unfeasible to use this cohort for the purposes of this study[20]. I am then left with a sample of 1.000 children ranging from 8 to 19 years along the period in order to carry out the study. Unlike the younger cohort data, older cohort data lack the degree of detail and singularity characterizing the younger cohort. Data on preschool attendance do not appear in an straightforward manner in the questionnaires, and information on prenatal care, delivery and early life measurements are nonexistent. On the contrary, children school histories are complete, including secondary school and even high education enrollment. Although it is not as complete as the young cohort, it is the best available data to answer the questions posted in this study.

Using the old cohort data, preschool is identified with a question in Round 3 where children are asked for their school histories back to 1994, date when they were born. The questions of interest read: *"Did you attend any school, preschool, early language program or kindergarten or similar for more than 6 months in year XXXX?"*, *What grade were you in XXXX?* and *What type of school was it in XXXX?*. Children could choose to tick a box that reads: *00 - Preschool* as an answer for the second question, so then I have retrospective information about preschool attendance. It is important to note that this specification of the questionnaire makes the whole study open to critique that one of the key variables is based on children recalling the school history. Young Lives questionnaire designers and researchers explained that they made sure that another relative, usually mother or caregiver, was present during the child's interview in order to make recalling issues less problematic.

[20]Carrying out a study of the same characteristics as the present one in the younger cohort is highly recommended when the last round of the study comes out in 2018

4.4 Descriptives

As the last part of this data section, I present a range of descriptives to give an idea of how the sample is distributed across preschoolers and non-preschoolers.

Table 3: Sample Descriptives

	Preschool			% Preschool	
	No	**Yes**	**Total**	**% No**	**% Yes**
Child's sex					
Male	222	250	472	47.0%	53.0%
Female	178	301	479	37.2%	62.8%
Region of residence					
Coastal Andhra	164	164	328	50.0%	50.0%
Rayalaseema	113	155	268	42.2%	57.8%
Telangana	117	224	341	34.3%	65.7%
Others	6	8	14	42.9%	57.1%
Father's Education					
No primary	170	157	327	52.0%	48.0%
Primary	82	111	193	42.5%	57.5%
Secondary	133	226	359	37.0%	63.0%
Higher Education	15	57	72	20.8%	79.2%
Wealth index					
Below than p25	97	139	236	41.1%	58.9%
Between p25-p50	123	126	249	49.4%	50.6%
Between p50-p90	164	204	368	44.6%	55.4%
Above p90	16	82	98	16.3%	83.7%
Total	400	551	951[21]	42.1%	57.9%

Source: Own elaboration using Young Lives data.

As table 3 shows, preschool attendance is evenly distributed in the sample, where

[21] Although the initial sample size was 1.000 children, only 951 remain after Round 4 and after cleaning the data. See Appendix D for an explanation of dataset creation, cleaning and missing data treatment.

58% of the children attended preschool at any point between 3 and 6 years old for at least 6 months, while 42% did not. More girls attend preschool than boys, with a 62% rate of attendance. There are also differences with respect to the regions of the sample, where Telangana seems to have higher preschool attendance rate than Coastal Andhra and Rayalaseema. Predictably, father's education conditions preschool attendance. The higher the child's father's education, the higher the rate of preschool attendance of that child. Lastly, another group of descriptives are presented according to wealth levels, measured by a wealth index[22]. Surprisingly, the proportion of children attending to preschool is larger in the group whose wealth index is below the 25^{th} percentile than in the groups between 25^{th} and 90^{th} percentiles. More than 80% of the children whose family's wealth index is above the 90^{th} percentile attended to preschool.

[22] The wealth index is calculated by the Young Lives as the average of three components: housing quality (HQ), consumer durable (CD), services (SV), all with values ranging between 0 and 1. The HQ Index is based on the number of rooms per person in the household and the main materials used for the walls, roof and floor. The CD Index is based on the number of assets owned by the household. A typical set of assets is considered like radio, refrigerator, bicycle, television, motorbike/scooter, car, mobile phone, landline telephone, fan, almairah (wardrobe), and clock. The SV Index is based on whether or not the dwelling has electricity, the source of drinking water, type of toilet facility and the main type of fuel used for cooking.

5 Methodology

This section will cover the empirical strategy followed in the present thesis in order to estimate the effects of preschool attendance on children's school completion.

5.1 Academic Outcomes

This subsection aims at explaining and defining the outcome variables used in this thesis. There are several ways to reflect academic performance and learning outcomes. The data from Young Lives offers a wide range of possible variables to reflect academic performance. As explained in Section 2, one of the problems that India is currently facing is a high dropout rate after completing primary school. Although current primary education enrollment levels are quite high and there has been large improvements on the basis of the *Education for all* initiative, attendance to secondary school is still below desired levels (UNESCO, 2015a). It has also been noted that dropout rates in primary education are not acceptable. Therefore, it seems reasonable that one of the main objectives of this study is to shed light on the effect that preschool might have on stay-on rates in primary and secondary education. Throughout this thesis, a child is considered as having dropped out when she does not complete Standard XII by the last round of interviews (Round 4, 2014) and she has been declared as ”not in school” in that questionnaire. This includes children not completing primary as well as those with primary completion but not secondary. Aware of the large extent of this definition, I will also show the effects on more specific outcomes like completion of primary or secondary school. Completion of primary has to be understood as finishing Std. VI and completion of secondary as finishing Std. XII.

Besides outcomes related with the extensive margin, like how many years of schooling, it would also be quite interesting to explore the effects on learning outcomes, in the sense of test scores or grades. This way, quality of education could also be studied. It is also well-known that the fact that children might not be prepared enough to fare properly at school might be one of the determinants of dropout rates[23]. Children are not able to follow the level in the class, falling behind and feeling that they do not belong to that level. Since in developing countries is common that every child moves on to the next course, this might eventually end up in dropouts as difficulty keeps scaling up. Hence, if attending preschool improve readiness to school, children might have more possibilities and resources to fare better at school without falling behind and feeling demoralized, and therefore, more likely to keep progressing in their education. To answer this question, however, I would need to use a different approach as the one used here. I would need data on the students progression through different grades, retention rates, promotion rates, test scores and a difference-in-differences approach to account for the gains earned through education between years. Therefore a deep analysis of the matter is beyond the scope of this thesis.

5.2 Preschool Attendance

The objective of the thesis is to discover differences on academic outcomes among children that attended preschool with those who did not. Hence, preschool has to be defined and delimited. In order to assign children to treatment and control groups, I use retrospective information about children school history from the survey questionnaire in round 3 (2009), explained in Section 4.

[23] Muralidharan (2013) mention the fact that curricula are designed to teach to the top of the class. See also Banerjee and Duflo (2012) and Pritchett and Beatty (2012) for more elaborate discussions about this topic.

I defined treatment as the preschool attendance for more than 6 months and compare it to a control group made up of children that have not attended preschool for such a period at least. The questionnaire explicitly indicates that attendance (school or preschool) must be declared only if it has lasted for at least 6 months, so that students that continuously join, quit and rejoin school are left out of the treatment group. This is a quite important characteristic of the questionnaire that needs to be considered for from now on. Hence, treatment group consists of children that attended preschool more than 6 months at any time in a year between 3 and 6 years old, and control group consists of children who did not meet that minimum preschool attendance to be categorized as ”preschoolers”. This means that there are children in the control group that might have been attending preschool, but not long enough to be included in the treatment group. The structure of the questionnaires and data available limit the identification of more groups here, being not possible to discern who of those children did not go to preschool at all or who went to preschool but less than six months. Therefore, through this thesis, children who attended preschool more than six months per year are compared to a control group formed of any other children that do not meet that criteria.

Along with that general concept of preschool, heterogeneous effects of preschool are also explored, like whether there exist differences between private and public preschool, and whether preschool duration matters. This is done simply by modifying the treatment variable, so that I compare e.g. two years of preschool vs. no preschool at all.

5.3 Identification Strategy

I am interested in exploring what are the effects on academic outcomes and dropout rates of attending preschool. In order to study this causal effect, I need a counterfactual to compare those individuals with - i.e. outcomes of the same students if they had not attended preschool. Ideally, one would like to observe the outcomes of the same individual when she has attended preschool and also when she has not, so that the only difference is preschool exposure. However, it is impossible to observe both outcomes in the same individual at the same time, since either she attends preschool or she does not. Formally, let's call $Y(0)$ the outcome of individual i if she has not attended preschool ($D = 0$) and $Y(1)$ if she has attended preschool ($D = 1$). In this case, we have a binary treatment D -preschool attendance- being equal 1 if individual i receives treatment and 0 otherwise. The focus here is in the "Average Treatment Effect on the Treated" (ATT), defined as:

$$\tau_{ATT} = E[Y(1)|D_i = 1] - E[Y(0)|D_i = 1] \tag{1}$$

But only one of the values is actually observed, since the conditional mean of the counterfactual for those treated -$E[Y(0)|D_i = 1]$- is not observed. Comparing conditional mean outcomes of individuals who have attended with those of individuals who have not is not advisable, since there might be other underlying factors that lead to selection of one group with certain characteristics into treatment. This is the so-called selection bias problem. Variables that affect selection into treatment might also affect outcome. For example, children whose parents are educated might have more chances to attend preschool because their parents are aware of the benefits of education, but also, and independent of attending preschool, it might be more likely that those children achieve further education levels because their parents can help them with school homework. If this is the case, even in the absence of treatment, outcomes of individuals from control and treatment group would differ - if the parents of a child are educated, as shown in the literature, will have higher incomes, and therefore more economic resources to keep the child in school rather than sending him to work to support family income. Equation 2 shows the selection bias problem when comparing conditional mean outcomes.

$$E[Y(1)|D_i = 1] - E[Y(0)|D_i = 0] = \tau_{ATT} + \{[E[Y(0)|D_i = 1] - E[Y(0)D_i = 0]\} \quad (2)$$

Where the last term on equation 2 is the "selection bias". In this case, the τ_{ATT} parameter is only identified through comparison of conditional mean outcomes of preschoolers vs. non-preschoolers if there is not selection bias ($[E[Y(0)|D_i = 1] - E[Y(0)|D_i = 0] = 0$). This condition is only met when treatment assignment is random. Randomized controlled trials have been quite popular since the past decades for evaluation of programs and policies, where individuals are randomly assigned to control and treatment groups so that there is no selection bias. Therefore, in the recent empirical research, randomized controlled trials have become the gold standard. However, this research requires a lot of resources and sometimes it is subjected to moral and ethical rules. The data at hand does not stem from a randomized trial experiment and as a result one has to rely on other methods, more specifically a non-experimental design. Another empirical strategy similar to randomized experiments are natural experiments. These strategies can be used when researchers have not access to a fully experimental design where control and treatment groups are randomly assigned, but there is in fact an exogenous variation in the data that introduced an unexpected change so that we can separate individuals into treatment and control groups in a close-to-randomized manner. That exogenous variation is usually introduced by new policies, regulations, natural events, or simply features of the particular dataset used. Unfortunately, the nature of the data does not allow for any of these strategies. Two of the most used methods used in empirical works have already been ruled out, particularly the ones whose estimates are less biased and free from common problems like unobserved variables or selection bias. The next set of empirical methods that aims to establish causal inference are quasi-experimental research designs. In those research designs, one applies different econometric methods to observational data in order to come a step closer to a causal relation. However, it is a complex and difficult task, and it is rarely the case when causality is assured due to the set of assumption taken in order to perform the analyses.

In this thesis, propensity score matching is used in order to compare individuals as similar as possible in terms of treatment probability from control and treatment groups. Next section covers the strategy in depth, its reliability, assumptions for a valid inference and how Young Lives data match this purpose.

5.4 Propensity Score Matching

As it has been mentioned at the beginning of this methodological section, propensity score matching aims to give estimates a causal interpretation, and reduce bias in the estimation of treatment effects using observational data.

Propensity score matching intends to match treatment and control observations based on their propensity scores, which are determined by covariates. So one compares mean outcomes of individuals that share similar characteristics, and therefore tries to overcome the selection problem explained above[24]. Rosenbaum and Rubin (1983) suggested in their popular paper to use balancing scores $b(X)$ instead of covariates, where $b(X_i) = E[D_i|X_i]$ are functions of the observed covariates X. An example of balancing scores is the propensity score, which indicates the probability of participating in a program given a set of covariates X. This way substitutes the previous way of matching

[24]This only holds if there is no omitted variable bias

on covariates, X, and saves the researcher from matching 2^X possible combinations, whose complexity increases with the number of covariates included in the model. This is known in the literature as the ”Curse of dimensionality”.

It is possible to estimate τ_{ATT} in non-experimental designs using propensity score matching by invoking two identifying assumptions[25]:

a) Conditional Independence Assumption (”Unconfoundedness”): this assumption implies that selection is based solely on observable characteristics and that one observes all variables that influence the outcome and treatment assignment simultaneously. Covariates, X, have to be independent of treatment, as it will be further discussed below. Formally,

$$Y(0),Y(1) \perp\!\!\!\perp D|P(X), \quad \forall \quad X \tag{3}$$

where $P(X)$ is the balancing score or propensity score suggested by Rosenbaum and Rubin (1983). They show that if, conditional on covariates, potential outcomes are independent of treatment, then they are also independent of treatment conditional on their propensity score $P(X)$.

Aware of the importance and restrictiveness of this assumption, its validity must be justified by the data. The dataset used in this thesis is extremely rich, and gives access to a large number of individual characteristics, parental background and community characteristics. All the available confounders are included in the model. For the rest of the thesis, it is assumed that CIA holds. However, Section 8 further discusses the validity of the assumption and its possible violations.

b) Common Support (”Overlap”): the second assumption ensures that identical sub-

[25]Another intrinsic assumption not mentioned here is Stable Unit-Treatment Value Assumption (SUTVA), requiring that treatment effect for each individual is not affected by other individuals' participation.

jects have at least some probability of participation in both treatment and control group. It implies that, given a set of X, the probability of participation is not 1 or 0 for any of the individuals. So we will always find individuals with similar characteristics in both groups. In other words, given a set of covariates, there is always a chance to observe units in both control and treatment groups.

$$0 < P(D = 1|X) < 1 \tag{4}$$

These assumptions are referred to Average Treatment Effects (ATE). In the present case, where the only parameter of interest is the Average Treatment Effect on the Treated (ATET), these assumptions are relaxed. So that CIA: $Y(0), \perp\!\!\!\perp D|P(X)$, only outcomes of control individuals have to be independent of treatment, and common support: $P(D = 1|X) < 1$, so there might be individuals with 0 probability of being in the treatment group.

Therefore, assuming CIA and common support and following Caliendo and Kopeinig (2008), the Propensity Score Matching estimator for the ATT for individuals on the common support region is:

$$\tau_{ATT}^{PSM} = E_{P(X)|D=1}\{E[Y(1)|D = 1, P(X)] - E[Y(0)|D = 0, P(X)]\} \tag{5}$$

So that it uses only the common support region for the estimation, and applies the propensity score to the mean differences between treatment and control. It is important to note and stress that this τ_{ATT} is only of those individuals falling under the common support region, and that this effect could be totally different (even opposite directions) if off-support individuals were taken into consideration.

Two more elements of propensity score matching have to be explained before finishing this section. Those are the estimation of the propensity score and the choice of the matching algorithm, which is the way matched pairs for the treated individual are chosen.

5.5 Estimating Propensity Score and Matching Methods

As described before, the propensity score $P(X)$ is intended to represent the probability of a particular individual of being treated based on his observables characteristics, X.

This value ranges within the unit interval and one as to decide which model is used to calculate it. Commonly, binary choice models are preferred over linear probability models due to the shortcomings of the latter when the outcome variable is dichotomous. Particularly, I use a logit model to estimate the scores. According to the literature (see Caliendo and Kopeinig (2008) or Smith (1997)), logit and probit models yield comparable results when variable is binary. Once propensity scores have been calculated, one need to choose the matching algorithm that will assign neighbours to treated individuals. It is useful to test some of them and then choose one based on the balance achieved between treatment and control groups in terms of the propensity score. The algorithms are explained in the next lines.

Nearest Neighbour Matching: Treated matched pairs are chosen according to the propensity score, so the individuals with the closest scores are paired. In this matching method, individuals from the control group can be used only once or more times. In the case they are used only one time, it is said that matching is without replacement and that individual cannot be used another time even though he could be the best match for more than one treated. In the other case, an individual from the control group can be used more than one time to match treated individuals, and matching is then called ”with replacement”. It is important to note that before matching, in the without-replacement case, data is sorted randomly so that matching is not based on non-random sorting like area of residence or birth year. In the case of this thesis, replacement is allowed. This way we relax the strictness in the matching algorithm, avoiding forced matches between distinct individuals. If replacement is not allowed, it might be the case that two distant individuals are matched together because more similar individuals have already been used. With replacement, the variance of the estimators is increased but the bias is reduced because average matching will be better. Replacement is needed in our case due to the distribution of the propensity scores[26]. Furthermore, a number of neighbours used for each treated need to be chosen. If more than one neighbour are used, variance will be reduced since there is more information at use to construct the counterfactual, but it increases bias from an average poorer matching since one treated individual is paired to more than one untreated. For the purpose of this thesis, I will present NN-matching 2 and 5 neighbours, as it is usually the case in the literature.

Radius Matching: The idea behind this matching algorithm is to use all the available neighbours within a predefined distance to the treated individual. It allows oversampling and prevents bad matches, since only allows individuals to match when their scores are within the caliper $\|P_i - P_j\| < \varepsilon$, where ε is the selected caliper, so that if the closest individual is farther than the caliper distance defined, that treated observation is dropped out, immediately applying the common support assumption. The counterfactual is then formed by the mean outcome of all the members of the control group that laid within the radius, contrary to the caliper matching, which just chooses the nearest individual(s) from within the caliper to form the counterfactual. The value of the radius has to be set. Garrido *et al.* (2014) and Austin (2011)[27] recommended, based on

[26] $E[P(X)|D=1] = 0.67$ and $E[P(X)|D=0] = 0.48$. And 25% of the distribution of treated have $P(X) > 0.8$ while only 5% of the non-treated have $P(X) > 0.8$.

[27] There is no real consensus in the literature as to how caliper must be set.

Monte Carlo simulations, to set the caliper equal 0.2 times the standard deviation of the logit of the propensity score when using caliper matching. Using this way of setting the caliper in the our case, the matched sample turn out to be imbalanced. A caliper of 0.01 report better balanced sample and do not restrict the common support region as much as the other option. Note that these guidelines were recommended for caliper matching, not for radius matching, and there are not specific ways to set the caliper when radius matching is used.

Kernel Matching: Kernel matching uses weighted averages of all individuals in the control group to construct the counterfactual. Kernel matching assigns a weight of one to the treated individual and the counterfactual is created by weighting all non-treated individuals on their distance in propensity scores from the treated within a range of the propensity score. This range is called bandwidth and it is an important parameter choice in this particular matching method. Higher bandwidth smoothes the estimated density function, which reduces the variance. However, it might not capture underlying characteristics of the untreated individuals, and therefore reporting biased estimates. In the estimation, I use a bandwidth of 0.06 as recommended by Heckman *et al.* (1998). Additionally, one have to choose the kernel function. I use a biweight kernel, although this choice is relatively unimportant in practice (DiNardo and Tobias, 2001).

Once the methods used in this thesis have been explained, standard errors have to be addressed since the interpretation of the coefficient depends on them. Standard errors calculated by the propensity score matching are likely to be biased and not asymptotically valid (Abadie and Imbens, 2016). It is the result of the fact that they are estimated without considering that propensity scores are not the real ones, but they have been previously estimated through a logit or probit model. Neither they account that only common support observations are used in the matching. While lots of empirical works rely on bootstrapping the whole matching process, Abadie and Imbens show, in a series of consecutive papers, that bootstrap standard errors are not valid when using simple nearest-neighbor matching estimators with replacement (Abadie and Imbens, 2008, 2006). They provide a consistent estimator of the sample variance that works in these settings. In the present paper, Abadie & Imbens (AI)[28] standard errors are used with neighbor matching, and bootstrap standard errors with 500 repetitions when other than NN matching are used. With bootstrap, the whole matching process is replicated 500 times, drawing a new sample of the same size with replacement each time, so that the distribution of the resampled estimates (including standard errors and ATTs) is used to approximate the sampling distribution of the original sample, and therefore obtaining more accurate estimates taking into account the fact that propensity scores are estimated.

The last part of the process is to check whether control and treatment matches and weights are balanced so that there are no after-match differences after conditioning on propensity scores. Balance is checked through different indicators explained below. Recent literature (Garrido *et al.*, 2014) has explicitly discouraged the use of t-tests to do so, since these are highly dependent on sample sizes. Mean bias, pseudo R^2 and standardized bias after matching are used as balance indicators. If matching is balanced, control and treatment groups have to be relatively similar on all covariates, so that mean bias of the matched sample is significantly lower than that of the unmatched sample. Mean bias is defined as the average standardized bias of all the covariates. Standardized bias for a continuous variable is defined as the difference in sample means between treated and control divided by the squared root of the average of sample variances in both treated and control groups.

[28] See Abadie and Imbens (2006) for an extensive explanation on how to construct a valid estimator of the asymptotic variance used to obtain AI standard errors.

$$SB_{cont} = 100 \cdot \frac{(\hat{x}_1 - \hat{x}_0)}{\sqrt{0.5 \cdot (\hat{s}_1^2 + \hat{s}_0^2)}} \tag{6}$$

where $\hat{x}_1$ ($\hat{x}_0$) denote the sample mean of X in the treatment (control) group and $\hat{s}_1^2$ ($\hat{s}_0^2$) is the sample variance of X in the treatment (control) group. However, when the covariate is dichotomous, only means of the underlying Bernoulli distributions are needed, as shown by the standardized bias equation:

$$SB_{dich} = 100 \cdot \frac{(\hat{p}_1 - \hat{p}_0)}{\sqrt{0.5 \cdot (\hat{p}_1(1-\hat{p}_1) + \hat{p}_0(1-\hat{p}_0))}} \tag{7}$$

where $\hat{p}_1$ ($\hat{p}_0$) is the mean of the binary variable in the treatment (control) group.

Some researchers (Caliendo and Kopeinig, 2008) recommend that after matching, this value has to be around 3% to 5% in empirical studies, while others like Austin (2009) allow maximum standardized differences for some covariates of around 10% to 25%. Mean bias reported in the results is the average of the standardized differences after matching. The pseudo R^2 indicates how well the covariates X explain the participation probability. It should be low since after the matching, there should be no differences in the distribution of covariates. Finally, the ratio of variances of the covariates of control and treatment group also serves as a balancing tests. According to Rubin (2001), it should be close to one, and a value below $\frac{1}{2}$ or above 2 is indicative of imbalances.

5.6 Choosing Covariates

Covariates included in the model have to be chosen according to some kind of rules. According to Caliendo and Kopeinig (2008), only variables that influence simultaneously the participation decision and the outcome variable should be included in the propensity score matching model. These variables have to be also independent of participation or anticipation of treatment. This means that only variables fixed over time or measured before treatment guarantee that the condition is met. However, if it is expected and feasible to believe that a variable is not affected by treatment, although it is not fixed or measured before it takes place, it can also be included. In the literature, there is no consensus about which variables to include or exclude, no rules of thumbs, and there are arguments in favor and against over-parametrization. On the one hand, Rubin and Thomas (1996) posed themselves in favor of including all the variables expected to influence the outcome variables, even if it is not statistically significant. They argue that variables might be excluded if they are highly correlated to variables already included in the model or if treatment and control groups are extremely similar in those variables. On the other hand, including more variables exacerbates the common support problem, making the common support condition more difficult to meet. Additionally, it has been demonstrated by Brookhart *et al.* (2006) and Bryson *et al.* (2002) that including non-significant covariates might increase the variance of the estimates without reducing bias significantly. In this thesis, only variables that are assumed to influence the participation and the outcome variable are included in the model. Those variables are either fixed over time (gender, parents education, ethnic group, region) or have occurred before treatment (number of children born to the mother). Additionally, those variables are assumed not to be influenced by participation. Table 4 shows the variables included in the main estimation. Certainly age and sex are associated to the outcome and treatment variable and need to be included. Father's and mother's literacy and education levels are also assumed to affect participation and outcome. Region and community type influence the decision of participation as well as outcomes since they reflect general fixed characteristics of specific areas, including preschool and school presence, transportation facilities or community aids. For the rest of variables included, there is a precise explanation for the reason of their inclusion in the next lines.

There are some concerns regarding some characteristics that are important to account for, but variables available seem not to be valid. Child height and child weight are measured in Round 1, when children are 8 years old, so they are probably affected by participation. The problem here is that one would like to have a variable that describes children's health before participation in preschool. The dataset does not provide information on birth weights, breastfeeding or weight measures before preschool. In order not to violate any matching assumption, an indicator of the presence of long term health is used as a proxy for children health status. However, this variable might not reflect children's health in the same way as prenatal information and newborn weights do. Similarly, family wealth is not fixed over time, neither collected before treatment occurs. It is collected in Round 1, when child is 8 years old, but it is difficult to believe that preschool attendance of a child affects family wealth index at the age of 8. Base on this argument, wealth index[29] is included as a covariate.

It is also important to mention that according to Brookhart *et al.* (2006), in small datasets, irrelevant covariates inclusion may introduce too much noise to the treatment effect estimates, increasing the variance. They recommend to exclude variables that are weakly associated with the outcome. In the present case, the final model has been decided as a combination of theory, data availability and "Hit or Miss" method (Heckman *et al.*, 1998). This method sequentially includes more covariates to the baseline model, which in this thesis was only age and region, and keep only those covariates that

[29]Wealth Index construction was explained in footnote 22

Table 4: Covariates used in the estimation

Variables	Description
Children Background	
sex	Child gender
agemon	Child age measured in months
agemon^2	Child age squared measured in months
hplong	Indicator for long term health problems
Parental Background	
wi	Wealth index
chdborn	Number of children born to the mother
dadedu	Father's maximum education achieved
momedu	Mother's maximum education achieved
dadlit	Father is literate (can read and write)
momlive	Indicator whether mother still alive and living in the household
Community	
region	Region
typesite	Indicator rural/urban area
schdis	Distance to preschool

Source: Own elaboration using Young Lives data.

increase the within-sample correct prediction rate, expressed as:

$$1 - \frac{1}{n}\sum_{i=1}^{n}(d_i - I_{\hat{p}_i(X)\geq\tilde{p}})^2 \tag{8}$$

where $\tilde{p}$ is the proportion of individuals in the sample that receive treatment. So that $I_{\hat{p}_i(X)\geq\tilde{p}}$ is an indicator of whether individual i has a higher propensity score than the proportion of treated individuals.

The final specification predicts the participation with around 68% correct prediction rate. It might seem low compared to other values in other different scenarios, like Heckman *et al.* (1998)'s study of the selection bias using data from the Job Training Partnership Act (JTPA), where the best prediction rate was above 80%. Although sample sizes are similar in this case, it can be expected from an actual randomized experiment, meant to study a particular issue, that the variables collected will have more power and meaningfulness than in the present case, where data from Young Lives is not designed exclusively to match this purpose. In addition, preschool is a parent's choice while the JTPA is purposely assigned.

6 Results

The aim of this section is to introduce the estimation results of the effects of preschool on different outcomes in Andhra Pradesh using the strategies explained in Section 5. After which it will be possible to compare results and have a better understanding of what are the mechanisms behind preschool attendance and dropouts.

6.1 Main Estimation

I now present the estimation of the propensity score matching model introduced in Section 5. As mentioned, I use four different matching algorithms in order to assess the balance of each and discuss the coefficient estimates in all of them.

Table 5: ATT of preschool attendance on dropout using 4 different propensity score matching methods.

		2-NN[a]	5-NN[a]	Radius[b]	Kernel[b]
ATT	Preschool	-0.073 **	-0.076**	-0.066	-0.054
		(0.0421)	(0.0387)	(0.0413)	(0.464)
Balancing Tests	Mean Bias	4.4	3.9	3.6	4.6
	Pseudo-R^2	0.029	0.022	0.019	0.039
	R	0.96	1.03	1.09	1.20
	Obs	927	927	918	927

Significance level: $***p \leq 0.01$, $**p \leq 0.05$, $*p \leq 0.1$.
Number of observations only show sample size after dropping off-support individuals. [a] AI Std. errors in parentheses (Abadie and Imbens, 2006).
[b] Bootstrap Std. errors in parentheses calculated with 500 repetitions.

Having a look at the different matching methods used, there are some differences in the results obtained. Table 5 show the ATT estimation using dropouts as dependent variable. Starting with 2 nearest neighbours, the effect of preschool on dropout before finishing secondary education is a significant -7.3%. Increasing the number of neighbours to 5, this effect increases to -7.6% at a 5% significance level. Balance tests in this matching method are somewhat better than in the previous 2-NN, with 3.9% mean bias, under the recommended threshold of 5%. Both of the neighbour matching estimations have similar ratio of covariances, being really close to 1, indicating that the samples are balanced. Next method used is radius matching, with a caliper of 0.01. The effect of preschool on dropouts is -6.6%, a value a bit lower than the obtained through neighbor matching, and insignificant at conventional levels. Radius matching achieves similar balance to the methods in previous columns, indicating a balanced matching. In the last column the results from the kernel matching are shown. The effect in this case is smaller than in the previous matching methods (-5.5%) and insignificant at convectional significance levels. Additionally, balance achieved by this method is slightly worse than in the others. Appendix H presents the standardized bias histograms of these models. In view of these results, it can be said that according to the model and methods used in this thesis, preschool has a significant effect on dropout reduction for children in the Indian region studied, being this effect around 7%. The insignificance of radius and kernel matching might come from the different parameters choices made in their estimation. With caliper and bandwidth as important parameter choices it might be that they are restricting the results. In Section 7 I will discuss this case further and check the robustness of the estimation.

It is worth to mention that differences in sample sizes in the different matching algorithms comes from the fact that the region of common support varies across them. There are 16 individuals off-support in Nearest Neighbor matching and Kernel matching, and 25 in Radius matching. Appendix E gives a brief overview of the nature of these off-support individuals and distribution of the propensity scores.

In addition to the effect on dropouts, I also check whether preschool attendance has positive effects on the completion of primary and secondary education. Results from this estimation are presented in Table 6 and 7, using the same four matching methods used previously.

Looking at Table 6, all four matching methods reflect similar significant effects from preschool on completion of primary school. Generalizing, the same pattern arises here among methods, the ATT is significant and around 4%. It can be concluded that attending preschool boosts primary completion.

Table 6: ATT of preschool attendance on completion of primary education using 4 different propensity score matching methods.

		2-NN[a]	5-NN[a]	Radius[b]	Kernel[b]
ATT	Preschool	0.039**	0.043**	0.036*	0.043*
		(0.0226)	(0.0214)	(0.0218)	(0.0242)
Balancing Tests	Mean Bias	4.4	3.9	3.6	4.6
	Pseudo-R^2	0.029	0.022	0.019	0.039
	R	0.96	1.03	1.09	1.20
	Obs	927	927	918	927

Significance level: ***$p \leq 0.01$, **$p \leq 0.05$, *$p \leq 0.1$.
Number of observations only show sample size after dropping off-support individuals. [a] AI Std. errors in parentheses (Abadie and Imbens, 2006).
[b] Bootstrap Std. errors in parentheses calculated with 500 repetitions.

Moving on to the results presented in Table 7, preschool seems to have a greater effect on completion of secondary, although it is barely significantly different from 0. These differences between primary and secondary are important. We have seen how primary education is less affected by dropout than secondary education, then it is interesting that preschool benefits most completion of secondary. However, only one of the four matching methods used is significant, only at a 10%, so one has to be cautious making final conclusions based on them. Further studies must be done in order to keep exploring the extent of these effects.

Since covariates and matching methods used in tables 5, 6 and 7 are the same, all three tables present the same values on their balance indicators. Kernel matching is the method that presents more difficulties to achieve balance.

Table 7: ATT effect of preschool attendance on completion of secondary education using 4 different propensity score matching methods.

		2-NN[a]	5-NN[a]	Radius[b]	Kernel[b]
ATT	Preschool	0.048	0.058*	0.044	0.021
		(0.0414)	(0.0364)	(0.0395)	(0.0444)
Balancing Tests	Mean Bias	4.4	3.9	3.6	4.6
	Pseudo-R^2	0.029	0.022	0.019	0.039
	R	0.96	1.03	1.09	1.20
	Obs	927	927	918	927

Significance level: ***$p \leq 0.01$, **$p \leq 0.05$, *$p \leq 0.1$.
Number of observations only show sample size after dropping off-support individuals. [a] AI Std. errors in parentheses (Abadie and Imbens, 2006).
[b] Bootstrap Std. errors in parentheses calculated with 500 repetitions.

6.2 Heterogeneous effects

In this subsection, I explore the effects on different subgroups to see whether some of them benefit more from preschool attendance than other. So ultimately, possible policies will be able to better target their treatment groups and be more efficient.

I show only results from 5 nearest neighbours matching estimation because it has been shown that it achieves fairly good balance. Radius matching and kernel matching require precise calibration of bandwidth and caliper radius in order to be reliable.

Table 8 first shows the estimation dividing girls and boys. It seems that girls are way much benefited by preschool than boys, in terms of coefficient size and significance. While girls that have attended preschool experience 12.4% less dropouts, this effect is only 6.7% for boys, and insignificant at conventional significance levels. This is expected to happen because female population is in higher disadvantage in developing countries. Families usually invest more in males for their education, since it is seen that they will be able to take care of their parents later in their lives (Banerjee and Duflo, 2012). More girls attend preschool than boys in the sample, and more girls dropped out of school too. If there is an effect of preschool years on dropouts, then girls will be much more affected by it, as it is shown from this result.

Table 8: ATT of preschool on dropouts using 5 nearest neighbours matching by gender.

		Female	Male
ATT	Preschool	-0.124**	-0.067
		(0.0535)	(0.0620)
Balancing Tests	Mean Bias	5.0	6.2
	Pseudo-R^2	0.035	0.044
	R	1.04	1.06
	Obs	441	417

Significance level: ***$p \leq 0.01$, **$p \leq 0.05$, *$p \leq 0.1$. AI standard errors in parentheses (Abadie and Imbens, 2006)

Although sample size are reduced due to the use of subgroups, the matched samples seem to be balanced as shown by the balancing tests, so these number are not a result of an imbalanced sampled.

I also check whether urban areas have larger preschool effects. Table 9 show the different effects of preschools in urban and in rural sites.

Using the urban sample alone, the matching algorithm does not achieve balance. Although the effect is high and significant, indicating that in fact preschools in urban

Table 9: ATT of preschool on dropouts using 5 nearest neighbours matching by community type.

		Urban	Rural
ATT	Preschool	-0.371***	0.008
		(0.1217)	(0.0475)
Balancing Tests	Mean Bias	11.5	3.0
	Pseudo-R^2	0.215	0.013
	R	-	1.00
	Obs	184	716

Significance level: $***p \leq 0.01$, $**p \leq 0.05$, $*p \leq 0.1$. AI standard errors in parentheses (Abadie and Imbens, 2006)

sites are likely to be associated with lower dropouts. However, we cannot draw any conclusion about causal urban preschool effects. With respect to rural sites, the effect is null. Since the balance is right, one can infer that rural preschools have not enough influence on children dropouts. One reason might be that preschools in rural areas might have a less stimulating effect on children due to their poor environment. As mentioned by Banerjee and Duflo (2012), teacher absence is high in urban areas, so that it could be the same for preschool teachers.

Next I explore to what extent the type of preschool influence the outcomes. Preschools can be either private preschools or government preschools. This does not imply that private preschools require payments by families and government preschools are free. Both of them can be either paid or non-paid, it depends on the area and the specific preschool center. But private preschools have their own private teachers, guidelines and principals, while public preschool rely on the government support to provide teachers and protocols. Table 10 shows these coefficients. As it was expected, it seems that private preschools contribute better to children school outcomes, with a significant 12.9% effect on dropout reduction. On the other hand, public preschools also have a significant effect on dropouts, but it is smaller (7.4%). However, the private preschool estimation have problems to balance the matched sample while the public one seem to be balanced. This is consistent with the official reports from ASER where private schools seem to have greater effects than public schools. In 2016, 75% of children in private school in the state of Haryana could read English sentences while only 30% could do it in the same state. This numbers are similar across all the states surveyed (UNESCO, 2015a). Since a large proportion of preschools were parts of school centers, this might well be the case for preschool.

Table 10: ATT of preschool on dropouts using 5 nearest neighbours matching by preschool type.

		Private	Public
ATT	Preschool	-0.129*	-0.074**
		(0.0836)	(0.0424)
Balancing Tests	Mean Bias	9.1	2.4
	Pseudo-R^2	0.077	0.007
	R	1.07	0.78
	Obs	603	764

Significance level: ***$p \le 0.01$, **$p \le 0.05$, *$p \le 0.1$. AI standard errors in parentheses (Abadie and Imbens, 2006)

Lastly, it is also important to explore to what extent duration of preschool exposure have an effect on dropout. Table 11 presents these coefficient estimates.

As expected, more preschool exposure is associated with better outcomes during school. One year of school has less effect than two years of preschool (−6.3% and −9.8%). Both seem to achieve balance on their matched samples. Furthermore, three years of preschool seem not to have an effect, but this is definitely driven by the overlap between preschool and school that occurs during the last year of preschool and first year of school, where some of the children have already started school with 5 years old.

Table 11: ATT of preschool on dropouts using 5 nearest neighbours matching by preschool duration.

		1 year preschool	2 years preschool	3 years preschool
ATT	Preschool	-0.063	-0.098**	-0.025
		(0.0529)	(0.0556)	(0.0739)
Balancing Tests	Mean Bias	3.1	4.5	6.0
	Pseudo-R^2	0.018	0.030	0.057
	R	1.22	0.77	1.20
	Obs	612	591	467

Significance level: ***$p \leq 0.01$, **$p \leq 0.05$, *$p \leq 0.1$. AI standard errors in parentheses (Abadie and Imbens, 2006)

Once results have been presented, it can be concluded that they are in line with what it is known so far. Disadvantaged groups of the population seem to benefit more from these extra years of education, or preparation, as well as rural and public preschools seem not to have such a great effect as private and urban-located preschools. Finally, it seems to be clear that the duration of preschool exposure is also an important determinant of future outcomes. Only one year of preschool has statistically non-significant effect on dropouts. However, with at least two years of preschool, the coefficient is statistically significant. Not only preschool exposure is important but also the duration of the same. In order to have a significant effect and stimulation in the child, treatment has to be persistent for more than a year.

7 Robustness and Alternative Specification

In this section several alternative estimations are presented in order to show that the estimation is robust to different model specifications and variables. In previous sections, it has been already demonstrated that the estimates do not vary significantly with different matching algorithms, using both matching and weighting methods. That is already a first indication that the specification is consistent to different matching methods. One additional weighting method (Inverse Propensity of Treatment Weighting) is included here due to the doubtful balance of Kernel Matching. I also show that results from simple OLS regression and logit are in line with the matching estimates. Lastly, I also run the same estimation but including some of the variables that have been left out due to both personal and theoretical considerations to see whether their inclusion makes a large difference in the results. Together with it, I also estimate the ATT using a crude model, consisting of only child's background characteristics.

7.1 Methodology Variation

First, I run an OLS regression using the same covariates as in the propensity score matching estimation. According to Angrist and Pischke (2008), regression should be a first shot for most empirical projects. One main explanation for this might be, as Angrist and Pischke (2008) points, the view that some researchers have of the Linear Probability Model as a Conditional Expectation Function (CEF) of Y. What it is tested is the probability of an outcome occurrence given that the subject belongs to one of the groups (control or treatment). As with the main estimation, I will be estimating $\tau_{ATT} = E[Y_i|X_i, D_i]$. Following Brand and Halaby (2006), assuming that our outcome variable is linear in covariates X, treatment effects are constant across all units and unconfoundedness hold, parameters of equation 9 can be estimated from the regression.

It can be assumed that the conditional expectation of the outcome is linear in its explanatory variable, so that:

$$\begin{aligned} \mathbf{E}(y_i|x_i;\beta_0,\beta_1) &= \beta_0 + \beta_1 x_i \\ \Leftrightarrow y_i &= \beta_0 + \beta_1 x_1 + \varepsilon_i \end{aligned} \tag{9}$$

The model can be written:

$$Y_i = \gamma + \alpha_{ATT} D_i + \beta X_i + \varepsilon_i \tag{10}$$

Standard errors are not homoskedastic in the case of cross-sectional data if the parameters of the covariates are different from 0, it is necessary to account for heteroskedastic standard errors. This is done by estimating robust standard errors, that provides a consistent estimate of the variance of the estimator when the variance of the error term is assumed to vary over i. However, besides heteroskedasticity, microeconometricians have been increasingly focusing on within-group dependence. Moulton (1990) was the first to notice that if observations share common locations, industries, or classes in general, also share unobservable characteristics that would make disturbances to be correlated. If within-group dependence is not accounted for in an OLS estimation, true OLS standard errors are quite likely underestimated. Therefore, there is a risk of finding statistically significant coefficients, when true coefficients are not. This problem arises from the fact that usually observations are not independent from each other within-clusters. And this is certainly the case in the Young Lives data. Since the main sampling units from Andhra Pradesh are communities, it is possible that individuals from the same community share some common characteristic and are not independent of each other, as basic OLS estimation assumes. So that simple random sampling is not the right assumption. I will assume that this is the case here, so that there are observations that depend on each other within-clusters and hence, iid standard errors assumption is not correct for the estimation of true standard errors. Even the robust standard errors mentioned earlier fall into problems when the asymptotic approximation is not good (Angrist and Pischke, 2008).

Discussions about clustering standard errors have taken place since the 90s. In order to control for clustering, one usually uses the cluster-robust variance matrix estimate[30]. Additionally, in order for this method to hold, one have to impose an additional assumption: individuals are correlated withing clusters, but uncorrelated across clusters. Cameron and Miller (2015) state that one should, at minimum, cluster at the level of primary sampling unit, which in our case is community level. For these reasons, I decide to use clustered robust standard errors at community level. Having over a 100 communities[31] in the first round is enough to use this technique.

Table 12 shows coefficient estimates of the OLS estimation, with standard errors clustered at community level in parentheses. What these results show is a similar pattern with the results obtained through propensity score matching. Two years of preschool have greater effects in dropout reduction and completion of primary than only one year of preschool. For completion of secondary school, it seems that the effect of preschool is similar disregarding its duration. Three years of preschool might not be reliable due to the overlap between school and preschool that occur when children are 5 years old. Also, they reflect, as in the propensity score matching case, that preschool has a significant effect on completion of primary and secondary. Additionally, effect sizes are

[30]For an technical explanation of cluster-robust variance matrix, see Appendix F.

[31]As for information purposes only, Angrist and Pischke (2008) and Cameron *et al.* (2008) both stressed the fact that there need to be a minimum number of clusters for valid inference. Cameron and Miller (2015) declare ”too few” clusters when they are in range of 20-50.

Table 12: OLS Estimates

	(1) Dropout	(2) Completion Primary	(3) Completion Secondary	(4) Higher Education
Preschool	-0.082***	0.039*	0.067**	-0.021
n=950	(0.0305)	(0.0203)	(0.0304)	(0.0238)
1 year	-0.068*	0.032	0.074*	-0.018
n=637	(0.0393)	(0.0264)	(0.0413)	(0.0289)
2 years	-0.103**	0.053**	0.071*	-0.025
n=615	(0.0411)	(0.0234)	(0.0426)	(0.0268)
3 years	-0.091	0.47	0.058	0.045
n=496	(0.0657)	(0.0325)	(0.0650)	(0.0452)

Significance level: ***$p \leq 0.01$, **$p \leq 0.05$, *$p \leq 0.1$.
Cluster-robust standard errors at community level in parentheses.

relatively similar between the main estimation and this OLS estimation, which indicates that results are not driven by the estimation strategy.

To validate the OLS estimation, I also estimate the coefficients using a logit model. This model assumes different error term distribution, and takes into account the fact that the dependent variable is binary. Results from this estimation show a coefficient of -0.538 significant at a 1% level. The predicted probabilities of this coefficient are similar to the OLS regression, with a -10.2% marginal effect of preschool on dropouts. Although this result is larger than the obtained through OLS and matching, it is still significant and negative, indicating that the estimation strategy is not solely driving the results.

7.2 Different covariates

To check whether the variables chosen to perform the analysis with are not extremely conditioning the results, I will also run the propensity score matching analysis using the variables not included in the main model due to failing to increase the within-sample prediction rate and other personal considerations, defined in Section 5. Those variables are mother's education, an indicator variables of whether the father lives in the household or not and primary occupation of the household. In addition to that specification, I run the same propensity score matching analysis using a crude model where I only include child characteristics as covariates. Appendix G presents the results from the same matching methods use for the main estimation but with these two different model specifications. Table G.3 presents the ATTs using the crude model. These coefficients are pretty similar to the ones in the main model. Kernel matching and 2-NN estimations are further distant from the main results, which imply they are more sensitive to the variable choice. Tabel G.4 presents the estimates of the other model specification. The coefficients of the estimation change minimally, being these last estimates a bit smaller and less significant. In this model specification it is Kernel Matching the method that achieves better balance, indicating that it could be a better choice for the ATT for this particular specification. The fact that these specifications produce similar results as the main estimation is a good way to make sure that previous estimates are not the result of a specific variables choice. However, as it has been mentioned, Brookhart *et al.* (2006) proved in a simulation study that the inclusion of less relevant variables that are not expected to have significant effects on the outcome might increase variance without reducing bias significantly. This might certainly be the case with the additional variables used here. They did not increase the within-sample correct prediction rate, were barely significant in the logit model used to estimate the propensity score and their potential effect can already be accounted for by other variables. Nonetheless, they serve us to present another robustness measure.

7.3 IPTW

Worried about how kernel matching has performed on the different samples, achieving doubtfully balanced samples, generating insignificant and slightly different results, I will lastly use Inverse Probability Treatment Weighting (IPTW). This method uses weighted averages of the observed outcome variable to estimate means of the potential outcomes. Each treated individual is assigned a weight of $\frac{1}{\hat{p}}$ and each individual in the control group is weighted by the inverse of the probability of treatment $\frac{1}{1-\hat{p}}$. The reason for using this method is that it is another weighting method, as kernel matching, contrary to neighbor matching and caliper, which are based on matching observations rather than weighting. The ATET estimate is -9.96% at a 1% confidence interval. This is a larger effect that the one found with the other matching methods, but it still in the same line, negative and significant.

Previous robustness checks have demonstrated that similar results emerge from different estimation strategies and different set of covariates. This is indicative that the results can be considered robust.

8 Discussion and Limitations of the thesis

This section will go through the most controversial part of the study, providing argumentation and intuition of several controversial points that have arisen throughout the study.

The most important point to note here are the assumptions on which this thesis is based. With the data available to me for this study, the most reasonable way to approximate the results to a causal relationship is by using propensity score matching. But as it has been mentioned throughout the thesis, this requires three important assumptions to hold. The most controversial one is the *conditional independence assumption* or unconfoundedness. It is difficult to believe that one can observe all and each one of the variables that affect treatment and outcome at the same time. This is a common characteristic of quasi-experimental design studies with observational data. Although the dataset used in this thesis is quite detailed and rich, the fact that the first round of interviews is done in 2002, when children are around 8 years old, and that the main purpose of the Young Lives project is poverty reduction and not exclusively early childhood care and education, hamper my ability to account for all possible confounders. Important determinants of participation are left out of the equation, or poorly controlled for. The most worrisome variable among this kind of studies is *children's ability*. If parents decide to send children to preschool based on their capabilities, social and cognitive abilities, then it is more likely that the preschool effect derived here is overestimated. If only ”high-ability” children attend preschool, it is natural that they perform better and stay longer at school because they are better able to. In order to capture and collect ability indicators, cognitive tests are usually administered. Given the age of preschoolers, these tests have to be undertaken before the age of 3. That information is not available in the old cohort of the Young Lives data. Other possible variables choices to account for ability might be parents' view of children abilities. Questions about child's aptitudes or expectations about the future of the child can be asked so that one can evaluate characteristics like curiosity, social interaction or playfulness, which are deemed to be indicators of ability in early years. This kind of questions are not present in the questionnaires either, at least before preschool participation. However, India is not considered a developed country, with a life expectancy of 68 years old at birth and GDP per capita of US$ 1.709[32]. It can be argued that there are other important reasons for a family to send their children to preschool, like acknowledgement of the returns to education, presence and availability of preschool in the region, preschool meals or community information and promotion of preschools. Ability has cornered the non-experimental studies about returns to schooling, starting from primary education, but it might be the case that the determinants of preschool are not the same as for school. It is reasonable to think that the main benefits of preschool to Indian families are not centralized around their children completing school. There are other feasible reasons for families to send their children to preschool like facilitating mother's labor force, being a safe place to keep the children during the day, to keep them busy, or to keep them fed. If this is the case, the fact that I cannot account for children's ability should only be considered as minor overestimation bias.

Another important factor has arisen in the previous lines and deserves a mention here: *parental view on education*. It is an undeniably confounder. Parents who do not value education enough will prefer to keep their children at home rather than send them to preschool. This will also influence the investment made on educating their children through the rest of their lives. Therefore, it surely influences preschool attendance and also completion rates or dropout rates. Parents who think education has great returns

[32]The World Bank data for 2016. Accessed: 24 July 2017.

will make sure their children complete primary and even secondary, while parents who do not trust the education system will see larger returns sending their children to work. In this thesis, it is assumed that this value of education is accounted for by parental education, which is not fully accurate but it is the best possible approximation.

With respect to the other two assumptions made throughout this thesis, concerns about common support region has already been mentioned. The reader has to take into account that the ATTs found are over the common support regions and that individuals left out of the common support might have different coefficients and even different signs.

A last mention regarding assumptions is about SUTVA. As in most of this kind of studies, SUTVA is assumed to hold. This assumption might be violated due to peer effects. For example, it is possible that even though a child has not attended preschool, the fact that his classmates in school did, affects him positively or negatively, either by getting help and encouragement from his classmates or by feeling depressed due to his lower level in the class. Then she will be affected by preschool even though she has not attended.

Besides the previous notes about the assumptions, it is also worth mentioning the structures of the control and treatment group. As it has been explained, treatment group consist of children that have attended school for at least six months in a year between 3 and 6 years old. Therefore, and by extension, control groups consist of the rest of children in the dataset. This broad category includes children that attended preschool for period of less than 6 months, children whose mothers are at home and might receive home-care education, and children whose mothers are not at home and that are taken care of by their siblings or other family members. This makes the control group a bit heterogeneous, but without a proper experiment where control group is specifically set to meet some criteria, it is difficult to homogenize it.

Another limitation of the thesis is the small sample size available for the study. Although it seems large enough to proper inference dropout at the population level, it is definitely small if one wants to do inference on subgroups. Hence, the heterogeneous effects can only give us an idea on how preschool attendance affect differently particular groups of the population, but by no means it can be stated that it is accurate.

Due to these limitations, interpretation and generalization of the results obtained in this thesis must be done cautiously. It gives a broad idea about the effects of preschool attendance on medium-term outcomes as completion of secondary. Although its true effect cannot be pinned down and it is likely to be slightly overestimated, the above discussion gives enough reasons to believe that this effect is somewhat larger than zero.

With the fifth round of Young Lives data available in 2018, a whole new range of possibilities opens up to improve the present thesis. Data on secondary education of younger cohort children will be available, making estimation method followed in here stronger since the younger cohort questionnaires include more information of pre-preschool years so that identification of confounders is easier, as well as it has a larger sample size (2.000 children instead of 1.000).

Also mentioning that this thesis has only studied the extensive margin of education - how preschool and duration of preschool affect completion of primary and secondary. But as it as been showed in the literature review, recent studies have highlighted the importance of the intensive margin, of the quality of education on the future of the children. So if the ultimate purpose of education is to eradicate poverty, create equal opportunities to everyone to develop, and bridge the inequality gap between different countries in the world, it would not be enough with guaranteeing completion of elementary or secondary education, nor achieving higher level of preschool enrollment. The goal has to be to guarantee the access to high-quality education, that actually gives a chance to every child. In fact, if primary school is capable of provide good quality education for children, preschool attendance should prepare and provide them with good set of skills so that they can stock more skills and progress during their primary school period. As mentioned by Heckman *et al.* (2012), skill begets skill.

One last point before finishing this discussion. Policies intended at improve preschool attendance and quality need to be framed in a cost-benefit analysis. Although it is besides the scope of the thesis, it is important to bear in mind that not all preschool policies will be efficient, although they achieve positive results. We have seen how the earlier the investment in children, the larger its returns are. But policymakers need to strive for cost reductions, and need to choose the most cost-efficient policies in order to generate larger returns.

9 Conclusion

In this thesis I have estimated the effects of preschool attendance on dropout of children in the state of Andhra Pradesh, in India. Results indicate that attending preschool between 3 and 6 years old have an effect of around 7% on dropout reduction. Particularly, it increases primary school completion significantly, by 4%, and it is still discernible in completion rates of secondary education with a slightly larger effect of around 5.8%. Additionally, it has been found that these effects are heterogeneous, where some subgroups benefit most from preschool attendance than others and where duration of preschool plays a role in children's future outcomes.

These results contribute to the literature of preschool effects in developing countries by increasing the number of non-experimental studies available that find positive preschool effects in academic outcomes, where academic literature is scarce. A novel part of the study is the focus on school completion or dropouts, rather than learning outcomes most used in the literature. Additionally, extra efforts have been made to determine causality.

Using non-experimental data provided by the Young Lives project and a propensity score matching model, I aimed at overcoming the typical self-selection bias problem. This problem arises when some individual characteristics are expected to influence the treatment participation and the outcome variable at the same time, therefore limiting the identification of the true treatment effect.

These non-experimental methods heavily rely on strong assumptions on the observed data. I have argued the extent to which it is feasible that these assumptions hold. Yet these results need to be interpreted cautiously, bearing in mind the limitations of the methodology and data used. With new data from the fifth round of the Young

Lives study coming out in 2018, it is extremely recommended to carry out a similar study on the younger cohort, with larger sample size and richer information on early years.

Based on the results of this thesis, one possible policy intended to reduce dropouts and increase completion of primary and secondary school could be one that promotes and boosts preschool attendance. As we have seen, policies based on inputs are less effective at keeping children in school than policies based on strengthen children abilities, meal provision or even conditional cash transfers. However, it is unclear whether preschool attendance is determined by the same factors as school is. This creates a possible research field in developing countries, where different policies targeted at increasing preschool attendance are to be tested to learn more about their efficiency.

Finally, these policies are deemed to fail if the additional years of preschool bring nothing to children's skills set. Ultimately, it is a two-fold task. On the one hand, making preschool available to everyone and promoting attendance, on the other hand making preschools a place where children actually learn.

Bibliography

ABADIE, A. and IMBENS, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, **74** (1), 235–267.

— and — (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, **76** (6), 1537–1557.

— and — (2016). Matching on the estimated propensity score. *Econometrica*, **84** (2), 781–807.

ABOUD, F. E., HOSSAIN, K. and O'GARA, C. (2008). The succeed project: challenging early school failure in bangladesh. *Research in Comparative and International Education*, **3** (3), 295–307.

AFRIDI, F. (2011). The impact of school meals on school participation: evidence from rural india. *Journal of Development Studies*, **47** (11), 1636–1656.

ANGRIST, J. D. and PISCHKE, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

AUSTIN, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, **28** (25).

— (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, **10** (2).

BANERJEE, A. and DUFLO, E. (2012). *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. PublicAffairs.

BANERJEE, A. V., COLE, S., DUFLO, E. and LINDEN, L. (2007). Remedying education: Evidence from two randomized experiments in india. *The Quarterly Journal of Economics*, **122** (3), 1235–1264.

BARNETT, W. S. (1995). Long-term effects of early childhood programs on cognitive and school outcomes. *The future of children*, pp. 25–50.

BERLINSKI, S., GALIANI, S. and GERTLER, P. (2009). The effect of pre-primary education on primary school performance. *Journal of public Economics*, **93** (1), 219–234.

BRAND, J. E. and HALABY, C. N. (2006). Regression and matching estimates of the effects of elite college attendance on educational and career achievement. *Social Science Research*, **35** (3), 749–770.

BROOKHART, M. A., SCHNEEWEISS, S., ROTHMAN, K. J., GLYNN, R. J., AVORN, J. and STÜRMER, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, **163** (12), 1149–1156.

BRYSON, A., DORSETT, R. and PURDON, S. (2002). The use of propensity score matching in the evaluation of active labour market policies.

BURDE, D. and LINDEN, L. L. (2012). *The effect of village-based schools: Evidence from a randomized controlled trial in Afghanistan*. Tech. rep., National Bureau of Economic Research.

CALIENDO, M. and KOPEINIG, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, **22** (1), 31–72.

CAMERON, A. C., GELBACH, J. B. and MILLER, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, **90** (3), 414–427.

— and MILLER, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, **50** (2), 317–372.

CARD, D. (1999). The causal effect of education on earnings. *Handbook of labor economics*, **3**, 1801–1863.

COPPLE, C. E. *et al.* (1987). Path to the future: Long-term effects of head start in the philadelphia school district.

CUNHA, F. and HECKMAN, J. (2007). *The technology of skill formation*. Tech. rep., National Bureau of Economic Research.

—, HECKMAN, J. J., LOCHNER, L. and MASTEROV, D. V. (2006). Interpreting the evidence on life cycle skill formation. *Handbook of the Economics of Education*, **1**, 697–812.

CURRIE, J. (2001). Early Childhood Education Programs. *Journal of Economic Perspectives*, **15** (2), 213–238.

DINARDO, J. and TOBIAS, J. L. (2001). Nonparametric density and regression estimation. *The Journal of Economic Perspectives*, **15** (4), 11–28.

DUFLO, E., DUPAS, P. and KREMER, M. (2015a). Education, hiv, and early fertility: Experimental evidence from kenya. *The American economic review*, **105** (9), 2757–2797.

—, — and — (2015b). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from kenyan primary schools. *Journal of Public Economics*, **123**, 92–110.

ELANGO, S., GARCÍA, J. L., HECKMAN, J. J. and HOJMAN, A. (2015). *Early Childhood Education*. Working Paper 21766, National Bureau of Economic Research.

ENGLE, P. L., FERNALD, L. C., ALDERMAN, H., BEHRMAN, J., O'GARA, C., YOUSAFZAI, A., DE MELLO, M. C., HIDROBO, M., ULKUER, N., ERTEM, I. *et al.* (2011). Strategies for reducing inequalities and improving developmental outcomes for young children in low-income and middle-income countries. *The Lancet*, **378** (9799), 1339–1353.

GARCES, E., THOMAS, D. and CURRIE, J. (2000). *Longer Term Effects of Head Start*. Working Paper 8054, National Bureau of Economic Research.

GARRIDO, M. M., KELLEY, A. S., PARIS, J., ROZA, K., MEIER, D. E., MORRISON, R. S. and ALDRIDGE, M. D. (2014). Methods for constructing and assessing propensity scores. *Health services research*, **49** (5), 1701–1720.

GLEWWE, P. and KASSOUF, A. L. (2012). The impact of the bolsa escola/familia conditional cash transfer program on enrollment, dropout rates and grade promotion in brazil. *Journal of development Economics*, **97** (2), 505–517.

—, KREMER, M. and MOULIN, S. (2009). Many children left behind? textbooks and test scores in kenya. *American Economic Journal: Applied Economics*, **1** (1), 112–35.

— and MURALIDHARAN, K. (2015). Improving school education outcomes in developing countries: evidence, knowledge gaps, and policy implications. *University of Oxford, Research on Improving Systems of Education (RISE)*.

HAENFLER, R., JOHNSON, B., KING, E. and KREMER, M. (2002). Vouchers for private schooling in colombia: Evidence from a randomized natural experiment. *The American Economic Review*, **92** (5), 1535–1558.

HANUSHEK, E. A. and WOESSMANN, L. (2015). *The knowledge capital of nations: Education and the economics of growth*. MIT Press.

HECKMAN, J., ICHIMURA, H., SMITH, J. and TODD, P. (1998). *Characterizing selection bias using experimental data*. Tech. rep., National bureau of economic research.

HECKMAN, J. J. *et al.* (2012). *The case for investing in disadvantaged young children*. European Expert Network on Economics of Education.

KREMER, M., BRANNEN, C. and GLENNERSTER, R. (2013). The challenge of education and learning in the developing world. *Science*, **340** (6130), 297–300.

MAGNUSON, K. A., RUHM, C. and WALDFOGEL, J. (2007). The persistence of preschool effects: Do subsequent classroom experiences matter? *Early Childhood Research Quarterly*, **22** (1), 18–38.

MASSE, L. N. and BARNETT, W. S. (2002). A benefit-cost analysis of the abecedarian early childhood intervention. *Cost-Effectiveness and Educational Policy, Larchmont, NY: Eye on Education, Inc*, pp. 157–173.

MOULTON, B. R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The Review of Economics and Statistics*, **72** (2), 334–338.

MURALIDHARAN, K. (2013). Priorities for primary education policy in india's 12th five-year plan. In *India Policy Forum*, National Council of Applied Economic Research, vol. 9, pp. 1–61.

MWAURA, P. A., SYLVA, K. and MALMBERG, L.-E. (2008). Evaluating the madrasa preschool programme in east africa: a quasi-experimental study. *International Journal of Early Years Education*, **16** (3), 237–255.

NORES, M. and BARNETT, W. S. (2010). Benefits of early childhood interventions across the world:(under) investing in the very young. *Economics of education review*, **29** (2), 271–282.

PRITCHETT, L. and BEATTY, A. (2012). The negative consequences of overambitious curricula in developing countries. *CESifo Working Paper Series*, (No.4040).

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70** (1), 41–55.

RUBIN, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, **2** (3), 169–188.

— and THOMAS, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, pp. 249–264.

SHORE, R. (1997). Rethinking the brain: New insights into early development. *ERIC. Non-journal report*.

SMITH, H. L. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological methodology*, **27** (1), 325–353.

TAIWO, A. and TYOLO, J. (2002). The effect of pre-school education on academic performance in primary school: a case study of grade one pupils in botswana. *International Journal of Educational Development*, **22** (2), 169–180.

UK Data Service (2016). Young lives: an internationa study of childhood poverty.

UNESCO (2006). *Education for all global monitoring report*. United Nations Educational, Scientific and Cultural Organization.

— (2015a). *Education for All 2015 National Review Report: India*. United Nations Educational, Scientific and Cultural Organization.

— (2015b). *Education for all global monitoring report*. United Nations Educational, Scientific and Cultural Organization.

VERAMENDI, G. and URZÚA, S. (2011). *The Impact of Out-of-Home Childcare Centers on Early Childhood Development*. Tech. rep., IDB Working Paper Series.

A EFA Goals

Education for All Goals (UNESCO, 2015b)

1. Expanding and improving comprehensive early childhood care and education, especially for the most vulnerable and disadvantaged children.

2. Ensuring that by 2015 all children, particularly girls, children in difficult circumstances and those belonging to ethnic minorities, have access to and complete free and compulsory primary education of good quality.

3. Ensuring that the learning needs of all young people and adults are met through equitable access to appropriate learning and life-skills programmes.

4. Achieving a 50 per cent improvement in levels of adult literacy by 2015, especially for women, and equitable access to basic and continuing education for all adults.

5. Eliminating gender disparities in primary and secondary education by 2005, and achieving gender equality in education by 2015, with a focus on ensuring girls' full and equal access to and achievement in basic education of good quality.

6. Improving all aspects of the quality of education and ensuring excellence of all so that recognized and measurable learning outcomes are achieved by all, especially in literacy, numeracy and essential life skills.

B Final Districts Selected

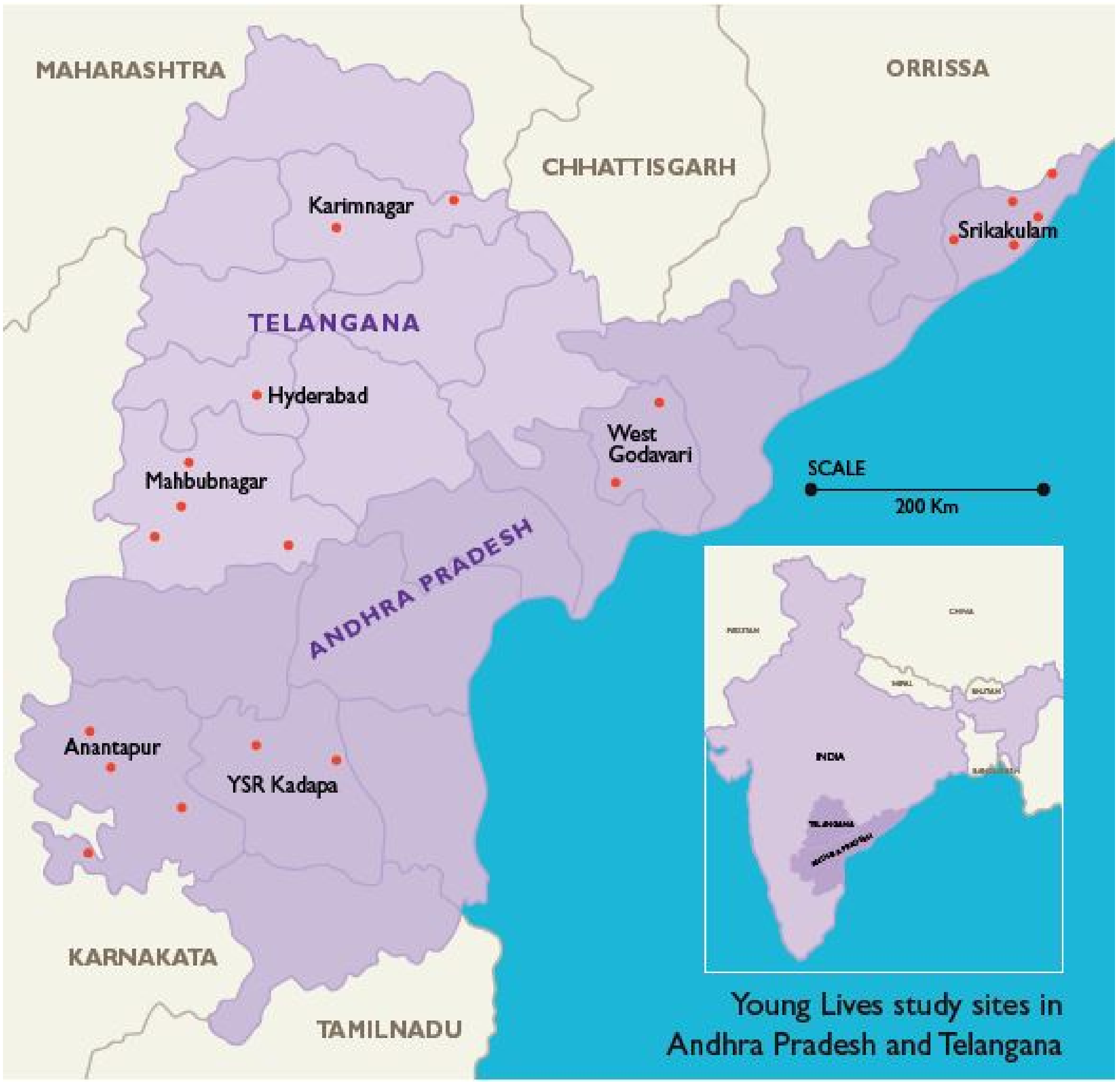
Table B.1: Young Lives Districts

Cluster ID	District	Anonymised name	Short Description
Coastal Andhra			
1	West Godavari	Sagar	An urban area in a well-developed coastal region
2	West Godavari	Raipur	A tribal mandal in a well-developed coastal district
3	Srikakulam	Patna	A town in north coastal Andhra Pradesh
4	Srikakulam	Manipur	A tribal mandal in north coastal Andhra Pradesh
5	Srikakulam	Puri	A tribal mandal in north coastal Andhra Pradesh
6	Srikakulam	Chandipur	A tribal mandal in north coastal Andhra Pradesh
7	Srikakulam	Angul	A rural mandal with a mix of tribes and non-tribes in north coastal Andhra Pradesh
Rayalaseema			
8	Kadapa	Bolangir	A rural mandal in the heart of the Rayalaseema region where agriculture is the main occupation
9	Kadapa	Kalahandi	A remote rural mandal in a forested part of the Rayalaseema region
10	Anantapur	Mayurbhanj	An urban site in the Rayalaseema region, which is a district headquarter
11	Anantapur	Katur	A poor rural mandal in Rayalaseema region affected by Naxalite movements
12	Anantapur	Sivakasi	A poor rural area spread across hilly areas and affected by Naxalite movements
13	Anantapur	Tondi	A rural mandal in the Rayalaseema region bordering the neighbouring state
Telangana			
14	Karimnagar	Dharmapuri	A medium-sized town in northern Telangana with people of mixed religion
15	Karimnagar	Kotagiri	A rural area in northern Telangana affected by Naxalite movements
16	Mababubnagar	Perambalur	A rural tribal mandal in the forest areas of southern Telangana
17	Mababubnagar	Nagore	A rural mandal in the southern Telanagana region with people moving in seasonal migration
18	Mababubnagar	Bhavara	A rural mandal in the southern Telanagana region with a high incidence of child labour and seasonal migration
19	Mababubnagar	Poompuhar	A very poor mandal in southern Telangana
State Capital			
20	Hyderabad	Polur	A densely crowded area in the state capital of Andhra Pradesh and Telangana

Source: Young Lives

C Young Lives. India Sampling Map

Figure C.1: Regions and Mandals used in Yound Lives project



Source: www.younglives.org.uk

D Data cleaning and structure

The data used in this paper comes from a collaborative research project coordinated out of the Department of International Development at the University of Oxford. The project, *Young Lives*, aims to study the drivers and impacts of child poverty in Ethiopia, Peru, India (state of Andhra Pradesh) and Vietnam. This study is following two groups of children through their youth, with a total of 12.000 children. The older cohort children are born in 1994-1995 and the younger cohort children are born in 2001-2002. The study was set to last for 15 years, so that by the end of the study (around 2017), the older and younger cohort children would be around 22 and 15 years old, respectively.

In order to construct the dataset used in this particular thesis, different dataset have been merged. Those are available at the UK Data Service[33]. This series consists of the four rounds of interviews, plus school surveys and a constructed dataset (*Constructed Files*) where specific questions common to all the rounds are merged together. The referred constructed dataset has been prepared by Young Lives quantitative research assistants. Additionally, three questionnaires are available per round. Child Questionnaire, Household Questionnaire and Community Questionnaire. Moreover, depending on the round, subsections of the questionnaires where also available, e.g. preschool subsection.

The final dataset used for estimation has been constructed based on the Constructed Files available from Young Lives. In addition, different questions that were not included in the constructed dataset have been included from other dataset due to their importance for the purposes of the thesis. These variables come from child, household and community questionnaires from every round. The identification variable common to most

[33]Date Series is available at: Series: Young Lives: an International Study of Childhood Poverty [https://discover.ukdataservice.ac.uk/series/?sn=2000060] GN 33379.

dataset was the children identification number (childid), but in some instances, it was a community identification number the link between them (commid). A total of 15 different dataset have been used in the construction of the final one.

Merging and estimation have been done using the statistical software package STATA, version 13. Do-files used to construct the final dataset as well as for the estimation are available in the Online Appendix or under request.

Missingness has been treated differently depending on the type. On the one hand, children not present in every round have been dropped out from the sample. On the other hand, individual missing observations due to randomness have been taken into account by creating a missingness dummy variable. This way, the effect of the missing value of that observation is taken away and reflected in a different coefficient, not influencing the coefficients of interest. Sometimes, the questions or the interviewer provided the respondent the option to answer with ”Not known” or ”Refused to answer”. These cases do not appear extensively in the variables used for the estimation. In the instances where those values appeared, they were categorized as ”Other” or ”Missing”.

E Common Support Regions

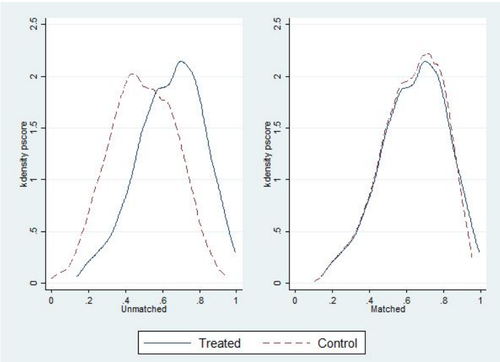
When one estimates the Average Treatment Effects using Propensity Score Matching, it may be the case that some treated individuals have particular propensity score that have no match in the control group. Therefore, they violate the common support condition explained in Section 5. These particular treatment individuals are dropped out of the sample and are not included in the estimation. If they form a small group, one do not have to worry about them, but if they are a significant part of the sample, it needs to be explored to get more information about that particular group.

I show here how is the distribution of the ”off-support” individuals dropped out from the main estimation and their characteristics. There are 16 treated observations not used in the estimation using 2 and 5 nearest neighbour matching and kernel matching, that is a 1.6% of the sample. On the contrary, there are 25 off-support observations using radius matching, 2.6%. Given that I have mainly used neighbour matching, the next lines are referred to that method in particular.

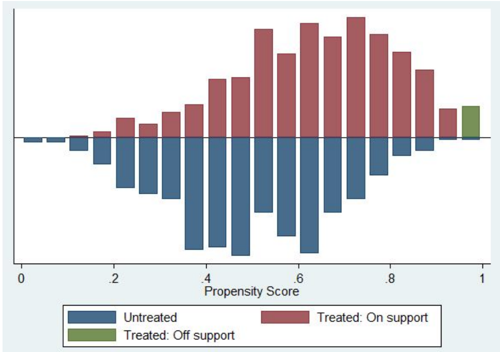
Figure E.2 a) shows the distribution of the propensity score across treatment and control group before and after matching. It is clear that individuals in the control group have lower propensity scores. It can also be noticed that there are some treated individuals with a propensity score of one or really close to one. This might be a problem, since the common support assumption establish that $P(D=1|X)<1$. In the matched chart, it is clear that it might be the case that some treated individuals are gonna have problems to find matches since no control observation scores that high. This is confirmed by Figure E.2 b). Treated individuals close to a propensity score of 1 are left out of the common support region. Those are the 16 observations mentioned before.

Next, in Table E.2, I present a brief description of these individuals that have been

Figure E.2: Propensity Score distribution of treatment and control group before and after matching



(a) Propensity Score distribution of treatment and control group before and after matching



(b) Propensity score and common support

left out of the main estimation to have an idea of who they are. On average, children left out of the common support region have literate parents, all of them obtained a bachelor degree or even masters/PhD degrees. Almost all of them live in urban areas and their average wealth index is almost twice as much as the one for on support children. They are also mostly males (gender takes value 1 for female). So, children dropped out of the estimation belong to some elite group, whose parents are well-educated, working withing business activities, public administration or education and living in urban areas. The exclusion should not pose any problem to the estimation.

Table E.2: Means of off/on-support individuals

	Age	Gender	Wealth Index	Father's Education
Off support	226.75	0.375	0.724	14.12
On support	225.96	0.506	0.399	4.11

	Father Literate	Mother's Education	Mother Literate	Rural/Urban
Off support	1	13	1	0.938
On support	0.381	3.51	0.209	0.289

Age measured in months. Education levels ranging from 1-14, covering Std.I-Std.XII plus higher education and vocational education. *Source*: Own elaboration from Young Lives data

F Cluster-Robust Variance Matrix

Starting with a simple linear model:

$$y_{ig} = x'_{ig}\beta + u_{ig} \tag{11}$$

where y_{ig} is the outcome variable for individual i in cluster g, x'_{ig} is a vector of covariates and u_{ig} is assumed to be correlated within clusters g but not across clusters and given that $E[y_{ig}|x_{ig}] = 0$.

Then, the cluster-robust variance matrix can be obtained from a model where all i observations are stacked in the g^{th} cluster (from equation 11):

$$y_g = X_g\beta + u_g, \qquad g = 1, \ldots, G \tag{12}$$

where there are N_g observations within cluster g.

Therefore, the OLS estimator is

$$\hat{\beta} = (X'X)^{-1}X'y = \left(\sum_{g=1}^{G} X'_g X_g\right)^{-1} \sum_{g=1}^{G} X'_g y_g \tag{13}$$

and the cluster-robust estimate of the variance matrix of the OLS estimator

$$\hat{V}_{clu}[\hat{\beta}] = (X'X)^{-1}\hat{\beta}_{clu}(X'X)^{-1} \tag{14}$$

where

$$\hat{\beta}_{clu} = \sum_{g=1}^{G} X'_g \hat{u}_g \hat{u}'_g X_g \tag{15}$$

and $\hat{u}_g = y_g - X_g\hat{\beta}$ is the vector of OLS residuals for the g^{th} cluster.

G Robustness Tables

Table G.3: ATT of preschool on dropout using 4 different propensity score matching methods in the crude model

		2-NN[a]	5-NN[a]	Radius[b]	Kernel[b]
ATT	Preschool	-0.045	-0.074**	-0.060	-0.036
		(0.0474)	(0.0387)	(0.0366)	(0.0564)
Balancing Tests	Mean Bias	6.1	4.9	2.0	4.9
	Pseudo-R^2	0.008	0.004	0.001	0.005
	R	1.40	1.22	1.21	1.12
	Obs	950	950	946	950

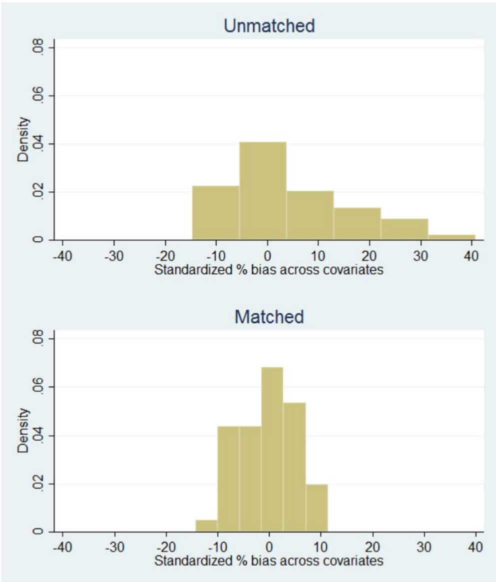
Significance level: $***p \leq 0.01$, $**p \leq 0.05$, $*p \leq 0.1$. Covariates: Gender, age, child's ethnic group and health problems Number of observations only show sample size after dropping off-support individuals.
[a] AI Std. errors in parentheses (Abadie and Imbens, 2006).
[b] Bootstrap Std. errors in parentheses calculated with 500 repetitions.

Table G.4: ATT of preschool on dropout using 4 different propensity score matching methods and including more variables than in the main model

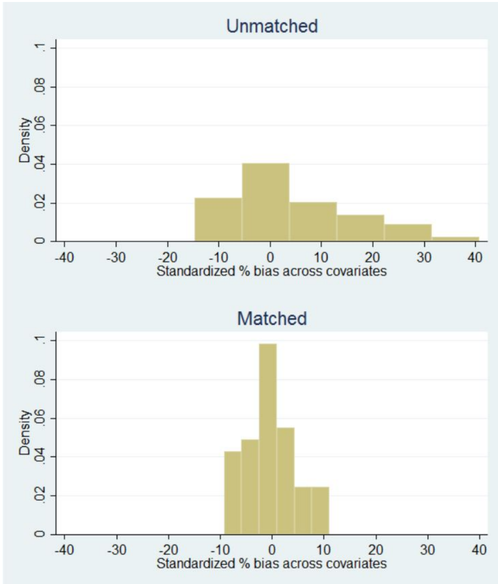
		2-NN[a]	5-NN[a]	Radius[b]	Kernel[b]
ATT	Preschool	-0.062	-0.057*	-0.076*	-0.057
		(0.0488)	(0.0436)	(0.0427)	(0.0527)
Balancing Tests	Mean Bias	5.4	3.6	4.1	6.1
	Pseudo-R^2	0.040	0.025	0.021	0.055
	R	1.07	1.42	1.09	1.05
	Obs	922	922	897	922

Significance level: $***p \leq 0.01$, $**p \leq 0.05$, $*p \leq 0.1$. Additional covariates: Mother's education, indicator whether father is alive and primary occupation of the household. Number of observations only show sample size after dropping off-support individuals. [a] AI Std. errors in parentheses (Abadie and Imbens, 2006). [b] Bootstrap Std. errors in parentheses calculated with 500 repetitions.

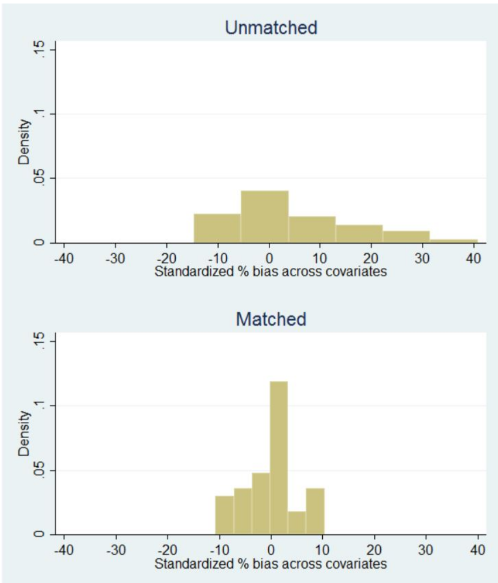
H Balancing Tests



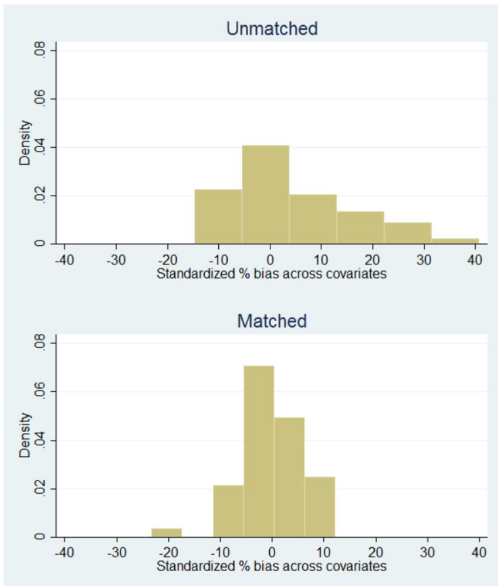
(a) 2-NN



(b) 5-NN



(c) Caliper



(d) Kernel

Source: Own elaboration from Young Lives Data