LUND UNIVERSITY
School of Economics and Management

Masters Programme in Economic Demography

Family size and short term educational attainment in Colombia

by

Juan Camilo Medina
ede15jme@student.lu.se

***Abstract**: Family related chracteristics are important determinants of individual outcomes. Also, fertility is a process closely related to development. As part of a body of literature investigating the effect of family size on schooling, this study develops an estimation using data for Colombia during 2010 to 2013. Building on exogenous variation in the number of siblings using multiple births and preferences over children's sex, this research exercise aims to provide some guidance and evidence about the relationship between children's quantity and quality.*

Key words: *Family size, fertility, education, quantity-quality trade-off, Colombia.*

EKHM51
Master's Thesis (15 credits ECTS)
June 2016
Supervisor: Björn Eriksson & Volha Lazuka
Examiner: Faustine Perrin
Word Count: 21287

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Research Problem

Colombia is currently experiencing a process of socioeconomic development characterized by high levels of inequality. Economic growth has been more dynamic than regional economic activity and the share of the population below the povery line has decreased from 34.1% in 2011 to 27.8% in 2015. Nonetheless, income distribution remained stable during this period; GINI measures rank Colombia as one of the 10 countries with worse income distribution (World Bank, 2016). More critical conditions related to armed conflict have had dramatic improvements during the past decades. Colombia had the highest homicide rate in the world, hosted the core of international drug industry and its state faced serious coercion from guerrilla and paramilitary groups. Heavy military intervention and a gradual increase in state presence generated an important effect upon governance and security. This also had significant implications over national and foreign investment, national debt, overall economic performance and social well-being (Robinson, 2013). Although strategic improvements in key areas of welfare such as health and education have taken place during the last decades, an important part of the population mostly associated with a rural environment remains relatively relegated from this development process (Parra-Peña, Ordóñez, & Acosta, 2013).

Although there is not a conclusive explanation describing the link between poverty and fertility, empirical studies for Latinamerica and Colombia typically reveal a positive association between poverty and higher levels of fertility (Martinez, 2013). In Colombia, as the conditions describing the socioeconomic gradient worsen, families tend to have a higher quantity of children. This becomes accentuated in rural environments. However, this gap has decreased over time; the difference between rural and urban Total Fertility Rate was 1,3 from 1987 to 1990 and 0,8 between 2007 and 2010 (Martinez, 2013). Also, the share of teenage pregnancies over the total amount of pregnancies has increased in recent years. Most of these women find themselves in difficult social conditions making pregnancy along with childbearing and human capital accumulation and labor market participation conflicting activities (Florez & Sánches, 2013). Finally, approximately 67% of all pregnancies in 2008 were unplanned (Prada, Singh, Remez, & Villarreal, 2011). The general characteristics of fertility are likely not to merely affect future mother's outcomes but are also likely to affect her children. Educational attainement in particular has important consequences over the course of an individual's life; higher educational attainment is associated with higher income and lower unemploymenht rates (Bureau of Labor Statistics, 2016). In Colombia individuals with higher education earn as nearly as six times as much as individuals with primary school (Pineda & Acosta, 2009).

Motivated by this situation, this study investigates the relationship between family size as the number of siblings and educational attainment in the contemporary Colombian context. Specifically, it addresses this issue by aiming to answer the following research question: what is the short term effect of the total number of siblings on educational attainment in Colombia?

Using cross-section and panel data for Colombia during the period between 2010 – 2013, different specifications employing exogenous variation in number of siblings based on multiple births and sex preferences over children are estimated. The time span encompassed in the study responds to the desire to develop an analysis that could describe the current scenario as closely as possible to the present and the availability of data that could provide the necessary information as well as some robust

1.2 Aim and Scope

Building on this research problem, the study aims to develop a quantitative research strategy using Colombian data from private and public sources based on the current literature's approach towards identification. Grounded on quasi-experimental methodologies, the study seeks to provide causal estimates and confront theoretical predictions about the relationship between family size and education. Specifically, two models based on multiple births and mixed sex preferences in children as instruments for number of siblings and a model including household level fixed effects are estimated. The working sample is dependent on complete information across key variables and a particular age range. Baseline estimations are grounded on this sample but the age restriction, which acts as a relatively arbitrary parameter, will explore different ranges in order to test the robustness of initial estimations. Finally, results are interpreted and related to previous findings and the research problem.

1.3 Outline of the Thesis

The study follows the following structure: Chapter 2 explores the theoretical proposals and observed findings that articulate the relationship between the size of the family and educational attainment or the relationship between quantity and quality of children (as it is commonly described) focusing on Gary Becker's contribution. The previous research section provides context and summarizes the literature concerned with the empirical estimations of this parameter in order to test the main theoretical expectations. This section also reviews the general methodological approach as it has been a key component in the evolution of this research topic. Chapter 3 describes the used data by commenting on its sources, purpose, limitations and weaknesses and the process by which variables were constructed and the final working sample was defined. Also, this section reflects upon representativity and possible selection processes that might by influencing the characteristics of the sample and therefore affect the validity of final estimates. The methodological aspects of the study are explored in chapter 4. Building on the controversies explored in previous research regarding identification strategies, instrument validity and specific variable related issues are discussed in this section. Also, a formal expression of the estimated models are provided. Chapter 5 presents and refects upon the results of these estimations and relates these findings to previous results, theoretical predictions and the stated research problem. Finally, chapter 6 summarizes the research outcomes and comments on the research agenda.

2 Theory

2.1 Theoretical Approach

One of the most insightful theoretical developments dealing with the relationship between the size of the family as number of children and their characteristics is Gary Becker's model explaining the demand for children and how this process is mediated by the parental investment in the quality of each child. He expanded the theoretical approach of neoclassical economics to the analysis of various social phenomenon by proposing rational choice models based on utility maximization. Becker developed his theory through various publication, the final version of his theory, as found in *A treatise of the Family* (Becker, 1973) published in 1973, which was continuously revised and adjusted through several publications, is the one exposed in the following pages. *Grosso modo*, Becker constructs a model of the demand for children around neoclassical consumer choice theory by maximizing utility subject to a budget constrain. He further develops the model by introducing quality and the framework of household production incorporating non-market goods through shadow prices and shadow income. The study will derive the main theoretical expectation of the relationship between family size and human capital accumulation of the children from this particular theory.

In this model each family maximizes a utility function[1] containing the quantity and quality of children and all other commodities which is consistent with completeness, transitivity and non-satiation. This implies that preferences over bundles are ordered, consistent and that 'more is better'. The utility function is described by:

$$U_f = U_f(n, q, Z_{1,} \dots, Z_m)$$

---

[1]This function must meet a positive first derivative with respect to each argument ($f'(U) > 0$) so that there is a positive marginal utility and must also display a negative second derivative ($f''(U) < 0$) for the function to represent diminishing marginal utility. This properties ensure that the tangency conditions are sufficient to guarantee a unique constrained optimization (Nicholson & Snyder, 2008).

Where

$U = Utility$

$f = family\ index$

$n = quantity$

$q = quality$

$Z = commodity$

$m = index\ of\ commodities$

All Z commodities are aggregated into a single Z commodity since children don't have substitutes (Becker, 1973). First, the demand for children is explained without regard for quality. This is consistent with the following utility function:

$$U_f = U_f(n, Z)$$

Also, families face a budget constrain equal to the unit cost of children multiplied by the amount of children and the total spending on all other commodities. Maximization implies that the totality of income I is exhausted; families allocate all of their resources to a combination of n and Z. This budget constrain can be described as:

$$p_n n + \pi_z Z = I$$

$I = full\ income$

$p_n = total\ cost\ of\ producing\ and\ rearing\ children$

$\pi_z = cost\ of\ aggregate\ commodity$

The optimization problem arising from this restriction is:

$$\max U_f(n, Z)\ \ s.t.\ \ p_n n + \pi_z Z \leq I$$

The First Order Conditions resulting from full income spending following Lagrange multipliers ($\mathcal{L} = U(n, Z) + \lambda[I - p_n n - \pi_z Z]$) are:

1. $\partial\mathcal{L}/\partial n = MU_n - \lambda p_n = 0$
2. $\partial\mathcal{L}/\partial Z = MU_Z - \lambda\pi_z = 0$
3. $\partial\mathcal{L}/\partial\lambda = I - p_n n - \pi_z Z = 0$

Where $\frac{\partial U(x)}{\partial x} \forall x \in (n, Z) = MU_x$

Solving for marginal utilizes in 2 and 3:

$$MU_n = \lambda p_n$$

$$MU_Z = \lambda\pi_z$$

Dividing the quantity's marginal utility by the aggregate commodities' marginal utility, we get the marginal rate of substitution (Nicholson & Snyder, 2008) where the slope of the indifference curve equals the slope of the budget constrain if the utility function is strictly quasi-concave:

$$\frac{MU_n}{MU_Z} = \frac{\cancel{\lambda} p_n}{\cancel{\lambda} \pi_z} = \frac{p_n}{\pi_z} = MRS_{n,Z}$$

From solving the maximization problem we conclude that the demand for children is a function of the relative price of children ($\frac{p_n}{\pi_z}$) and total income (I)[2].

$$n = n(\frac{p_n}{\pi_z}, I)$$

According to this solution, *ceteris paribus*, the demand for children decreases if the relative price of children increases or income decreases.

However, several factors can influence p_n. Becker states that rural fertility is higher compared to urban fertility because the costs of several goods and services such as housing and food are relatively smaller. Also, other variables like the profitability of child labor explained this difference. Nonetheless, advances in productivity and compulsory schooling laws have had an impact on the demand for children because the once attractive cost differential has decreased. Also, aid programs aimed towards helping mothers, especially single mothers who have several children, decrease the cost of rearing children. Likewise, changing opportunity costs arising from time allocation also affect the demand for children. This becomes particularly relevant in the face of increasing female labor market participation.

Becker argues that his framework applies to the desired number of children rather than the actual number of children and that families might face difficulties in attaining this desired quantity because of frictions in reproductive behavior such as sterile partners and unwanted children. Nonetheless, he claims that non-modern contraception can and has had important implications over reproductive behavior and that it alone can explain large changes in fertility.

---

[2] "A consequence of the assumption of constrained utility maximization is that the individual's optimal choices will depend implicitly on the parameters of his or her budget constraint. That is, the choices observed will be implicit functions of all prices and income. Utility will therefore also be an indirect function of these parameters" (Nicholson & Snyder, 2008)

Furthermore, the demand for children is also affected by income. Children and wealth have been historically positively correlated. Nevertheless, during the nineteenth century this relationship reversed. Becker claims that this happended because i) in lower income levels the quantity of children and income remains positively correlated (as in a malthusian dynamic) while in higher levels of income the relation is negative because of increasing direct and relative costs (particularly because of women's potential earnings). He believes that the main reason for this increase in rearing costs is the interaction between quantity and quality.

In order to explain the interaction between quantity and quality, Becker states that childre's quality has no substitutes. Hence, it can be introduced in utility functions just as it was initially specified. One important assumption is that children share the same level of quality which is produced by the family through market goods and time. Mathematically, quality is introduced as an element of the utility function and has a unit price of p_c with each children having a total level of quality q. Subsequently, the total amount of resources spent of each child is $p_c qn$. The approach in which quantity and quality interact is based on the notion of shadow prices[3].

As a part of a general household production model, suitable for the analysis of children production (a non-market production), beyond the problem of constrained optimization, the concepts of substitutes and complementary good are explored by considering interactions between produced goods following an input-output logic. The introduction of a quality component synthesizes the costs of rearing children and enables the model to provide conclusions about the interaction between quantity and quality. Considering the initial utility function

$$U_f = U_f(q, n, Z)$$

And the following budget constrain

$$p_c qn + \pi_z Z = I$$

We get the following FOC:

$$\max U_f(n, Z) \;\; s.t. \;\; p_c qn + \pi_z Z \leq I$$

$$\mathcal{L} = U(n, Z) + \lambda[I - p_c qn - \pi_z Z]$$

3 "A commodity's shadow price is the ratio at which a household can transform one commodity into another, or, more precisely, into the standard or numeraire commodity" (Pollak, 2002)

FOC:

1. $\partial \mathcal{L}/_{\partial n} = MU_n - \lambda p_c q = 0$
2. $\partial \mathcal{L}/_{\partial q} = MU_q - \lambda p_c n = 0$
3. $\partial \mathcal{L}/_{\partial Z} = MU_z - \lambda \pi_z = 0$

$$MU_n = \lambda p_c q$$

$$MU_q = \lambda p_c n$$

$$MU_Z = \lambda \pi_z$$

These marginal utilities help define the shadow prices for n and q ; $\pi_n = p_C q$ and $\pi_q = p_C n$. Shadow prices are not only a function of the price of quality but also of children in case of quality and of quality in the case of children. Through this reasoning, equilibrium values are determined by demand functions of shadow prices and shadow income R.

$$x = d_x(\pi_z, \pi_n, \pi_q, R) \forall x \in (n, q, Z)$$

Shadow income is then equal to the sum of shadow spending on each good

$$R = \pi_z Z + \pi_q q + \pi_n n$$

Nonetheless, there is only explicit interaction of n and q in the market price version of the restriction:

$$I + p_c qn = \pi_z Z + (p_c n)q + (p_c q)n$$

Ceteris paribus analysis of exogenous changes in some variable reveal the only preference structure in which the model can provide positive solutions to both quantity and quality:

"If p_c, π_z, and I were held constant, an exogenous increase in *n* would raise the shadow price of q, $\pi_q (= np_c)$, and thereby would reduce the demand for q. The reduction in q lowers the shadow price of *n* because it depends on *q*, which further increases the demand for *n.* But this raises π_q and lowers q still further, which lowers π_n and raises *n* still further, and so on. " (p. 146 (Becker, 1973))

According to shadow prices, the quality of the interaction on quantity and quality describes these arguments as substitutes. The degree or the intensity of the interaction between quantity and quality is determined by the elasticity of substitution; if they held a very close substitute relation the model would not have positive solutions. This interaction states that there is a negative relationship because of the substitute quality of the relationship. Because of this relation, theoretical expectations about the relation between number of siblings and short term educational attainment is that this relation is negative.

2.2 Previous Research

The broader links between fertility and schooling in the direction this study is concerned with have been stablished as part of overall fertility analysis and the demand for children. One of the earliest effort in this subject is found in the Malthusian proposal. Either by affecting nuptiality and fertility or mortality, population growth was strongly linked to fluctuations in the economic environment through real wages (Malthus, 1826). In particular, the relationship was described by a positive elasticity of fertility in relation to income. Following the first demographic transition, population growth was no longer strongly tied to income (Guinnane, 2011).

A large amount of literature in historical demography is concerned with explaning the critical shift in fertility through the analysis of "the impact of modernization on the demand for children." (Cleland & Wilson, 1987). In spite of several efforts, there is not a clear, unifying and empirically consistent explanation to this phenomena mainly because of the simultaneity of several processes and heterogeneous aspects of the general dynamics amongst populations (Guinnane, 2011). Nonetheless, one of the key approaches linking fertility and education is a consumer choice based proposal in which children's education influences the desired number of children through relative price changes. The relative price approach allows a systematic approach to the inclusion of different and apparently distant factors influencing fertility decisions: "Several significant changes in the relevant period plausibly altered the costs of and returns to children in ways that would reduce fertility. These include housing costs due to urbanization, changes in child-labor law, increases in the opportunity costs of childbearing because of better labor-force opportunities for women, the introduction of free or compulsory primary education, and the development of social-insurance systems." p. 21 (Guinnane, 2011).

The specific link between children's education and their parent's fertility imposes an empirical research challenge because the scarcity of proper historical data obscures the estimation of returns to schooling and children's quality does not follow a straight forward interpretation without a clear idea of the consequences of education. Additionally, although schooling-encouraging policies are consistent with changes in relative prices in the direction of lower fertility, the opportunity cost in schooling versus working became less relevant because of child-labor laws (Guinnane, 2011). Nonetheless, one key contribution to the development of growth theories explaining the transition from a malthusian society to a development process grounded upon modern economic growth articulates human capital accumulation through parent investment in children's quality. Galor and Weil's *Unified Growth Theory* builds on human capital accumulation and technological progress to explain how a steady-state economy gravitated towards a regime of sustained economic growth joined by previously unseen population growth. Speciffically, they claim that the slow pace of technological development in the malthusian epoch did not provide enough incentives for parents to reallocate resources away from habitual expenditure in order to invest in children's quality (Galor, 2005). As this pace increased, income restrictions upon quality investment eased. With technology demanding human capital qualifications and alleviating consumption, a virtuous-circle dynamic involving technological innovation and human capital accumulation developed. One of the building blocks of this proposal is parental choice over children's quantity and quality; this section of the theory is approached though Becker's contribution.

Most of the developing world began a rapid decline in fertility after the second world war. In the decade of 1950 developing countries displayed high levels of fertility and the vast majority have experienced very important reductions (some reaching replacement levels) with steady declines associated to economic growth (Bongaarts, 2008). In relationship to schooling, this process is characterized by a trend reversal in human capital accumulation associated with economic development; a study analysing the relationship of fertility, human capital and development in 48 developing countries finds that previous to a significant decline in fertility, children in large families displayed high levels of schooling and subsequently, mostly due to the effects of economic growth, small families had more educated children (Vogl, 2016).

During the period between 1950 and the present, countries in Latinamerica and the Caribbean experienced a significant decline in fertility from a TFR of 5,9 in 1950 - 1955 to a current near-replacement level (Saad, 2009). In 2000, this region had mean years of schooling equal to the world average and Colombia had the lowest measure when compared to its neighboring countries (Giménez, 2005). Figure 1 plots mean schooling for individuals in age 25+ and TFR for the period between 1990 and 2010 in Colombia. Since 1995, fertility has decreased by almost 1 children and mean schooling has consistently increased from 5,5 mean years of schooling to 7,3. In 2010 mean schooling for the population in ages 5 and 18 was 6,42 and in 2013 it had a slight increase to 6,49[4].

[4] Based on the National Quality of Life Survey (ENCV – Encuesta Nacional de Calidad de Vida). Autor's calculations

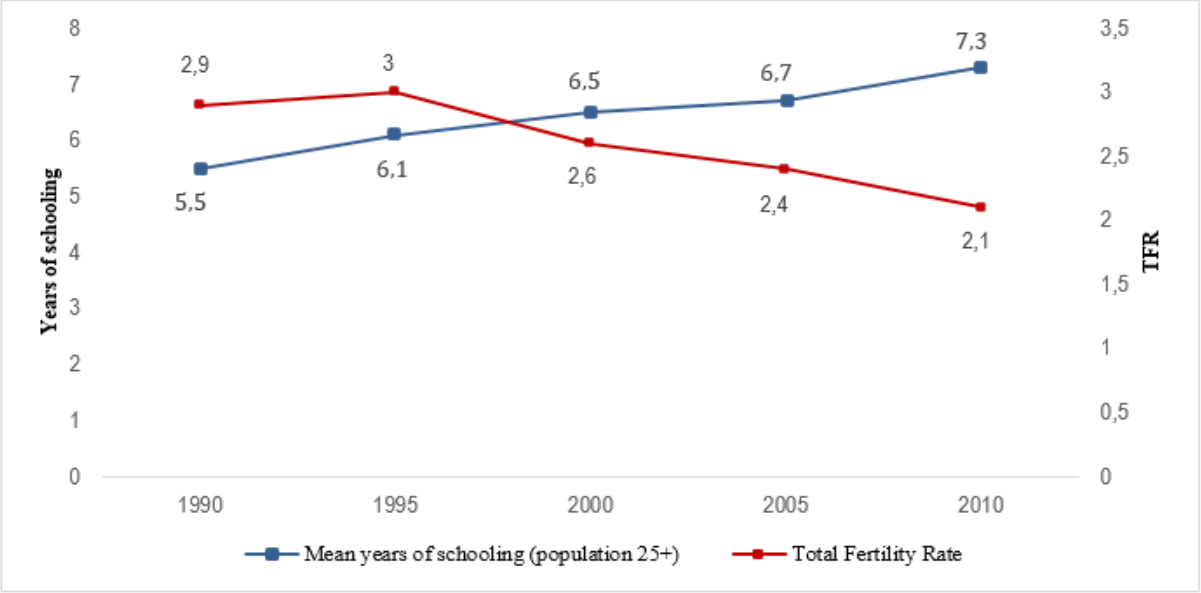

Figure 1 - Schooling and feritlity in Colombia (1990 - 2010) based on *(Flórez & Sánchez, 2013)*

The current state of fertility in Colombia partially displays signs of a Secondary Demographic Transition according to the conditions postulated by Ron Lesthaegh (Lesthaeghe, 2010). These are i) sustained fertility leveles below replacement linked to postponement and ii) an increase in the age at first marriage and the prevalence of informal arrangements. However, this is only the case for specific segments of the population associated to relatively high levels of income and education. There are important and persisten regional and socioeconomic divergences in fertility within the country. In particular across rural and urban environments (Flórez & Sánchez, 2013).

As far as this study can tell, there are no analyses of the effect of the family size on the educational attainment in the Colombian context. Nonetheless, this link has been studied in several different settings mostly using individual level data. The following paragraphs synthesize the evolution of this body of literature.

A study from 1974 by Leibowitz analyzing the determinants of ability, schooling and income for high IQ individuals in California found that, based on OLS estimates, male educational achievement, measured as years of schooling, at age 29 in 1940 is negatively related (-0,118) to the number of siblings and significant at ten percent level of statistical significance. For the same individuals the coefficient decreases by half (-0,058) ten years later (Leibowitz, 1974). The author argues that larger families tend to delay final levels of education because of competition for household resources. He also concludes that final levels of education is independent of family size. Nonetheless, as it is stated in the study, this sample is highly unrepresentative of the overall population. Another study from 1981 (Blake, 1981) employing several surveys and basic control strategies finds similar though larger effects of sibling size over years of schooling for men and, particularly, women. She concludes that as the final level of schooling increases the effect of sibling size decreases. Other study measuring the effect of the number of children over more direct measures of quality like reading comprehension and vocabulary tests also find negative associations (Hauser & Sewell, 1985). Several authors claim that until the employment of quasi-experimental methodologies this type of findings where

standard not only in educational and quality related children outcomes but also in mother's earnings and labor market participation (Schultz, 2005) (Angrist, Lavy, & Schlosser, 2010).

The main difficulty in assessing these previous results as causal lies on the high probability of an omitted variable bias (Angrist, Lavy, & Schlosser, 2010). This issue raises several questions that don't seem to be addressed in these previous studies. In particular, as Black et al. 2005 notice "is it true that having a larger family has a causal effect on the "quality" (in our case IQ) of the children? Or is it the case that families who choose to have more children are (inherently) have less children in the first place?" Such questions imply that educational attainment might not be affected so much by the size of the family but by other factors correlating with it. Moreover, family size is, to a reasonable extent, a decision. The nature of this variable raises concern about endogeneity in OLS estimations because of its relation to unobserved factors associated to the error term.

According to Angrist, research on women's earnings (in which "hundreds of empirical studies" p. 450 (Angrist J. D., 1998) reported the usual important negative associations) provided the firsts alternatives to evaluate the effect of the number of children through exogenous variation. Rosenzweig and Wolpin were amongst the first to use a quasi-experimental design in order to evaluate these effects by instrumenting family size using multiple births. This type of instrumentation proved relevant to the analysis of schooling because it provides a feasible exogenous variation since the allocation of multiple birth seems to be as good as random. The use of empirical strategies based on exogenous variation of the number of children has mixed evidence but is consistent in reporting very different estimates when compared to previous studies repoting negative, important and significant estimates. Relatively recent studies ((Caceres, 2006) (Angrist, Lavy, & Schlosser, 2010), (Booth & Joo Kee, 2009), (Black, Devereux, & Salvanes, 2005), (Marteleto, 2012)) have focused on exogenous variation of family size mainly through multiple births as an instrument of the number of children which is said to provide exogenous variation. Another usual source of exogenous variability has been uniform sex composition. The logic behind this instrument is based on the idea that parents who have a uniform sex composition in their two first births and have strong preferences for mixed sex composition will experience an important exogenous incentive to have a third child.

The results for this type of research are heterogeneous. Black et al. (Black, Devereux, & Salvanes, 2005) report large negative and significant OLS effects for sibship over education in Norway even when controlling for birth order. Second stage results for families having twins in the second and third birth become positive and insignificant. Angrist et al. (Angrist, Lavy, & Schlosser, 2010) using Israeli census data find that the sibship size is statically significant and negatively related to highest grade completed in OLS estimations. 2SLS coefficients using twin births and several versions of uniform sex composition are positive and not statistically significant. Finally, using data for adolescents in Brazil, Marteleto finds similar effects of family size on years of complete education; large negative and significant coefficients for OLS and positive but small and insignificant 2SLS estimates (Marteleto, 2012).

3 Data

In order to develop a quantitative analysis investigating the effect of family size on short term educational attainment in the Colombian context, data has to meet several characteristics. The main features of a dataset that would provide the necessary information to estimate this effect are i) individual level data about kinship relations to other members of the household (this allows for a precise definition of a nuclear family which is valuable given the recurrent extended family structures of Colombian households) and ii) information about schooling with as much variation as possible. The exogenous variation estimations employed in the analysis also require some specific information on age or the date of the birth and sex of children. Additional variables relevant to educational attainment are also fairly common information found in household datasets such as income or if the household is located in a rural or urban environment. Additionally, information on kinship is also very useful since parent's characteristics are very likely to influence children's schooling and this enables this characteristics to be traced amongst observations through parental links.

Nevertheless, the absence of an ideal system of information based on register data that offers data on individual characteristics not only in Colombia but in most developing countries has been filled with different surveys with all sorts of different objectives. Furthermore, in the Colombian context there is only one current data recollection aiming at creating a longitudinal dataset gathering information on households and individuals suitable for socioeconomic and demographic research. This is *Longitudinal Survey of Wealth, Income, Labor and Land* (Encuesta Longitudinal Colombiana de la Universidad de los Andes) abbreviated in Spanish as ELCA. However, there are relatively abundant cross-section surveys with similar purposes developed by different institutions such as the Demographic and Health Survey (DHS) and the different surveys developed by the National Administrative Department of Statistics (DANE – Departamento Administrativo Nacional de Estadística). The DHS dataset focuses on women's reproductive behavior and lacks information about schooling and other characteristic that are important for this analysis. A continuous attempt to gather information about the Colombian population reggarding living standards at household and individual level by the name of National Quality of Life Survey (ENCV – Encuesta Nacional de Calidad de Vida) has been operating since 1997 for every year since 2010, has served as input for hundreds of social studies in recent year. This survey was a product of the Living Standards Measurement Study (LSMS) described as "… a household survey program housed within the Surveys & Methods Unit of the World Bank's Development Research Group that provides technical assistance to national statistical offices (NSOs) in the design and implementation of multi-topic household surveys" (World Bank, 2016).

Both the ELCA and the ENCV provide the formation needed for required estimations and they encompass a common time span and both were used in the analysis. The ELCA began in 2010

and plans to follow over 10,800 households and its members over 10 years. So far this has translated into a first wave in 2010 and a second wave in 2013. As stated, the ENCV has had a continuous operation since 2010 so there is datasets available for the years 2010, 2011, 2012 and 2013. The empirical analysis developed in the upcoming chapters is based on the information provided by these two surveys for the 2010 - 2013 period and aims to construct estimates both on cross-section data and, to the extent in which is possible, longitudinal data.

3.1 Source Material

The following pages describe the data used in the analysis and how the final samples where defined. However, the defined research problem naturally limits the study to a certain base-line age group defined as individuals between 5 and 18 years of age because this interval encompasses the ages in which children and teenagers normally are part of the basic education system either through public or private schooling (Ministry of National Education, 2014). This strongly restricts the overall size of the sample to an important fraction of the complete datasets. Furthermore, in order to have the most comparable estimates across datasets, the sample in each data was restricted to observations that could be used in every estimation. This means that individuals who might be suitable for bivariate models could not be part of the final sample because they present missing information on variables used in other models. For robustness purposes, the conditions defining the baseline sample (consistency across estimations and 5 – 18 age range) will become more flexible.

Estimations are based upon 7 samples. Cross-section samples correspond to the 2010, 2011, 2012 and 2013 ENCV datasets and the first wave (2010) of the ELCA dataset. Samples employing longitudinal information correspond to i) the combination of the 2010 and 2013 ELCA datasets and ii) a subsample of this information based on an specific characteristic of the individual's siblings. Specifically, one panel sample consists of individuals for which their number of siblings did not change between 2010 and 2013 and a second sample pools these individuals along with the rest. This dataset is referred as encompassing a constant sibling sample (or CSS when space was scare).

3.1.1 ELCA

As discussed, the ELCA survey is a response to an informational vacuum regarding data that could study the dynamics of certain social processes across time with an emphasis on poverty. Specifically, this survey is aimed at providing an input to the study of poverty issues so that the design of public policy on household income dynamics, asset accumulation and young individual's human capital accumulation is based on stronger evidence. The secondary objectives of the survey are providing information on labor market conditions, early childhood, cognitive development and rural environment amongst other (Universidad de los Andes, 2011). Following these objectives, the dataset is based on a multilevel structure which provides household and community level information for all individuals. Moreover, a set of variables interrelating the household members by kinship and offspring allow the identification of nuclear

families, parents and siblings. Each household member has an individual identifier within the household. The dataset reports, if it's the case, the identifier for father or mother allowing sibling identification through common parents. Also, schooling information is available through several questions enquiring for last approved grade, coursing grade, completed education level (pre-school, primary, secondary, technical, technological, undergraduate and graduate) and coursing educational level. This mitigates the likeliness of an individual with absolutely no schooling information. The type of household networking provided by this survey and the questions examining broad measures of schooling are based on the design of the ENCV.

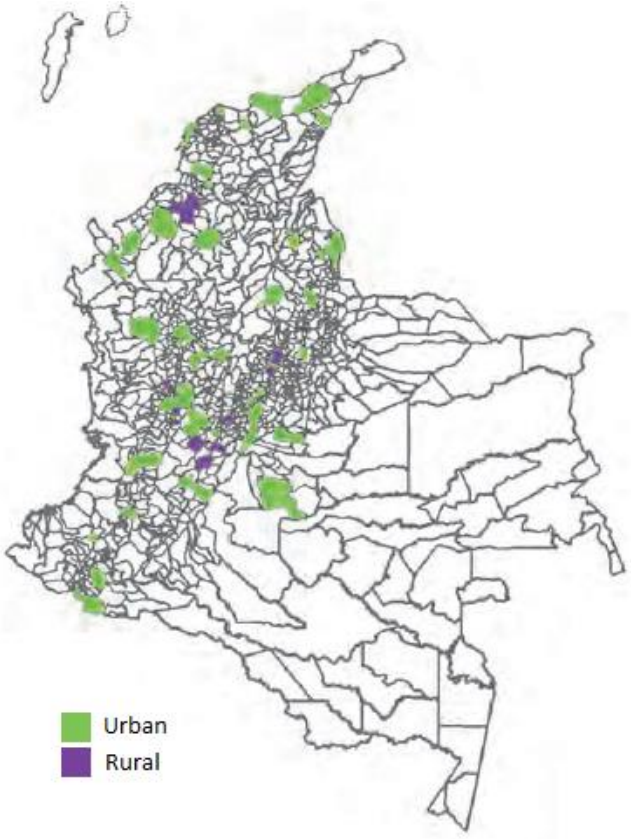

Figure 2 - Municipalities part of ELCA Source: *(Universidad de los Andes, 2011)*

The survey is based upon a 10,800 households that in 2010 contained 44,255 individuals divided in 6,000 urban and 4,800 rural households. The initial sample had representativity at urban level and the rural households represented the population of small production agricultural poor households in 4 micro-regions. Figure 2 identifies all the municipalities which are part of the survey. Areas in south and western regions of the country are largely underpopulated, and don't present key urban concentrations.

The data is fragmented into 8 different dataset for each wave divided by either rural or urban environment. For each environment there is a dataset for person level information registering all relevant individual characteristics, a household level dataset, a household expenditure dataset and a community level dataset. Information was linked across datasets first appending the rural and urban files for each one of the 4 cases described above since they have the exact same structure. Next, the household level datasets were merged with the individual level dataset by household identifier so that each observation for an individual also displays household's

characteristics. Community level information was also linked by a community identifier present in one of the household level dataset. This process resulted in two databases, one for each wave, which were appended since they share the same variable structure. Identification across waves was possible through labeling each household member and a unique household identifier that remain constant over time.

Although the survey is aimed at following households and individuals belonging to the initial households set, the second wave reports no observations for 10,543 individuals that were present in the 2010 wave. Additionally, the second wave reports new 5,760 individuals (26,78% because they were born). Individuals leaving the sample have been classified into two categories presumed to follow different dynamics; either individuals leave the survey '*on their own*' or they leave the survey as a consequence of the household leaving the survey. Individuals belonging to the first category belong to a *household with partial attrition* because a only a portion of the individuals belonging to the household are absent in 2013. Individuals belonging to the second group belong to *households with total attrition* because in that case every member of the household is absent. Around half (50,9%) of these 10,543 individuals absent in the 2013 wave belong to households of the second group.

Even though it is hard to define the specific causes of attrition, there are some associated factors that might provide information useful in assessing what selection processes are active at the moment of defining a sample observed in both periods. The first factor that has an apparent important influence on attrition at is age. Figure 3 shows the age distribution for the three populations by attrition status: i) individuals observed in both periods, ii) individuals not observed in the second wave belonging to a household with partial attrition and iii) individuals not observed in the second wave belonging to a household with total attrition. Age distribution

is similar for individuals belonging to the first and third group. Individuals who left the survey 'on their own' tend to be highly concentrated around 19 years of age (Figure 3).

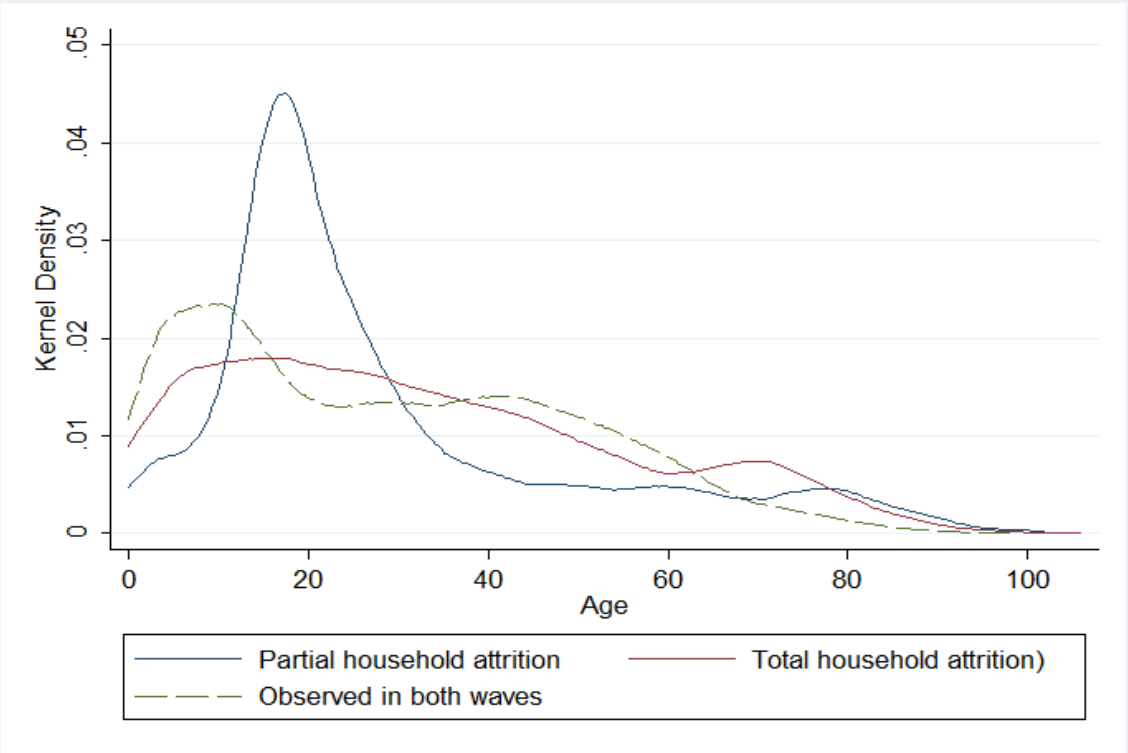

Figure 3 - Age distribution by attrition status

While individuals belonging to households with total household attrition might leave the survey because of household level dynamics, those in group *ii* are likely to do so because of individual level processes. This age group is associated with several events that are likely to influence the permanence of individuals in their hosting households; at age majority (18) individuals are likely to leave their households because of a very different set of motives all being very recurrent in young Colombians. These include pursing higher education in other geographical setting, initiating their own family, desire of independence, engagement in the labor market or military service. Individuals in these ages who are observed in both periods might display certain characteristics that could be negatively and positively correlated with schooling. In other words, individuals leaving their home probably have strong reasons to pursue activities incompatible with remaining at their initial household. A sample including individuals beyond 18 years might trigger a selection process in the first wave because of survival bias. If individuals are more likely to leave their households around age 20 because of factors negatively correlated with schooling, the set of people in that age group present in the survey might display higher average levels of schooling biasing the effect of family size. According to theoretical expectations, if the effect of family size on scholing is negative, the selection bias will produce a diminished estimation.

Table 1 (page 18) displays descriptive statistics for the three baseline samples based on the ELCA datasets. The first sample (*ELCA 2010*) corresponds to the set of individuals with complete information in all listed variables in the 2010 wave. Here *schooling* represents the data for years of complete schooling in 2010. So is the case for *ELCA 2013* with *schooling* representing the data for years of complete schooling in 2013. Those individuals part of this

last sample who did not experience any change in the number of siblings across waves constitute the definitive sample referred as 'ELCA 2013 constant sibling sample'.

Most variables follow similar descriptive statistics and don't display significant differences across samples except for schooling which is time-dependen. A particular difference lies in the share of the sample for which the first two births in a nuclear family shared sex in the case of the sample for individuals with a constant amount of siblings across waves. Although this is more formally tested in the upcoming sections, this may be explained because households with uniform sex composition in births one and two could be more prone to having a third child. If this is the case, these families are more likely not to have additional children since they have already had third child (almost 1 unit above the TFR: which was estimated in 2010 as 2.1 children per woman (Flórez & Sánchez, 2013)) and therefore keep a constant number of children between the observed periods. The average individual is 11 years old, has five complete years of schooling in 2010 and around eight in 2013 and has two siblings. Its parents are around age 40, have completed primary school and, though there is a wide variance in income, the typical household earns 95% of a minimum wage ($515.000 COP in 2010 (Central Bank of Colombia (Banco de la República), 2016)) per month. Pndividuals in this sample are then quite likely to belong to a poor household and face important budget restrictions.

Table 1 - Descriptive statistics (ELCA)

	ELCA 2010					ELCA 2013					ELCA 2013 (Constant sibling sample)				
	N	**Mean**	**SD**	**Min**	**Max**	**N**	**Mean**	**SD**	**Min**	**Max**	**N**	**Mean**	**SD**	**Min**	**Max**
Age	7,638	11.05	3.706	5	18	7,759	11.14	3.520	5	18	4,291	11.16	3.480	5	18
Age (sq.)	7,638	135.8	84.39	25	324	7,759	136.6	80.48	25	324	4,291	136.7	79.51	25	324
Aggregate sibling exposure	7,638	15.27	14.19	0	103.0	7,759	15.78	14.29	0	103.0	4,291	15.63	11.32	0.0833	78.50
Birth order	7,638	2.029	1.182	1	9	7,759	2.055	1.194	1	10	4,291	1.996	1.018	1	8
Birth order 2 or 3	7,638	0.475	0.499	0	1	7,759	0.480	0.500	0	1	4,291	0.554	0.497	0	1
Birth order 4 or more	7,638	0.0403	0.197	0	1	7,759	0.0429	0.203	0	1	4,291	0.0235	0.152	0	1
Father's Age	7,638	42.45	9.304	20.33	91	7,759	42.62	9.156	20.33	91	4,291	41.86	8.326	22.42	90.92
Father's education	7,638	6.706	4.276	0	18	7,759	6.607	4.248	0	18	4,291	6.754	4.244	0	18
Firstborn	7,638	0.417	0.493	0	1	7,759	0.407	0.491	0	1	4,291	0.367	0.482	0	1
Girl	7,638	0.483	0.500	0	1	7,759	0.484	0.500	0	1	4,291	0.486	0.500	0	1
Household income (log)	7,638	13.14	1.010	4.585	16.52	7,759	13.12	0.991	4.585	16.52	4,291	13.12	1.003	4.585	16.52
Mother's Age	7,638	37.48	7.723	18	78	7,759	37.60	7.592	12.67	78	4,291	36.99	7.041	12.67	60.67
Mother's education	7,638	7.394	4.129	0	18	7,759	7.323	4.095	0	18	4,291	7.535	4.015	0	18
Number of siblings	7,638	1.931	1.530	0	9	7,759	1.982	1.554	0	9	4,291	2.059	1.198	1	8
Same-sex composition (births 1 & 2)	7,638	0.281	0.449	0	1	7,759	0.281	0.450	0	1	4,291	0.367	0.482	0	1
Schooling	7,638	5.215	2.893	0	16	7,759	8.283	2.870	1	17	4,291	8.377	2.868	1	17
Twin	7,638	0.0156	0.124	0	1	7,759	0.0157	0.124	0	1	4,291	0.0182	0.134	0	1
Twin instr.	7,638	0.00262	0.0511	0	1	7,759	0.00219	0.0468	0	1	4,291	0.00210	0.0458	0	1
Uniform sex composition instr.	7,638	0.358	0.479	0	1	7,759	0.359	0.480	0	1	4,291	0.450	0.498	0	1
Urban environment	7,638	0.499	0.500	0	1	7,759	0.481	0.500	0	1	4,291	0.496	0.500	0	1

3.1.2 ENCV

The ENCV is a cross-section survey developed in accordance to the World Bank's methodology for measuring living standards implemented in 1993, 1997 and continuously from 2010 to 2015. The survey's purpose is to construct a wider characterization of Colombian household's living standards beyond a measure of income. With this in mind, it presents modules for education, labor market, housing, household expenditure, utilities, health and other more specific subjects. Its general objective is to obtain information that allows the analysis of socioeconomic conditions of Colombian households which permit monitoring of variables key to the development and implementation of public policy and the Millennium Development Goals (National Administrative Department of Statistics (DANE), 2014). The modules needed to obtain the information consistent with the research objectives are *household characteristics and composition*, *education* and *labor force*. Each dataset was merged using an individual unique identifier based on the dataset structure, as described in the previous section, consisting of the household code and the numbering of each individual within the household. The sample in each survey is representative at national level.

The total number of observations correspond to 53,453 individuals distributed in 14,268 households in 2010. The average household contains 4,8 individuals and the share of the sample living in a rural environment is close to 26%. In 2011 the sample size considerably increased to 92,188 observations and, in 2012 and 2013, 74,172 and 73,155.

Table 2- Final ENCV sample relative size

	Final sample	Eligible (Population aged 5 -18)	Total observations	Final sample as a share of total individuals aged 4-18	Final sample as a share of total observations
2010	4169	14215	53453	29.33%	7.80%
2011	6723	25088	92188	26.80%	7.29%
2012	5392	19226	74172	28.05%	7.27%
2013	5053	18443	73155	27.40%	6.91%

Additionally, the relative size of the sample with respect to the total amount of observations (Table 2) fluctuates in each year reaching a maximum in 2010 and a minimum in the following year. The final sample size relative to the sample ages 5 to 18 is somewhat more stable. Decreasing relative measures of the final sample indicates an increase in aggregate missing information. Even though there is a significant observation loss because of missing information, final sample size in each edition resembles the overall size of the datasets in absolute numbers.

Average household size had a decreasing trend in every year; 4,65 in 2011 to 4,43 and 4,34 in 2012 and 2013. Table 3 (page 21) displays descriptive statistics for the four ENCV final samples restricted by the 5 to 18 age interval and non-missing information in each one of the listed variables. The average individual displays a somewhat constant age but has slightly higher

education following a positive trend since 2011. Its parents are, on average, similar; while mean age does not present important changes, average complete years of schooling tend to increase since 2011. Regarding aggregated household variables, income increased from 1,04 to 1,23 minimum wages. This represents a general improvement in purchasing power; in combination with a decreasing number of siblings, average absolute real income per household member increases therefore lessening the degree of competition for household resources. In the context of Becker's theory, quality depends positively on income. This is consistent with aggregate measures. The family is also more likely to dwell in an urban environment.

Table 3 –Descriptive statistics (ENCV)

	ENCV 2010					ENCV 2011					ENCV 2012					ENCV 2013				
	N	**Mean**	**SD**	**Min**	**Max**	**N**	**Mean**	**SD**	**Min**	**Max**	**N**	**Mean**	**SD**	**Min**	**Max**	**N**	**Mean**	**SD**	**Min**	**Max**
Age	4,169	11.47	3.676	5	18	6,723	11.40	3.666	5	18	5,392	11.51	3.713	5	18	5,053	11.58	3.684	5	18
Age (sq.)	4,169	145.0	85.48	25	324	6,723	143.3	85.09	25	324	5,392	146.2	86.41	25	324	5,053	147.7	86.10	25	324
Aggregate sibling exposure	4,169	14.57	13.30	0	109.4	6,723	15.38	13.88	0	104.5	5,392	14.51	13.44	0	95.92	5,053	13.75	13.10	0	93.58
Birth order	4,169	1.973	1.102	1	8	6,723	2.006	1.139	1	10	5,392	1.959	1.095	1	9	5,053	1.906	1.042	1	9
Birth order 2 or 3	4,169	0.496	0.500	0	1	6,723	0.497	0.500	0	1	5,392	0.488	0.500	0	1	5,053	0.491	0.500	0	1
Birth order 4 or more	4,169	0.0317	0.175	0	1	6,723	0.0353	0.184	0	1	5,392	0.0287	0.167	0	1	5,053	0.0247	0.155	0	1
Father's Age	4,169	43.11	9.224	20.42	83.75	6,723	42.62	8.645	20.50	86.58	5,392	42.67	8.793	20	79.08	5,053	42.69	8.981	18.50	93.75
Father's education	4,169	7.825	4.553	0	18	6,723	7.737	4.567	0	18	5,392	7.862	4.489	0	18	5,053	8.122	4.430	0	18
Firstborn	4,169	0.414	0.493	0	1	6,723	0.406	0.491	0	1	5,392	0.422	0.494	0	1	5,053	0.432	0.495	0	1
Girl	4,169	0.486	0.500	0	1	6,723	0.487	0.500	0	1	5,392	0.483	0.500	0	1	5,053	0.474	0.499	0	1
Household income (log)	4,169	13.19	1.704	1.792	17.03	6,723	13.38	1.041	2.565	16.71	5,392	13.42	1.038	4.654	16.45	5,053	13.50	1.084	4.644	16.68
Mother's Age	4,169	38.59	8.210	20.08	94.75	6,723	38.39	7.602	19.17	72.58	5,392	38.36	7.735	9.583	66.50	5,053	38.57	7.839	18.58	65.33
Mother's education	4,169	8.303	4.370	0	18	6,723	8.291	4.566	0	18	5,392	8.491	4.315	0	18	5,053	8.822	4.401	0	18
Number of siblings	4,169	1.780	1.376	0	8	6,723	1.896	1.441	0	9	5,392	1.748	1.367	0	8	5,053	1.631	1.296	0	8
Same-sex composition (births 1 & 2)	4,169	0.318	0.466	0	1	6,723	0.310	0.463	0	1	5,392	0.289	0.453	0	1	5,053	0.298	0.457	0	1
Schooling	4,169	5.965	3.168	0	17	6,723	5.917	3.141	0	17	5,392	6.089	3.168	0	17	5,053	6.189	3.131	0	17
Twin	4,169	0.0134	0.115	0	1	6,723	0.0173	0.130	0	1	5,392	0.0137	0.116	0	1	5,053	0.0158	0.125	0	1
Twin instr.	4,169	0.00768	0.0873	0	1	6,723	0.00476	0.0688	0	1	5,392	0.00297	0.0544	0	1	5,053	0.00376	0.0612	0	1
Uniform sex composition instr.	4,169	0.410	0.492	0	1	6,723	0.389	0.487	0	1	5,392	0.368	0.482	0	1	5,053	0.379	0.485	0	1
Urban environment	4,169	0.701	0.458	0	1	6,723	0.688	0.463	0	1	5,392	0.702	0.457	0	1	5,053	0.728	0.445	0	1

3.1.3 External validity

Due to the important difference in observations between the eligible set of individuals in ages between 5 and 18 who have some missing information and the actual working sample, it is relevant to check for any structural differences between these groups. Table 4 shows the difference in means for each variable in each dataset between mean values for those in the final sample and mean values of those within the considered age interval who, due to missing information in one or several variables, could not be included in the estimations. Because of the consistency of very significant mean difference between these two groups it is not probable that these populations are comparable and factors leading to missing information might be correlated with observables. In particular, the gaps in *number of siblings* are relatively large and significant at 1% level of statistical significant. Even though these surveys are representative of the national population, this differences heavily undermine the possibility of results having external validity. Hence, the results and conclusions derived from the quantitative analysis are more likely to be restricted to the working sample and efforts were aimed towards internal validity.

Table 4 - Difference in means by participation in final sample (Excluded - included)

Variable	ELCA 2010	ELCA 2013	ELCA 2013 (CSS)	ENCV 2010	ENCV 2010	ENCV 2010	ENCV 2010
Age	1.468***	1.254***	.449***	.258***	.201***	.16***	.132**
Age (sq.)	32.767***	31.658***	12.989***	6.571***	5.351***	4.062***	3.556**
Aggregate sibling exposure	-2.273***	-3.724***	.2530	-3.083***	-2.629***	-3.503***	-2.785***
Birth order	-.356***	-.435***	-.136***	-.257***	-.224***	-.278***	-.226***
Birth order 2 or 3	-.131***	-.147***	-.084***	-.123***	-.121***	-.12***	-.125***
Birth order 4 or more	-.018***	-.025***	-.0010	-.01***	-.007***	-.009***	-.005**
Father's Age	1.06***	.569***	.4960	.2120	.432***	.325*	.402**
Father's education	-.32***	-.0310	-.2980	-1.603***	-1.571***	-1.211***	-1.145***
Firstborn	.176***	.207***	.087***	.148***	.138***	.152***	.142***
Girl	-.0040	-.0080	-.0220	-.0020	.0020	.0110	.018**
Household income (log)	.036*	.106***	.0220	-.161***	-.187***	-.152***	-.172***
Mother's Age	.3**	-.030	-.3950	-.525***	-.957***	-1.123***	-1.189***
Mother's education	-.739***	-.549***	-.829***	-.936***	-1.316***	-.791***	-.892***
Number of siblings	-.364***	-.511***	.088*	-.341***	-.268***	-.36***	-.263***
Same-sex composition (births 1 & 2)	-.024***	-.026***	.0310	-.064***	-.053***	-.026***	-.035***
Schooling	-.317***	-.3***	-.0240	.0440	-.092**	-.0340	-.050
Twin	.0010	.0010	0	.0010	-.003*	.0020	.0010
Twin instr.	.002*	.003***	0	-.004***	-.001*	.002**	00
Uniform sex composition instru.	-.064***	-.068***	-.0250	-.099***	-.085***	-.056***	-.068***

3.1.4 Variable construction

Age / Age (squared) – In both surveys age was calculated as the sum of a yearly part and a monthly par. Both were found in the original datasets. The monthly part of age was expressed as number of months and was divided by twelve so that the final variable for age was expressed in year units. *Age squared* was constructed by elevating the value in the variable *Age* to the power of two.

Number of siblings – As it was previously commented, both datasets (ELCA and ENCV) provide information about offspring relationships within the household. This information is contained in variables displaying the household identifier for father and mother. Table 5 displays a hypothetical example employing the structure used in the datasets. In this example,individuals with identifiers 1 and 2 are the father and mother of individuals labeled as 3, 4 and 5 . If an individual shares at least one parent with another member of the household it becomes part of the set of individuals related as siblings. *Number of siblings* is the sum of individuals that meet this condition.

Table 5 - Household, personal and kinship identifier (example)

House hold identifier	Individual identifier in the household	Father identifier	Mother identifier	Number of siblings
1	1	.	.	
1	2	.	.	
1	3	1	2	2
1	4	1	2	2
1	5	1	2	2

Aggregate sibling exposure – Once siblings are identified, it is possible to calculate the amount of time that an individual has been exposed to the different amount of siblings that birth sequence implies and for how long has an individual been exposed to its siblings. For example, if a particular household has three children aged five, three and one, the exposure for the child aged five is two years exposed to cero siblings, two years to one sibling and one year to two siblings. The child aged 3 has been exposed to one sibling for two years and to two siblings for one year. Finally, the child aged one has been exposed to two siblings for one year. Following the example in Table 5, Table 6 displays the hypothetical sibling exposure scheme described above.

Table 6 - Sibling exposure (example)

Individual identifier in the household	Father identifier	Mother identifier	Number of siblings	Age	Exposure to 0 siblings	Exposure to 1 sibling	Exposure to 2 siblings	Aggregate sibling exposure	Weighted sibling exposure
1	.	.		35	.	.	.		
2	.	.		30	.	.	.		
3	1	2	2	5	2	2	1	3	4
4	1	2	2	3	0	2	1	3	4
5	1	2	2	1	0	0	1	1	2

Birth order – Building on age and the criteria for sibling identifier, birth order can be constructed by ordering the set of siblings around age. The relative position of the individual in this rank will provide its value for birth order.

Birth order 2 or 3 / Birth order 4 or more– Once birth order has been constructed, a dummy variable taking the value of 1 if the birth order was either 2 or 3 and 0 if the birth order variable had any other value was generated to differentiate the effect of this position in the birth rank. So was the case if the birth order was 4 or more (although this variable was mostly used as a reference category).

Father's Age / Mother's age – Based on father and mother indicator, it is also possible to 'import' the information from that observation. Although variables containing the information for mother's and father's age were present in the dataset before any variable construction, this information was only used if the 'imporing' mechanism found no information because it was considered more reliable and had less missing values.

Father's education / Mother's education – The exact same information conditions for father and mother's age describe the ones for father and mother education so the same mechanism described above was used.

Firstborn – Using birth order, a dummy variable taking the value of 1 if the individual was a firstborn and 0 if otherwise was used to indicate a first born condition.

Girl – The variable containing the information for sex was recoded so that 1 represents that the individual is a female and 0 if otherwise.

Household income (log) – By aggregating the income for all the individuals in one household and taking the logarithm of this quantity, a measure of household income was constructed.

Urban environment – If the household is part of an urban environment, then a dummy variable takes the value of 1 and 0 if otherwise.

Schooling - Schooling is defined as the number of complete years within the formal education system encompassing 12 grades ranging from transition grade (a grade articulating pre-school and formal education) until the last high-school grade, technical, technological, undergraduate and graduate education. Although the initial age restriction is likely to have an effect over the maximum levels of schooling, a measure of education was constructued for everyone in the data set. The basic strategy to interpret education related data was to prioritize precision. The process involved several categories and educational levels. Because of extension and pertinence details are provided in Appendix B.

Twin – Individuals are defined as twins if they were born to the same mother according to mother identifier and were born in the same date (or have the same age). This definition is robust for multiple births resuling in more than two children.

Same-sex composition (births 1 & 2) – If the first two births in a sibling set share sex then a dummy for same-sex composition of births 1 and 2 takes the value of 1 and 0 if this condition is not met.

4 Methods

4.1 The Approach

Empirical research analyzing the relationship between family size and educational attainment and its conclusions are widely divided by methodological approach. In accordance to conventional theoretical predictions, early quantitative studies estimating this relation found relatively large negative associations. Later, when exogenous variation methodologies were applied, the large evidence supporting the predictions of the neoclassical model became questioned.

The present empirical strategy is based in the now standard approach the literature investigating the relationship between schooling and number of sibling has developed. This strategy resembles the development in the application of methodological tools this research has had since the decade of 1970 but also serves a practical purpose particularly relevant in estimations aiming to provide causal effects using instrumental variables. Following Angrist et al. (Angrist, Lavy, & Schlosser, 2010), Black et al. (Black, Devereux, & Salvanes, 2005) and Marteleto et al. (Marteleto, 2012), this type of analysis compares results from Ordinary Least Squares and Two Stage Least Squares for each instrument in different samples according to specific research questions.

Also, birth order effects have increasingly gained importance in recent studies. Growing concerns about estimations reporting a number of siblings effect not disentangeling the birth order effect have motivated models assessing the specific effect of this variable on educational attainment. Black, Devereux, & Salvanes (2005) estimate a household fixed-effects model to analyze the role of birth order. The present study approaches this issue by i) including birth order and ii) interpreting the number of siblings effect through an impact intensity or exposure perspective. Altough the number of siblings is extremely relevant, specially when considering Becker's proposal about children's fixed costs and economies of scale having an effect on the distribution of parental resource investment, a degree of exposure to the effect can be insightful. Individuals can have the exact same number of siblings but, depending on spacing, the time they have been exposed to these circusmstances may vary.

In a hypothetical scenario, lets imagine two households having extremely similiar socioeconomic conditions but different preferences over spacing. Both households have a ten year old first born and two other children. In the first household one of these is a 1 year old and the other is 2 years old and in the second household 5 and 8. Both households have the same number of children and every children has the same number of siblings. Nonetheless, the amount of time each first born has been exposed to the effect of two siblings is significantly different in each case. Part of this effect is captured by birth order and number of siblings.

However, a great part of the variation in the variable of interest or the intensity of the number of siblings can still be rescued from the information needed for the standard analysis. Considering this issue, estimation are extended to calculate the effect of the aggregate or total sibling exposure. An additional analysis controlling for household fixed effects preserving exogenous variation is estimated. Although estimation through instrumental variables provides causal estimates based on variation that is as good as random, controlling for other environment characteristics might produce more precise estimates. Because household level unobserved dynamics might be responsible for processes influencing schooling (i.e. genetic or parent-behavior factors related to ablity), an additional specification controlling for household fixed effects is estimated for aggregate sibling exposure in order to capture internal variation in the effect of siblings. This cannot be done when the causal variable of interest is number of siblings because everits a fixed factor.

Finally, because of the rareness of multiple births, estimations employing twin births as a source of exogenous variation become highly dependent on a small amount of observations. This motivated a robustness analysis based on flexible sample entry parameters. Because of possible survivor bias the results of this analysis are likely to present a downward bias.

The general form of the basic control strategy through an OLS estimation is based on the model:

$$Y_i = \beta_0 + \beta_1 X_{i,j} + A_i'\gamma + \varepsilon_i$$

Where, for every individual i, Y_i represents schooling, $X_{i,j}$ the causal variable of interest for every variable j such that $j \in (Number\ of\ siblings, Aggregate\ siblings\ exposure)$, A_i' a set of covariates and ε_i an error term.

2SLS models estimate the following first stage for the multiple birth instrument (Z_i)[5]:

$$X_{i,j} = \pi_0 + \pi_1 Z_i + A_i'\gamma + v_i$$

A family level first stage for the uniform sex composition of births 1 and 2 (U) for every family f, where TC is a dummy variable taking the value of 1 if a family in the set of families who had

[5] The model for first stage is based on: $Familiy\ Size = \alpha_0 + \alpha_1 TWIN + X\alpha_2 + v$ (Black, Devereux, & Salvanes, 2005) and $c_i = X_i'\beta + \alpha t_{2i} + \eta_i$ (Angrist, Lavy, & Schlosser, 2010)

either 2 or 3 children had a third child and 0 if otherwise and ϕ represents a cumulative normal distribution function is estimated through the following probit model:

$$\Pr\left(TC_f = 1\right) = \phi(\eta + \theta U_f)$$

Predicted values for the causal variable of interest ($\widehat{X_\iota}$) for every instrument $I_{i,k}$ where $k \in Z, U$ where defined as

$$\widehat{X_{\iota,J}} = \widehat{\pi_1} I_{i,k} + A_i' \hat{\gamma}$$

The second stage is defined as

$$Y_i = \alpha_0 + \alpha_1 \hat{X}_{i,j} + A_i' \delta + \xi_i$$

Finally, the fixed effects model through within estimation was specified as:

$$\widetilde{Y_{\iota,h}} = \lambda_0 + \lambda_1 \widetilde{\hat{X}_{\iota,J,h}} + \widetilde{A_{\iota,h}'\rho} + \widetilde{\xi_{\iota,h}}$$

$$\widetilde{\omega_\iota} = \omega_{i,h} - \overline{\omega_\iota} \; \forall \; \omega \in (Y, \hat{X}, A', \xi)$$

4.2 Instrument validity

4.2.1 Multiple birth instrument

The twin instrument, which acts as a source of exogenous variation in the number of siblings, is affecting not only twins but every child in the family as well. As mentioned previously, the last estimate for Total Fertility Rate in Colombia was 2.1 births per woman. This makes it is more likely that a twin birth will exogenously affect the number of siblings in families that already have at least one child. Also, it is important to consider that multiple births followed by additional births cannot be used as an instrument because this situation suggests that the target number of children had not been completed because of the multiple birth. This might be an exogenous variation in the timing of births but not in the number of children or number of siblings.

Although it is also possible that this event would exogenously affect families aiming at higher quantities of total children, the degree of likeliness of this situation decreases as the total amount of children becomes larger. For example, if a family has five children and the sixth birth is a twin birth, the assumption that their desired total amount of children is six (since it could also be four) becomes less realistic as this situation gradually differs from the average situation implied in the TFR. A real strong exogenous variation is present if, because of this event, the real number of children becomes larger that the desired or targeted number of siblings. Likewise, a twin birth in the first birth might not exogenously alter the number of children because, based on the TFR and descriptive statistics, it is more probable that the targeted

number of children is more than one. Following these concerns, it is more probable that the multiple birth exogenously affects the number of children and the number of siblings is when it occurs in the second or third birth (descriptive statistics suggest that individuals have, on average, an average number of siblings larger than 1). Moreover, the instrument represents being part of a family in which this type of multiple birth has occurred. Table 7 provides an example.

Table 7 - Twin instrument example

House hold identifier	Individual identifier in the household	Father identifier	Mother identifier	Age	Number of siblings	Twin	Twin instrument
1	1	.	.	35		.	.
1	2	.	.	30		.	.
1	3	1	2	5	2	0	1
1	4	1	2	3	2	1	1
1	5	1	2	3	2	1	1

Black et al. (Black, Devereux, & Salvanes, 2005), based on Rosenwein and Wolpin (Rosenzweig & Wolpin, 1980) define their twin instrument as a dummy indicating in the individual was part of a multiple birth arguing that multiple birth are "unplanned and therefore exogenous variation in family size". I argue that there is real exogenous variation when the multiple birth contributes to a total number of children beyond the target number of children which, although unobserved, is likely to be close either two or three as average conditions from descriptive statistics and the TFR suggest.

Angrist et al (Angrist, Lavy, & Schlosser, 2010) define their instrument as "The variable t_{2i} (which we call twins2) indicates multiple second births (…)." Black et al. define their instrument variable according to the following statement: "The *TWIN* indicator is equal to 1 if the *n*th birth is a multiple birth and equal to 0 if the *n*th birth is a singleton. We restrict the sample to families with at least *n* births and study the outcomes of children born before the *n*th birth. In practice, we estimate the specification for values of *n* between 2 and 4. By restricting the sample to families with at least *n* births, we make sure that, on average, preferences over family size are the same in the families with twins at the *n*th birth and those with singleton births." p .13 (Black, Devereux, & Salvanes, 2005). Based on this last approach, the multiple birth instrument is define as a dummy taking the value of 1 if in a family that has three children the second birth was multiple or if in a family of four children the third birth was multiple. The fulfilling of a exclusionary restriction is based on the fact that the educational system has no restrictions affecting children born in a multiple birth and therefore all the effect over schooling should be associated to family size.

4.2.2 Uniform sex composition instrument

Following a similar discussion to the one justifying the specifications of the twin instrument, same-sex composition or uniform sex composition of births number one and two acts as an instrument affecting the probability of having a third child based on the assumption of preference for a set of children with mixed sex. This variation is argued to be exogenous because

parents cannot control the sex of their children and have fixed sex preferences. This event works as an instrument in families willing to engage in a trade-off between their target number of children (only across totals of 2 and 3) and mixed sex preferences. I interpret that families with more than 3 children have a revealed preference for more than 3 children and are therefore not motivated by this event to have a third child because they initially already desired a third or more children. An example of how data was structured around this instrument is provided in

Table 8. Similarly, as this event is extrememly common there are no institutional restrictions to individuals part of this type of uniform sex composition.

House hold identifier	Individual identifier in the household	Father identifier	Mother identifier	Age	Number of siblings	Sex	Same sex composition of births 1 and 2	Uniform sex composition of births 1 and 2 instrument
1	1	.	.	35		Male	.	.
1	2	.	.	30		Female	.	.
1	3	1	2	5	2	Female	0	1
1	4	1	2	3	2	Female	1	1
1	5	1	2	3	2	Male	1	1

Table 8 - Example for uniform sex composition of births 1 and 2

5 Empirical Analysis

5.1 Results

According to mean schooling by number of siblings (Table 9) individuals who have two siblings display higher average measures. Table 9 displays the average years of schooling for each number of siblings for each sample identified in the a first row. Although there in no control strategy in this analysis, individuals with quite high numbers of siblings have higher levels of schooling than individuals with no siblings. Nonetheless, according to broad measures of fertility and family composition it is likely that individuals reporting cero silings are very young individuals in families that have not completed the fertility process. Additionally, very large families might display different producer/consumer profiles: older children might shift from a net consumer function to a net producer function. The results for the models described in the previous section are displayed in the following pages and a synthesis of these is provided in the first appendix.

Table 9 - Mean schooling by number of siblings

Number of siblings	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010	ELCA 2013	ELCA 2013 (CSS)	Average
0	5.9	5.6	5.7	5.9	5	8.1	-	6.033
1	6.1	6.1	6.2	6.2	5.3	8.5	8.4	6.686
2	6.1	6.1	6.3	6.5	5.4	8.5	8.5	6.771
3	5.7	5.7	6.1	6.1	5.1	8.1	8.2	6.429
4	5.7	5.5	5.6	6	5.3	8.1	8.4	6.371
5	5.2	5.2	5.5	5.7	4.8	7.8	8.4	6.086
6	4.9	5	5.5	5.4	4.6	7.4	7	5.686
7	4.7	4.1	5.9	6.8	4.7	7.1	6.5	5.686
8	4.5	5.4	6.4	7.2	4.8	6.6	4.5	5.629
9	-	6.4	-	-	5.4	7.8	-	6.533

Tables 10 to 33 present different specifications of models explaining schooling. The first row indicates to which sample the estimation corresponds to. Variables involved in the model are listed to the left of the tables. Corresponding coefficients, the standard errors and the statisticial significance according to the criteria found at the bottom of the tables are presented. Also, the number of observations and a fit measure are displayed. Throughtout various specificiations the relationship between family size as number of siblings and schooling as years of schooling is tested. Both theoretical expectations based on the literatured explored in section 2 suggest that these variables display a negative relation; negative coefficients for the *number of siblings* variable are consistent with this consideration and support these claims; positive coefficients

don't support such claims. A second approach to the problem is to measure the net exposure to siblings as defined previously instead of number of siblings. Results for this estimation are presented just as described above as well.

5.1.1 OLS ESTIMATIONS

Table 10 and Table 11 report OLS estimations for bivariate regressions. For *number of siblings* highly significant coefficients display negative and relatively large relationships. Less significant coefficients are closer to cero and the ENCV 2013 sample reports a small but positive estimation. Coefficients reggarding *aggregate sibling exposure* exhibit less divergence and high statistical significance although they all have positive sign.

Table 10 - Bivariate OLS (Number of siblings)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Number of siblings	-0.144***	-0.131***	-0.0262	0.0102	-0.0370*	-0.122***	-0.168***
	(0.0356)	(0.0265)	(0.0316)	(0.0340)	(0.0216)	(0.0209)	(0.0365)
Constant	6.220***	6.167***	6.135***	6.172***	5.286***	8.525***	8.724***
	(0.0801)	(0.0632)	(0.0700)	(0.0708)	(0.0533)	(0.0527)	(0.0869)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291
Adjusted R-squared	0.004	0.003	-0.000	-0.000	0.000	0.004	0.005

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

Table 11 - Bivariate OLS (Aggregate sibling exposure)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Aggregate sibling exposure	0.0297***	0.0285***	0.0385***	0.0403***	0.0309***	0.0250***	0.0392***
	(0.00366)	(0.00274)	(0.00317)	(0.00332)	(0.00231)	(0.00226)	(0.00382)
Constant	5.532***	5.479***	5.531***	5.635***	4.743***	7.889***	7.765***
	(0.0722)	(0.0568)	(0.0626)	(0.0630)	(0.0481)	(0.0482)	(0.0737)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291
Adjusted R-squared	0.015	0.016	0.027	0.028	0.023	0.015	0.024

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

Subsequent OLS estimations (Table 12 to Table 14) address different sets of covariates; multivariate specifications are estimated with different birth order specifications and the inclusion of education related subsidies (*Familias en acción*). All three models report negative, relatively large and highly statistical coefficients for *number of siblings* with the exception of the ELCA 2010 sample that reports a small and less significant effect. These estimations are consistent with the main theoretical prediction: a negative relationship between family size as number of siblings and schooling.

Table 12 - Multivariate OLS 1 (Number of siblings)

	ENCV	ENCV	ENCV	ENCV	ELCA 2010	ELCA 2013	ELCA 2013

	2010	2011	2012	2013	Outcome	Outcome	Outcome (CSS)
Number of siblings	-0.125***	-0.139***	-0.107***	-0.103***	-0.0313**	-0.100***	-0.128***
	(0.0176)	(0.0131)	(0.0151)	(0.0156)	(0.0141)	(0.0117)	(0.0198)
Age	0.151***	0.237***	0.205***	0.0701*	0.238***	1.688***	1.681***
	(0.0427)	(0.0331)	(0.0362)	(0.0362)	(0.0378)	(0.0321)	(0.0422)
Age (sq.)	0.0260***	0.0221***	0.0238***	0.0298***	0.0172***	-0.0448***	-0.0438***
	(0.00183)	(0.00142)	(0.00155)	(0.00154)	(0.00166)	(0.00140)	(0.00185)
Girl	0.372***	0.284***	0.339***	0.308***	0.277***	0.510***	0.573***
	(0.0452)	(0.0353)	(0.0388)	(0.0380)	(0.0395)	(0.0334)	(0.0438)
Father's Age	0.00411	0.00378	0.00325	0.00241	-0.00735**	0.00406*	0.00274
	(0.00341)	(0.00297)	(0.00327)	(0.00313)	(0.00288)	(0.00245)	(0.00351)
Mother's Age	0.00501	0.00805**	0.00453	0.00264	0.00822**	0.00491	0.00325
	(0.00397)	(0.00344)	(0.00382)	(0.00369)	(0.00365)	(0.00309)	(0.00435)
Father's education	0.0499***	0.0379***	0.0433***	0.0331***	0.0227***	0.0441***	0.0423***
	(0.00652)	(0.00496)	(0.00555)	(0.00545)	(0.00599)	(0.00506)	(0.00682)
Mother's education	0.0487***	0.0612***	0.0511***	0.0439***	0.0602***	0.0782***	0.0667***
	(0.00677)	(0.00494)	(0.00585)	(0.00558)	(0.00641)	(0.00544)	(0.00722)
Household income (log)	0.0155	0.0589***	0.0643***	0.0575***	-0.0440*	-0.0143	-0.0115
	(0.0137)	(0.0183)	(0.0201)	(0.0182)	(0.0250)	(0.0217)	(0.0291)
Urban environment	0.351***	0.102**	0.0373	0.153***	-0.153***	-0.164***	-0.159***
	(0.0556)	(0.0431)	(0.0480)	(0.0488)	(0.0490)	(0.0418)	(0.0551)
Familias en accion	0.122	0.194***	0.286**	0.204**	-0.0129	0.0332	-0.0285
	(0.0987)	(0.0665)	(0.116)	(0.0953)	(0.0448)	(0.0377)	(0.0497)
Constant	-1.114***	-1.967***	-1.715***	-0.755**	0.233	-5.435***	-5.191***
	(0.307)	(0.300)	(0.336)	(0.319)	(0.384)	(0.333)	(0.449)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291
Adjusted R-squared	0.790	0.789	0.798	0.815	0.645	0.739	0.752

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

Table 13 - Multivariate OLS 2 (Number of siblings)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Number of siblings	-0.186***	-0.202***	-0.144***	-0.146***	-0.0448**	-0.144***	-0.155***
	(0.0246)	(0.0188)	(0.0221)	(0.0229)	(0.0204)	(0.0168)	(0.0246)
Age	0.165***	0.250***	0.209***	0.0737**	0.242***	1.701***	1.690***
	(0.0428)	(0.0332)	(0.0362)	(0.0362)	(0.0380)	(0.0323)	(0.0425)
Age (sq.)	0.0259***	0.0221***	0.0240***	0.0299***	0.0172***	-0.0449***	-0.0439***
	(0.00183)	(0.00142)	(0.00155)	(0.00154)	(0.00166)	(0.00140)	(0.00185)
Girl	0.371***	0.285***	0.338***	0.308***	0.276***	0.508***	0.572***
	(0.0451)	(0.0352)	(0.0388)	(0.0379)	(0.0395)	(0.0333)	(0.0438)
Father's Age	0.00313	0.00198	0.00258	0.00179	-0.00759***	0.00330	0.00207
	(0.00342)	(0.00299)	(0.00328)	(0.00314)	(0.00289)	(0.00245)	(0.00353)
Mother's Age	0.000345	0.00311	0.00164	-0.000565	0.00693*	0.000574	9.68e-05
	(0.00417)	(0.00360)	(0.00402)	(0.00390)	(0.00391)	(0.00331)	(0.00468)
Father's education	0.0501***	0.0385***	0.0436***	0.0334***	0.0228***	0.0444***	0.0423***
	(0.00651)	(0.00495)	(0.00555)	(0.00545)	(0.00599)	(0.00506)	(0.00682)
Mother's education	0.0495***	0.0618***	0.0514***	0.0443***	0.0605***	0.0794***	0.0680***
	(0.00677)	(0.00494)	(0.00585)	(0.00558)	(0.00642)	(0.00544)	(0.00726)
Household income (log)	0.0152	0.0587***	0.0646***	0.0579***	-0.0455*	-0.0192	-0.0139
	(0.0137)	(0.0183)	(0.0201)	(0.0182)	(0.0250)	(0.0218)	(0.0291)
Urban environment	0.344***	0.0924**	0.0296	0.146***	-0.153***	-0.165***	-0.163***
	(0.0555)	(0.0431)	(0.0481)	(0.0488)	(0.0490)	(0.0417)	(0.0552)
Familias en accion	0.117	0.194***	0.285**	0.206**	-0.0126	0.0344	-0.0270
	(0.0986)	(0.0664)	(0.116)	(0.0953)	(0.0448)	(0.0377)	(0.0497)
Birth order	0.114***	0.118***	0.0674**	0.0765**	0.0258	0.0846***	0.0572*
	(0.0322)	(0.0254)	(0.0291)	(0.0298)	(0.0281)	(0.0234)	(0.0316)
Constant	-1.153***	-1.971***	-1.708***	-0.753**	0.245	-5.398***	-5.177***
	(0.307)	(0.300)	(0.336)	(0.319)	(0.384)	(0.333)	(0.449)

Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291
Adjusted R-squared	0.790	0.790	0.799	0.815	0.645	0.739	0.752

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

Table 14 - Multivariate OLS 3 (Number of siblings)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Number of siblings	-0.173***	-0.173***	-0.123***	-0.123***	-0.0401**	-0.130***	-0.149***
	(0.0227)	(0.0170)	(0.0200)	(0.0208)	(0.0186)	(0.0153)	(0.0233)
Age	0.166***	0.245***	0.207***	0.0714**	0.238***	1.696***	1.689***
	(0.0429)	(0.0332)	(0.0362)	(0.0363)	(0.0380)	(0.0323)	(0.0424)
Age (sq.)	0.0257***	0.0220***	0.0239***	0.0299***	0.0174***	-0.0448***	-0.0438***
	(0.00183)	(0.00142)	(0.00155)	(0.00154)	(0.00166)	(0.00140)	(0.00185)
Girl	0.372***	0.284***	0.339***	0.308***	0.275***	0.509***	0.573***
	(0.0451)	(0.0352)	(0.0388)	(0.0380)	(0.0395)	(0.0333)	(0.0438)
Father's Age	0.00335	0.00291	0.00303	0.00209	-0.00769***	0.00343	0.00201
	(0.00342)	(0.00298)	(0.00328)	(0.00313)	(0.00288)	(0.00245)	(0.00353)
Mother's Age	0.00144	0.00558	0.00357	0.000832	0.00663*	0.00159	0.000253
	(0.00413)	(0.00356)	(0.00397)	(0.00385)	(0.00384)	(0.00326)	(0.00462)
Father's education	0.0500***	0.0382***	0.0434***	0.0333***	0.0226***	0.0442***	0.0424***
	(0.00651)	(0.00495)	(0.00555)	(0.00545)	(0.00599)	(0.00506)	(0.00682)
Mother's education	0.0491***	0.0615***	0.0512***	0.0442***	0.0603***	0.0791***	0.0679***
	(0.00677)	(0.00494)	(0.00585)	(0.00558)	(0.00641)	(0.00544)	(0.00725)
Household income (log)	0.0153	0.0590***	0.0644***	0.0582***	-0.0441*	-0.0165	-0.0132
	(0.0137)	(0.0183)	(0.0201)	(0.0182)	(0.0250)	(0.0218)	(0.0291)
Urban environment	0.349***	0.0970**	0.0344	0.148***	-0.155***	-0.166***	-0.163***
	(0.0555)	(0.0431)	(0.0481)	(0.0489)	(0.0490)	(0.0417)	(0.0552)
Familias en accion	0.115	0.194***	0.283**	0.204**	-0.0124	0.0341	-0.0265
	(0.0986)	(0.0665)	(0.116)	(0.0953)	(0.0448)	(0.0377)	(0.0497)
Firstborn	-0.404***	-0.295***	-0.134	-0.137	-0.0426	-0.240***	-0.197*
	(0.117)	(0.0903)	(0.102)	(0.106)	(0.0995)	(0.0828)	(0.114)
Birth order 2 or 3	-0.328***	-0.252***	-0.129	-0.0587	0.0708	-0.124*	-0.112
	(0.0994)	(0.0757)	(0.0851)	(0.0891)	(0.0816)	(0.0677)	(0.0986)
Constant	-0.663**	-1.613***	-1.551***	-0.591*	0.295	-5.127***	-4.944***
	(0.334)	(0.319)	(0.360)	(0.345)	(0.404)	(0.349)	(0.471)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291
Adjusted R-squared	0.790	0.789	0.798	0.815	0.646	0.739	0.752

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

Multivariate estimations for the effect of aggregate sibling exposure on schooling (Table 15, Table 16 and Table 17) are always negative at 1% level of statistical significance. The effect is particularly weak in the ELCA 2010 sample and quite strong in the ENCV samples for 2010 and 2011. The second model, in which birth order is introduced as a continuous variable, reports particularly high coefficients for aggregate sibling exposure. These results suggest that as exposure to siblings increases, even when controlling for age and the rest of listed covariates, there is an associated decrease in schooling.

Table 15 - Multivariate OLS 1 (Aggregate sibling exposure)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Aggregate sibling exposure	-0.0150***	-0.0161***	-0.0111***	-0.0112***	-0.00407***	-0.0102***	-0.0121***
	(0.00188)	(0.00142)	(0.00160)	(0.00160)	(0.00157)	(0.00132)	(0.00218)
Age	0.169***	0.255***	0.218***	0.0825**	0.245***	1.702***	1.697***

	(0.0428)	(0.0332)	(0.0364)	(0.0363)	(0.0380)	(0.0323)	(0.0424)
Age (sq.)	0.0256***	0.0217***	0.0235***	0.0294***	0.0171***	-0.0451***	-0.0443***
	(0.00183)	(0.00142)	(0.00155)	(0.00154)	(0.00166)	(0.00141)	(0.00186)
Girl	0.375***	0.283***	0.341***	0.308***	0.276***	0.511***	0.577***
	(0.0451)	(0.0352)	(0.0388)	(0.0379)	(0.0395)	(0.0334)	(0.0438)
Father's Age	0.00469	0.00511*	0.00352	0.00267	-0.00714**	0.00438*	0.00330
	(0.00341)	(0.00297)	(0.00327)	(0.00313)	(0.00288)	(0.00245)	(0.00353)
Mother's Age	0.00761*	0.0112***	0.00712*	0.00527	0.00912**	0.00705**	0.00664
	(0.00396)	(0.00345)	(0.00384)	(0.00370)	(0.00368)	(0.00312)	(0.00441)
Father's education	0.0495***	0.0382***	0.0437***	0.0333***	0.0226***	0.0445***	0.0427***
	(0.00650)	(0.00494)	(0.00555)	(0.00544)	(0.00599)	(0.00507)	(0.00683)
Mother's education	0.0489***	0.0614***	0.0520***	0.0442***	0.0600***	0.0794***	0.0681***
	(0.00674)	(0.00493)	(0.00583)	(0.00556)	(0.00639)	(0.00543)	(0.00723)
Household income (log)	0.0171	0.0622***	0.0670***	0.0601***	-0.0423*	-0.0115	-0.00847
	(0.0136)	(0.0183)	(0.0201)	(0.0182)	(0.0250)	(0.0218)	(0.0292)
Urban environment	0.353***	0.116***	0.0476	0.161***	-0.152***	-0.156***	-0.146***
	(0.0555)	(0.0431)	(0.0480)	(0.0487)	(0.0490)	(0.0418)	(0.0552)
Familias en accion	0.125	0.195***	0.284**	0.203**	-0.0135	0.0188	-0.0502
	(0.0986)	(0.0664)	(0.116)	(0.0953)	(0.0446)	(0.0376)	(0.0494)
Constant	-1.415***	-2.369***	-2.007***	-1.021***	0.125	-5.716***	-5.581***
	(0.305)	(0.297)	(0.334)	(0.317)	(0.386)	(0.335)	(0.449)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291
Adjusted R-squared	0.790	0.790	0.798	0.815	0.646	0.739	0.751

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

Table 16 - Multivariate OLS 2 (Aggregate sibling exposure)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Aggregate sibling exposure	-0.0321***	-0.0344***	-0.0213***	-0.0260***	-0.00866***	-0.0197***	-0.0207***
	(0.00316)	(0.00250)	(0.00294)	(0.00307)	(0.00275)	(0.00232)	(0.00349)
Age	0.222***	0.310***	0.240***	0.110***	0.261***	1.737***	1.731***
	(0.0433)	(0.0335)	(0.0367)	(0.0366)	(0.0388)	(0.0330)	(0.0437)
Age (sq.)	0.0249***	0.0211***	0.0234***	0.0294***	0.0168***	-0.0457***	-0.0448***
	(0.00182)	(0.00141)	(0.00155)	(0.00154)	(0.00166)	(0.00141)	(0.00186)
Girl	0.378***	0.282***	0.340***	0.307***	0.275***	0.507***	0.575***
	(0.0449)	(0.0350)	(0.0388)	(0.0378)	(0.0395)	(0.0333)	(0.0438)
Father's Age	0.00322	0.00257	0.00247	0.00141	-0.00754***	0.00360	0.00244
	(0.00340)	(0.00297)	(0.00328)	(0.00313)	(0.00289)	(0.00245)	(0.00353)
Mother's Age	-0.000128	0.00323	0.00301	-0.000532	0.00666*	0.00194	0.00206
	(0.00410)	(0.00354)	(0.00396)	(0.00383)	(0.00387)	(0.00328)	(0.00464)
Father's education	0.0492***	0.0393***	0.0443***	0.0342***	0.0227***	0.0447***	0.0425***
	(0.00647)	(0.00492)	(0.00554)	(0.00543)	(0.00599)	(0.00506)	(0.00683)
Mother's education	0.0507***	0.0621***	0.0528***	0.0449***	0.0609***	0.0810***	0.0699***
	(0.00671)	(0.00490)	(0.00583)	(0.00554)	(0.00640)	(0.00543)	(0.00724)
Household income (log)	0.0181	0.0646***	0.0686***	0.0644***	-0.0443*	-0.0155	-0.0114
	(0.0136)	(0.0182)	(0.0200)	(0.0181)	(0.0250)	(0.0218)	(0.0291)
Urban environment	0.339***	0.107**	0.0391	0.152***	-0.152***	-0.155***	-0.151***
	(0.0552)	(0.0428)	(0.0480)	(0.0486)	(0.0490)	(0.0417)	(0.0551)
Familias en accion	0.116	0.196***	0.281**	0.208**	-0.0138	0.0190	-0.0497
	(0.0981)	(0.0660)	(0.116)	(0.0950)	(0.0445)	(0.0375)	(0.0493)
Birth order	0.261***	0.276***	0.151***	0.219***	0.0692**	0.142***	0.128***
	(0.0387)	(0.0312)	(0.0367)	(0.0390)	(0.0341)	(0.0287)	(0.0408)
Constant	-1.842***	-2.802***	-2.218***	-1.342***	0.0370	-5.912***	-5.769***

	(0.310)	(0.300)	(0.337)	(0.321)	(0.388)	(0.337)	(0.453)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291
Adjusted R-squared	0.793	0.792	0.799	0.816	0.646	0.739	0.752

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

Table 17 - Multivariate OLS 3 (Aggregate sibling exposure)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Aggregate sibling exposure	-0.0266***	-0.0259***	-0.0154***	-0.0187***	-0.00766***	-0.0168***	-0.0188***
	(0.00278)	(0.00214)	(0.00248)	(0.00256)	(0.00239)	(0.00201)	(0.00302)
Age	0.208***	0.283***	0.227***	0.0959***	0.254***	1.724***	1.723***
	(0.0432)	(0.0334)	(0.0365)	(0.0365)	(0.0385)	(0.0327)	(0.0432)
Age (sq.)	0.0250***	0.0214***	0.0235***	0.0295***	0.0170***	-0.0454***	-0.0446***
	(0.00183)	(0.00142)	(0.00155)	(0.00154)	(0.00166)	(0.00141)	(0.00186)
Girl	0.377***	0.283***	0.340***	0.308***	0.274***	0.508***	0.577***
	(0.0450)	(0.0351)	(0.0388)	(0.0379)	(0.0395)	(0.0334)	(0.0438)
Father's Age	0.00350	0.00379	0.00312	0.00206	-0.00760***	0.00370	0.00225
	(0.00341)	(0.00297)	(0.00328)	(0.00313)	(0.00288)	(0.00245)	(0.00353)
Mother's Age	0.00201	0.00666*	0.00533	0.00185	0.00663*	0.00312	0.00209
	(0.00407)	(0.00353)	(0.00393)	(0.00379)	(0.00381)	(0.00323)	(0.00459)
Father's education	0.0492***	0.0389***	0.0440***	0.0338***	0.0225***	0.0444***	0.0426***
	(0.00648)	(0.00493)	(0.00555)	(0.00544)	(0.00598)	(0.00506)	(0.00682)
Mother's education	0.0501***	0.0617***	0.0524***	0.0447***	0.0604***	0.0804***	0.0694***
	(0.00672)	(0.00491)	(0.00583)	(0.00555)	(0.00639)	(0.00543)	(0.00723)
Household income (log)	0.0175	0.0644***	0.0671***	0.0633***	-0.0418*	-0.0118	-0.0102
	(0.0136)	(0.0182)	(0.0201)	(0.0182)	(0.0250)	(0.0218)	(0.0291)
Urban environment	0.348***	0.110**	0.0437	0.153***	-0.156***	-0.159***	-0.153***
	(0.0553)	(0.0430)	(0.0480)	(0.0487)	(0.0490)	(0.0417)	(0.0551)
Familias en accion	0.115	0.195***	0.280**	0.204**	-0.0118	0.0206	-0.0454
	(0.0982)	(0.0662)	(0.116)	(0.0952)	(0.0445)	(0.0375)	(0.0493)
Firstborn	-0.741***	-0.633***	-0.273**	-0.417***	-0.173	-0.387***	-0.374***
	(0.131)	(0.103)	(0.117)	(0.123)	(0.113)	(0.0956)	(0.132)
Birth order 2 or 3	-0.479***	-0.418***	-0.188**	-0.211**	-0.000580	-0.188***	-0.166
	(0.102)	(0.0797)	(0.0901)	(0.0958)	(0.0862)	(0.0723)	(0.104)
Constant	-0.806**	-1.837***	-1.755***	-0.679**	0.262	-5.399***	-5.266***
	(0.323)	(0.309)	(0.351)	(0.335)	(0.398)	(0.344)	(0.463)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291
Adjusted R-squared	0.792	0.791	0.798	0.816	0.646	0.739	0.752

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

5.1.2 INSTRUMENTAL VARIABLE ESTIMATIONS

OLS results support the idea that family size and shooling hold a negative relation. However, as previously explained, these variables are likely to be simultaneously determined generating biased OLS results. Using different instruments as a source of exogenous variation, this segment of the study presents the results of estimations in the context of causal explanations. Initially, first stage estimations testing the existence of a strong relationship between the instruments and the causal variable of interest, either number of siblings or aggregate sibling exposure, are presented for bivariate and multivariate specifications. Afterwards, second stage results are displayed according to the models described in section 4.

5.1.2.1. First stage regressions

Multiple birth instrument

Table 18 and Table 19 display the output for first stage estimation of the multiple birth instrument on number of siblings. Twin instrument coefficients are always positive though only statistically significant at 1% level of statistical significance for two samples. In the ELCA samples multivariate estimations are not significant and the effect comes close to cero in the ELCA 2013 sample. Based on these results it is likely that the instrumentation will work better in the samples in which the effect is stronger; that is ENCV 2010, 2012 and 2013.

Table 18 - Bivariate First stage (twin birth on number of siblings)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Twin instr.	0.663***	0.104	0.692**	0.952***	0.521	0.667*	0.721*
	(0.244)	(0.255)	(0.342)	(0.298)	(0.343)	(0.377)	(0.400)
Constant	1.775***	1.896***	1.746***	1.627***	1.929***	1.980***	2.057***
	(0.0214)	(0.0176)	(0.0186)	(0.0183)	(0.0175)	(0.0177)	(0.0183)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291
Adjusted R-squared	0.002	0.000	0.001	0.002	0.000	0.000	0.001

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

Table 19 - Multivariate First Stage (twin birth on number of siblings)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Twin instr.	0.522***	0.222	0.574**	0.654***	0.211	0.0588	0.109
	(0.176)	(0.184)	(0.244)	(0.210)	(0.241)	(0.267)	(0.318)
Age	0.202***	0.151***	0.104***	0.104***	0.178***	0.185***	0.135***
	(0.0291)	(0.0237)	(0.0247)	(0.0245)	(0.0236)	(0.0241)	(0.0281)
Age (sq.)	-0.00527***	-0.00279***	-0.000721	-0.000849	-0.00396***	-0.00424***	-0.00292**
	(0.00125)	(0.00102)	(0.00106)	(0.00104)	(0.00103)	(0.00105)	(0.00123)
Girl	0.0406	0.0118	-0.00831	0.00804	-0.0201	-0.0268	-0.0419
	(0.0308)	(0.0253)	(0.0265)	(0.0257)	(0.0246)	(0.0250)	(0.0290)
Father's Age	0.00480**	-0.00297	-0.000461	0.000922	-0.000805	-0.00107	-0.000723
	(0.00234)	(0.00214)	(0.00224)	(0.00212)	(0.00180)	(0.00184)	(0.00234)
Mother's Age	-0.0393***	-0.0344***	-0.0351***	-0.0375***	-0.0380***	-0.0395***	-0.0300***
	(0.00276)	(0.00252)	(0.00267)	(0.00255)	(0.00235)	(0.00239)	(0.00302)

Father's education	-0.0289***	-0.0193***	-0.0202***	-0.0112***	-0.0173***	-0.0184***	-0.0141***
	(0.00442)	(0.00354)	(0.00378)	(0.00368)	(0.00371)	(0.00376)	(0.00449)
Mother's education	-0.0327***	-0.0347***	-0.0269***	-0.0331***	-0.0375***	-0.0394***	-0.0411***
	(0.00460)	(0.00351)	(0.00397)	(0.00375)	(0.00396)	(0.00403)	(0.00475)
Household income (log)	-0.0162*	-0.0463***	-0.0549***	-0.0141	-0.0255*	-0.0244	-0.0500***
	(0.00932)	(0.0131)	(0.0137)	(0.0123)	(0.0153)	(0.0160)	(0.0188)
Urban environment	0.00340	-0.123***	-0.108***	-0.151***	-0.145***	-0.159***	-0.131***
	(0.0379)	(0.0309)	(0.0328)	(0.0330)	(0.0306)	(0.0312)	(0.0366)
Firstborn	-3.251***	-3.371***	-3.286***	-3.296***	-3.538***	-3.536***	-2.617***
	(0.0622)	(0.0500)	(0.0530)	(0.0542)	(0.0471)	(0.0472)	(0.0640)
Birth order 2 or 3	-2.279***	-2.360***	-2.231***	-2.257***	-2.343***	-2.343***	-2.063***
	(0.0579)	(0.0460)	(0.0495)	(0.0511)	(0.0433)	(0.0431)	(0.0573)
Constant	4.701***	5.691***	5.699***	5.173***	5.363***	5.415***	5.341***
	(0.216)	(0.218)	(0.233)	(0.221)	(0.238)	(0.246)	(0.291)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291
Adjusted R-squared	0.481	0.486	0.496	0.507	0.506	0.501	0.373

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

Likewise, Table 20 and Table 21 display the results of the first stage for multiple births instrumenting aggregate sibling exposure. All coefficients are also positive for this instrument and very significant for ENCV 2010 and 2013 as well. Additionally and contrary to first stages for the instrumentation of number of siblings, the effect is particularly strong and highly significant the two versions of ELCA 2013 but only in the bivariate version. The effect is quite consistent across both estimations for ENCV 2013.

The twin instrument seems to have strong first stages instrumenting both the number of siblings and aggregate sibling exposure in both bivariate and multivariate specifications in the ENCV 2010 and 2013 samples. In the instrumentation for number of siblings, in the ENCV 2012 sample there is also quite an effect though not as significant. This is case for the ELCA 2010 sample in the aggregate sibling sample instrumentation. Other samples display less consistent first stages. In particular, estimations in all ELCA and ENCV 2011 samples for the twin effect over number of siblings instrumentation are weak. So is the case for ENCV 2012 in the aggregate sibling effect instrumentation.

Table 20 - Bivariate First Stage (twin birth on aggregate sibling exposure)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Twin instr.	5.943**	1.288	2.563	8.887***	7.551**	11.23***	14.52***
	(2.358)	(2.459)	(3.366)	(3.008)	(3.177)	(3.468)	(3.772)
Constant	14.52***	15.38***	14.50***	13.72***	15.25***	15.76***	15.60***
	(0.207)	(0.170)	(0.183)	(0.184)	(0.163)	(0.162)	(0.173)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291
Adjusted R-squared	0.001	0.000	0.000	0.002	0.001	0.001	0.003

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

Table 21 - Multivariate First Stage (twin birth on aggregate sibling exposure)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)

Twin instr.	4.232***	3.149**	2.529	5.044***	3.788**	3.537*	4.614*
	(1.436)	(1.454)	(1.961)	(1.700)	(1.862)	(2.025)	(2.428)
Age	2.903***	2.502***	2.098***	1.993***	3.028***	3.120***	2.899***
	(0.237)	(0.188)	(0.199)	(0.198)	(0.182)	(0.183)	(0.215)
Age (sq.)	-0.0625***	-0.0434***	-0.0302***	-0.0277***	-0.0667***	-0.0690***	-0.0648***
	(0.0102)	(0.00805)	(0.00852)	(0.00845)	(0.00799)	(0.00797)	(0.00939)
Girl	0.456*	0.0547	0.0162	0.0683	-0.224	-0.258	-0.111
	(0.251)	(0.200)	(0.213)	(0.208)	(0.190)	(0.189)	(0.222)
Father's Age	0.0373*	0.0135	0.00264	0.00468	0.00704	0.00740	0.00428
	(0.0190)	(0.0169)	(0.0180)	(0.0172)	(0.0139)	(0.0139)	(0.0179)
Mother's Age	-0.235***	-0.187***	-0.165***	-0.193***	-0.201***	-0.208***	-0.131***
	(0.0224)	(0.0200)	(0.0215)	(0.0207)	(0.0182)	(0.0181)	(0.0231)
Father's education	-0.219***	-0.103***	-0.119***	-0.0476	-0.114***	-0.120***	-0.0837**
	(0.0360)	(0.0281)	(0.0304)	(0.0298)	(0.0287)	(0.0286)	(0.0343)
Mother's education	-0.172***	-0.226***	-0.136***	-0.193***	-0.193***	-0.214***	-0.233***
	(0.0374)	(0.0278)	(0.0320)	(0.0304)	(0.0306)	(0.0306)	(0.0364)
Household income (log)	-0.0249	-0.101	-0.263**	0.179*	0.155	0.178	-0.105
	(0.0759)	(0.104)	(0.110)	(0.0998)	(0.119)	(0.121)	(0.144)
Urban environment	-0.0159	-0.341	-0.266	-0.764***	-0.793***	-0.777***	-0.500*
	(0.309)	(0.245)	(0.264)	(0.267)	(0.236)	(0.237)	(0.280)
Firstborn	-33.77***	-35.60***	-35.15***	-36.64***	-35.61***	-35.92***	-30.07***
	(0.507)	(0.396)	(0.427)	(0.440)	(0.364)	(0.358)	(0.489)
Birth order 2 or 3	-20.49***	-22.23***	-21.56***	-23.00***	-21.59***	-21.78***	-19.12***
	(0.472)	(0.364)	(0.398)	(0.415)	(0.335)	(0.327)	(0.438)
Constant	25.20***	29.42***	32.14***	29.32***	23.85***	23.54***	22.41***
	(1.758)	(1.724)	(1.877)	(1.795)	(1.835)	(1.867)	(2.225)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291
Adjusted R-squared	0.632	0.652	0.662	0.682	0.657	0.661	0.590

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

Uniform sex composition instrument

The results for first stage estimations for the uniform sex composition are displayed in Table 22 and Table 23. These models estimated the probability of having a third child at household level for a sample of families who either had two or three children. The outcome is having a third child and the depedent variable is a dummy variable taking the value of 1 if the first two births shared sex and 0 if otherwise. The first model does not differentiate the effect by the sex of the first pair of children and reports important positive and highly statistically significant coefficients in all samples but in the ELCA 2013 constant sibling sample. When the estimation distinguishes between female and male sex allocation for the first two births results remain consistent for all samples except for the ENCV 2012. Apparently a male composition of the

first two births has a slighty stronger effect on the probabily having a third child. This suggest a particular prefence for girls over boys in this situation.

Table 22 - First stage Probit (Uniform sex composition of births 1 and 2 on the probability of having a third child). Only for households with at least one observation used in the 2SLS estimation

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Uniform composition (Births 1 & 2)	0.276***	0.241***	0.168***	0.287***	0.339***	0.316***	0.0139
	(0.0565)	(0.0450)	(0.0513)	(0.0522)	(0.0416)	(0.0425)	(0.0557)
Constant	-0.802***	-0.790***	-0.822***	-0.895***	-0.815***	-0.819***	-0.499***
	(0.0349)	(0.0279)	(0.0300)	(0.0312)	(0.0249)	(0.0256)	(0.0380)
Observations	2,518	3,957	3,299	3,198	4,782	4,565	2,220

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

Table 23 - First stage Probit (Uniform sex composition of births 1 and 2 by sex on the probability of having a third child). Only for households with at least one observation used in the 2SLS estimation)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Male composition (Births 1 & 2)	0.282***	0.316***	0.115	0.288***	0.361***	0.334***	0.0155
	(0.0737)	(0.0588)	(0.0711)	(0.0715)	(0.0551)	(0.0561)	(0.0709)
Female composition (Births 1 & 2)	0.271***	0.177***	0.208***	0.286***	0.321***	0.300***	0.0125
	(0.0702)	(0.0557)	(0.0623)	(0.0634)	(0.0517)	(0.0530)	(0.0674)
Constant	-0.802***	-0.790***	-0.822***	-0.895***	-0.815***	-0.819***	-0.499***
	(0.0349)	(0.0279)	(0.0300)	(0.0312)	(0.0249)	(0.0256)	(0.0380)
Observations	2,518	3,957	3,299	3,198	4,782	4,565	2,220

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

5.1.2.2. Second stage estimations

5.1.2.2.1. Number of siblings instrumented by multiple births

Results for 2SLS estimates employing multiple births as an instrument for number of siblings (Table 24 and Table 25) vary in sign and magnitude. There are no significant effects for any sample. Nonetheless, coefficients with large magnitude tend to display large standard errors while small magnitudes are associated to smaller errors. Errors and magnitudes in ELCA samples are relatively large in the multivariate estimation. In particular, ENCV samples for 2010, 2012 and 2013 display the smaller errors. Based on first stage estimates, multiple birth had an important effect in ENCV samples in 2010 and 2013. In these samples, coefficients display small standard errors in both models but signs in the multivariate model differ. Also, coefficients with small errors in the second model have either negative and smaller or positive magnitudes compared to OLS estimates.

Table 24 - Bivariate 2SLS (twin birth as instrument of number of siblings)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010	ELCA 2013	ELCA 2013

					Outcome	Outcome	Outcome (CSS)
Number of siblings	-0.327	-8.837	-2.757	-1.198	2.089	0.812	3.029
	(0.850)	(21.97)	(1.772)	(0.845)	(1.872)	(1.169)	(2.213)
Constant	6.546***	22.67	10.91***	8.142***	1.181	6.674***	2.141
	(1.513)	(41.65)	(3.098)	(1.379)	(3.614)	(2.316)	(4.557)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291

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

Table 25 - Multivariate 2SLS (twin birth as instrument of number of siblings)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Number of siblings	0.533	-0.819	-0.368	-0.757	2.000	-2.562	2.389
	(0.549)	(1.273)	(0.630)	(0.515)	(2.957)	(12.58)	(8.613)
Age	0.0241	0.347*	0.234***	0.139**	-0.124	2.146	1.347
	(0.121)	(0.196)	(0.0743)	(0.0656)	(0.528)	(2.327)	(1.163)
Age (sq.)	0.0294***	0.0201***	0.0237***	0.0293***	0.0254**	-0.0551	-0.0364
	(0.00356)	(0.00389)	(0.00163)	(0.00172)	(0.0120)	(0.0533)	(0.0253)
Girl	0.347***	0.292***	0.336***	0.314***	0.316***	0.444	0.680*
	(0.0544)	(0.0416)	(0.0397)	(0.0415)	(0.0872)	(0.344)	(0.370)
Father's Age	0.000242	0.000789	0.00284	0.00280	-0.00604	0.000844	0.00378
	(0.00452)	(0.00497)	(0.00333)	(0.00343)	(0.00526)	(0.0145)	(0.00930)
Mother's Age	0.0289	-0.0167	-0.00482	-0.0229	0.0843	-0.0948	0.0765
	(0.0218)	(0.0441)	(0.0224)	(0.0197)	(0.113)	(0.497)	(0.258)
Father's education	0.0701***	0.0257	0.0384***	0.0259***	0.0581	-0.000894	0.0786
	(0.0173)	(0.0253)	(0.0137)	(0.00822)	(0.0520)	(0.231)	(0.122)
Mother's education	0.0724***	0.0388	0.0448**	0.0234	0.137	-0.0174	0.173
	(0.0197)	(0.0445)	(0.0181)	(0.0180)	(0.111)	(0.496)	(0.355)
Household income (log)	0.0260	0.0286	0.0512	0.0499**	0.00936	-0.0796	0.117
	(0.0174)	(0.0622)	(0.0405)	(0.0209)	(0.0855)	(0.310)	(0.434)
Urban environment	0.344***	0.0203	0.00751	0.0504	0.142	-0.552	0.170
	(0.0616)	(0.162)	(0.0851)	(0.0953)	(0.437)	(1.997)	(1.140)
Firstborn	1.889	-2.471	-0.941	-2.225	7.180	-8.846	6.450
	(1.787)	(4.291)	(2.070)	(1.704)	(10.46)	(44.47)	(22.55)
Birth order 2 or 3	1.279	-1.774	-0.678	-1.489	4.853	-5.827	5.127
	(1.254)	(3.003)	(1.406)	(1.168)	(6.929)	(29.47)	(17.77)
Constant	-3.974	2.069	-0.159	2.673	-10.68	8.130	-18.57
	(2.599)	(7.255)	(3.614)	(2.691)	(15.88)	(68.11)	(46.01)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291
R-squared	0.742	0.744	0.793	0.782	0.072		0.049

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

5.1.2.2.2. Aggregate sibling exposure instrumented by multiple births

Table 26 and Table 27 display the results for 2SLS estimates using the multiple birth instrument on aggregate sibling exposure. When compared to OLS results, 2SLS coefficients tend to display divergence and no statistical significance except for ELCA 2013 CSS sample. Also, OLS estimates for bivariate models were positive; for ENCV samples the direction of the coefficient changed. In the bivariate case, samples in which the first stage effect was significant at 1% level of significant display positive, negative, significant and not significant coefficients. Although the multivariate estimation presents no statistical significance in aggregate sibling exposure signs don't follow a clear pattern.

Table 26 – Bivariate 2SLS (twin birth as instrument of aggregate sibling exposure)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Aggregate	-0.0364	-0.716	-0.744	-0.128	0.144	0.0482	0.150**
sibling exposure	(0.0975)	(1.484)	(1.072)	(0.0981)	(0.0973)	(0.0620)	(0.0712)
Constant	6.495***	16.93	16.89	7.954***	3.015**	7.522***	6.028***
	(1.421)	(22.84)	(15.56)	(1.350)	(1.486)	(0.979)	(1.114)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291
R-squared						0.002	

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

Table 27 - Multivariate 2SLS (twin birth as instrument of aggregate sibling exposure)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Aggregate	0.0658	-0.0576	-0.0835	-0.0982	0.112	-0.0426	0.0563
sibling exposure	(0.0683)	(0.0823)	(0.151)	(0.0669)	(0.117)	(0.102)	(0.111)
Age	-0.0591	0.367*	0.370	0.256*	-0.106	1.805***	1.506***
	(0.205)	(0.209)	(0.317)	(0.138)	(0.357)	(0.318)	(0.324)
Age (sq.)	0.0307***	0.0198***	0.0214***	0.0272***	0.0250***	-0.0472***	-0.0398***
	(0.00477)	(0.00386)	(0.00481)	(0.00247)	(0.00803)	(0.00713)	(0.00743)
Girl	0.338***	0.285***	0.340***	0.314***	0.301***	0.502***	0.586***
	(0.0589)	(0.0359)	(0.0415)	(0.0416)	(0.0525)	(0.0428)	(0.0483)
Father's Age	0.000346	0.00400	0.00323	0.00256	-0.00843**	0.00391	0.00181
	(0.00451)	(0.00323)	(0.00352)	(0.00342)	(0.00342)	(0.00258)	(0.00380)
Mother's Age	0.0234	0.000656	-0.00570	-0.0134	0.0306	-0.00243	0.0123
	(0.0165)	(0.0160)	(0.0251)	(0.0134)	(0.0239)	(0.0213)	(0.0152)
Father's	0.0692***	0.0356***	0.0358*	0.0298***	0.0362**	0.0410***	0.0497***
education	(0.0166)	(0.00991)	(0.0187)	(0.00668)	(0.0150)	(0.0132)	(0.0117)
Mother's	0.0663***	0.0542***	0.0434**	0.0295**	0.0836***	0.0745***	0.0875***
education	(0.0142)	(0.0193)	(0.0215)	(0.0142)	(0.0239)	(0.0225)	(0.0270)
Household	0.0191	0.0607***	0.0495	0.0782***	-0.0590*	-0.00953	0.00377
income (log)	(0.0153)	(0.0203)	(0.0453)	(0.0233)	(0.0336)	(0.0281)	(0.0326)
Urban	0.347***	0.101**	0.0251	0.0899	-0.0598	-0.179**	-0.115
environment	(0.0621)	(0.0510)	(0.0660)	(0.0745)	(0.110)	(0.0900)	(0.0823)
Firstborn	2.377	-1.762	-2.668	-3.328	4.075	-1.319	1.890
	(2.310)	(2.932)	(5.293)	(2.455)	(4.183)	(3.658)	(3.337)
Birth order 2 or 3	1.412	-1.123	-1.658	-2.037	2.576	-0.753	1.276
	(1.403)	(1.830)	(3.246)	(1.543)	(2.537)	(2.218)	(2.124)
Constant	-3.125*	-0.895	0.431	1.635	-2.616	-4.737*	-7.076***
	(1.754)	(2.445)	(4.864)	(1.995)	(2.840)	(2.425)	(2.530)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291
R-squared	0.737	0.784	0.771	0.781	0.529	0.734	0.716

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

5.1.2.2.3. Number of siblings instrumented by uniform sex composition

When using the uniform sex composition of the first two birts as an instrument for number of siblings in the ENCV samples, coefficients, when compared to the OLS estimations, lose statistical significance (Table 28 and Table 29). In the ELCA samples this is only true for the multivariate version of the second stage. Very large coefficients, either positive or negative, tend to display larger errors than coefficients closer to zero except when there is statistitical significance which is the case in the ELCA samples.

Table 28 - Bivariate 2SLS (Uniform sex composition of births 1 and 2 as instrument of Number of siblings)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Number of siblings	-2.206	-0.415	-3.077	0.552	-1.001*	-0.745**	0.0321
	(1.846)	(0.355)	(3.381)	(0.997)	(0.550)	(0.328)	(0.185)
Constant	9.892***	6.704***	11.47*	5.288***	7.147***	9.759***	8.311***
	(3.286)	(0.674)	(5.910)	(1.627)	(1.063)	(0.650)	(0.384)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291

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

Table 29 - Multivariate 2SLS (Uniform sex composition of births 1 and 2 as instrument of Number of siblings)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Number of siblings	-2.408	-0.201	-0.476	1.668	-1.164	-0.516	0.0856
	(2.757)	(0.384)	(0.562)	(1.892)	(1.110)	(0.419)	(0.174)
Age	0.623	0.254***	0.245***	-0.108	0.437**	1.768***	1.657***
	(0.567)	(0.0669)	(0.0686)	(0.201)	(0.202)	(0.0845)	(0.0487)
Age (sq.)	0.0137	0.0218***	0.0236***	0.0312***	0.0130***	-0.0464***	-0.0432***
	(0.0151)	(0.00178)	(0.00164)	(0.00283)	(0.00482)	(0.00230)	(0.00194)
Girl	0.462***	0.285***	0.335***	0.294***	0.252***	0.499***	0.584***
	(0.136)	(0.0355)	(0.0402)	(0.0615)	(0.0531)	(0.0365)	(0.0448)
Father's Age	0.0135	0.00260	0.00279	0.000793	-0.00862**	0.00305	0.00211
	(0.0139)	(0.00319)	(0.00338)	(0.00516)	(0.00363)	(0.00259)	(0.00357)
Mother's Age	-0.0857	0.00460	-0.00863	0.0676	-0.0360	-0.0140	0.00755
	(0.108)	(0.0137)	(0.0201)	(0.0709)	(0.0424)	(0.0169)	(0.00694)
Father's education	-0.0143	0.0377***	0.0362***	0.0528**	0.00340	0.0366***	0.0462***
	(0.0801)	(0.00893)	(0.0125)	(0.0226)	(0.0205)	(0.00930)	(0.00726)
Mother's education	-0.0254	0.0602***	0.0419**	0.103	0.0184	0.0633***	0.0779***
	(0.0925)	(0.0142)	(0.0164)	(0.0628)	(0.0423)	(0.0175)	(0.0102)
Household income (log)	-0.0202	0.0572**	0.0452	0.0823**	-0.0716*	-0.0297	0.00200
	(0.0500)	(0.0255)	(0.0374)	(0.0381)	(0.0413)	(0.0245)	(0.0300)
Urban environment	0.361***	0.0953	-0.00456	0.423	-0.318*	-0.228***	-0.133**
	(0.102)	(0.0635)	(0.0794)	(0.301)	(0.172)	(0.0794)	(0.0602)
Firstborn	-7.677	-0.390	-1.299	5.782	-4.015	-1.612	0.421
	(8.970)	(1.295)	(1.847)	(6.249)	(3.928)	(1.485)	(0.465)
Birth order 2 or 3	-5.424	-0.318	-0.921	3.996	-2.560	-1.034	0.375
	(6.286)	(0.907)	(1.255)	(4.281)	(2.602)	(0.985)	(0.369)
Constant	9.836	-1.447	0.464	-9.878	6.293	-2.948	-6.271***
	(12.96)	(2.205)	(3.229)	(9.807)	(5.972)	(2.297)	(1.028)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291
R-squared	0.302	0.789	0.787	0.544	0.473	0.718	0.746

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

5.1.2.2.4. Aggregate sibling exposure instrumented by uniform sex composition

Similarly, when instrumenting the sibling exposure employing uniform sex composition coefficients for ENCV don't report any statistical significance and, with the exception of the 2013 sample in the bivariate model, are all negative. Coefficients that were significant in the bivariate estimation display reduced magnitudes but keep the same direction. These results are reported in Table 30 and in Table 31.

Table 30 - Bivariate 2SLS (Uniform sex composition of births 1 and 2 as instrument of aggregate sibling exposure)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Aggregate sibling exposure	-0.207	-0.0481	-0.307	0.0884	-0.0990*	-0.0857**	0.00422
	(0.181)	(0.0428)	(0.360)	(0.157)	(0.0571)	(0.0407)	(0.0243)
Constant	8.974***	6.657***	10.54**	4.973**	6.727***	9.636***	8.311***
	(2.635)	(0.660)	(5.223)	(2.156)	(0.872)	(0.643)	(0.382)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291

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

Table 31 - Multivariate 2SLS (Uniform sex composition of births 1 and 2 as instrument of aggregate sibling exposure)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Aggregate sibling exposure	-0.0729	-0.0162	-0.0279	-0.143	-0.0495	-0.0377	0.0154
	(0.0471)	(0.0309)	(0.0320)	(0.125)	(0.0394)	(0.0296)	(0.0314)
Age	0.345**	0.264***	0.254***	0.346	0.380***	1.789***	1.624***
	(0.144)	(0.0841)	(0.0761)	(0.252)	(0.125)	(0.0977)	(0.101)
Age (sq.)	0.0220***	0.0216***	0.0231***	0.0260***	0.0143***	-0.0468***	-0.0424***
	(0.00351)	(0.00195)	(0.00183)	(0.00388)	(0.00311)	(0.00248)	(0.00276)
Girl	0.400***	0.283***	0.339***	0.318***	0.265***	0.503***	0.582***
	(0.0508)	(0.0352)	(0.0389)	(0.0467)	(0.0412)	(0.0344)	(0.0445)
Father's Age	0.00519	0.00342	0.00308	0.00274	-0.00734**	0.00387	0.00199
	(0.00388)	(0.00301)	(0.00328)	(0.00382)	(0.00295)	(0.00248)	(0.00358)
Mother's Age	-0.00878	0.00847	0.00346	-0.0221	-0.00163	-0.00141	0.00699
	(0.0117)	(0.00681)	(0.00656)	(0.0244)	(0.00877)	(0.00693)	(0.00615)
Father's education	0.0389***	0.0399***	0.0424***	0.0277***	0.0179**	0.0416***	0.0463***
	(0.0122)	(0.00589)	(0.00671)	(0.00877)	(0.00753)	(0.00617)	(0.00734)
Mother's education	0.0418***	0.0635***	0.0510***	0.0208	0.0525***	0.0756***	0.0780***
	(0.0108)	(0.00853)	(0.00730)	(0.0249)	(0.00999)	(0.00835)	(0.0103)
Household income (log)	0.0161	0.0649***	0.0643***	0.0865***	-0.0342	-0.0104	-0.000621
	(0.0141)	(0.0185)	(0.0218)	(0.0319)	(0.0258)	(0.0221)	(0.0290)
Urban environment	0.347***	0.115***	0.0405	0.0545	-0.189***	-0.175***	-0.137**
	(0.0570)	(0.0441)	(0.0489)	(0.114)	(0.0590)	(0.0479)	(0.0582)
Firstborn	-2.309	-0.290	-0.713	-4.985	-1.660	-1.142	0.661
	(1.592)	(1.102)	(1.128)	(4.596)	(1.405)	(1.067)	(0.950)
Birth order 2 or 3	-1.431	-0.204	-0.459	-3.078	-0.902	-0.646	0.494
	(0.968)	(0.689)	(0.694)	(2.886)	(0.854)	(0.648)	(0.607)
Constant	0.364	-2.114**	-1.361	2.961	1.232	-4.853***	-6.160***
	(1.227)	(0.959)	(1.086)	(3.697)	(1.018)	(0.774)	(0.833)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291
R-squared	0.779	0.790	0.798	0.730	0.632	0.736	0.745

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

5.1.2.2.5. Household fixed effects estimations

Household fixed effects models for both instruments (Table 32 and Table 33), report some positive statistical significance coefficients for *aggregate siblings exposure* when the instrumentation is based on multiple births. The coefficients for the second stage based on uniform sex compotision have no statistical significance and have different signs.

Table 32 - FE, 2SLS (Twin birth as instrument of aggregate sibling exposure)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Aggregate sibling exposure	0.00348	0.0421***	0.00260	0.0160*	-0.00317	0.0111	0.0156**
	(0.00966)	(0.00872)	(0.00795)	(0.00848)	(0.0102)	(0.00787)	(0.00792)
Age	0.208***	0.0387	0.305***	0.105*	0.226***	1.593***	1.600***
	(0.0678)	(0.0591)	(0.0610)	(0.0633)	(0.0697)	(0.0523)	(0.0579)
Age (sq.)	0.0222***	0.0296***	0.0190***	0.0275***	0.0174***	-0.0413***	-0.0405***
	(0.00290)	(0.00254)	(0.00260)	(0.00271)	(0.00304)	(0.00230)	(0.00258)
Girl	0.400***	0.313***	0.325***	0.311***	0.245***	0.515***	0.493***
	(0.0638)	(0.0520)	(0.0547)	(0.0580)	(0.0609)	(0.0451)	(0.0542)
Constant	0.121	0.436	-0.382	0.542*	0.284	-4.260***	-4.439***
	(0.321)	(0.268)	(0.289)	(0.310)	(0.308)	(0.232)	(0.278)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291
Number of panel units	2,518	3,957	3,299	3,198	4,782	4,565	2,220

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

Table 33 - FE, 2SLS (Uniform sex composition of births 1 and 2 as instrument of aggregate sibling exposure)

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA 2010 Outcome	ELCA 2013 Outcome	ELCA 2013 Outcome (CSS)
Aggregate sibling exposure	0.0175	-0.0758	-0.211	-0.00466	0.114	-0.0446	-0.0836
	(0.221)	(0.339)	(0.228)	(0.132)	(0.281)	(0.194)	(0.229)
Birth order	-0.0930	0.716	1.526	0.147	-0.659	0.307	0.688
	(1.460)	(2.057)	(1.632)	(0.936)	(1.579)	(1.070)	(1.587)
Age	0.134	0.670	1.489	0.208	-0.421	1.896*	2.138*
	(1.160)	(1.814)	(1.267)	(0.656)	(1.552)	(1.055)	(1.239)
Age (sq.)	0.0243	0.0106	-0.0156	0.0245	0.0377	-0.0508	-0.0568
	(0.0334)	(0.0547)	(0.0370)	(0.0190)	(0.0487)	(0.0333)	(0.0376)
Girl	0.398***	0.301***	0.306***	0.309***	0.276***	0.495***	0.473***
	(0.0689)	(0.0600)	(0.0678)	(0.0586)	(0.0977)	(0.0820)	(0.0711)
Constant	0.640	-3.652	-8.839	-0.203	4.214	-6.069	-8.025
	(8.143)	(11.75)	(9.047)	(4.763)	(9.418)	(6.314)	(8.272)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291
Number of panel units	2,518	3,957	3,299	3,198	4,782	4,565	2,220

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

5.1.3 Robustness check

The main restriction that has been imposed on the analysis is the age range that conditions the working samples. Although the lower limit of this restriction is less arbitrary since children under age 5 rarely reported formal schooling, the educational process continues beyond age 18. As has been discussed, in the ELCA datasets individuals in ages above 18 might present a form of survivor bias due to the high probability of capturing individuals with preferences for schooling when evaluating the 2013 outcome. Also, extending samples for individuals older that 18 aggregates mandatory education and higher education situations in one estimation. With this in mind, if there is a common process related to family size simoultaneously affecting all stages of education then estimates should reflect it.

For each marginal change between ages 18 and 25 a sample was defined and all the previously estimated models where estimated again. This resulted a set of coefficients for the samples in ages 5 – 18, another for the samples in ages 5 – 19, 5 – 20 and so on until age 25. Figure 4 displays the distribution of these resulting coefficients by each instrumented variable. Coefficients for *number of siblings* tend to be concentrated around negative numbers; coefficients product of multiple birth instrumentation display an important concentration to the left of the distribution for coefficients basaed on the union sex composition instrument. The distribution of *aggregate sibling exposure* display the opposite situation. However, while coefficients for *number of siblings* are concentrated around negative values, coefficients for *aggregate sibling exposure* display concentrations that are not clearly associated with particular sign.

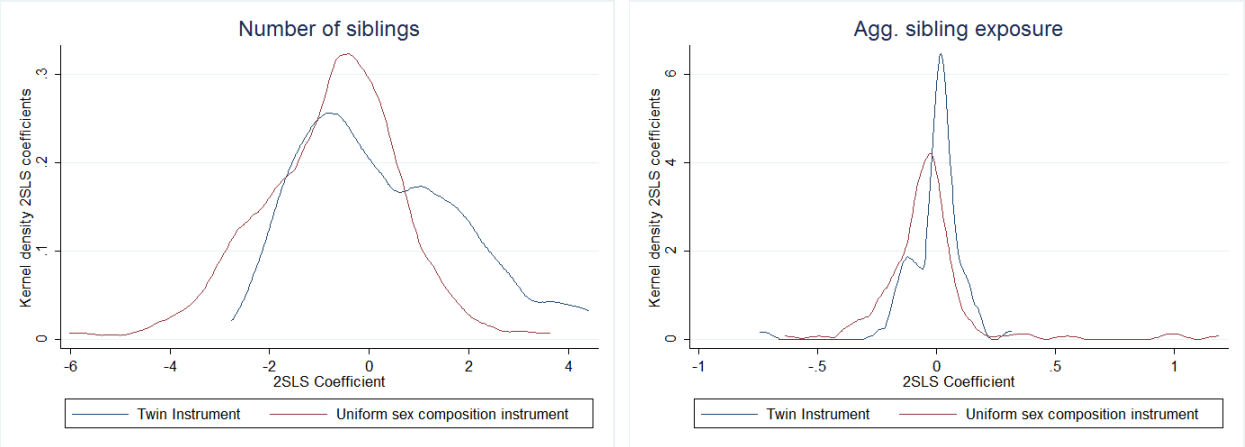

Figure 4 - Kernel density function of 2SLS pooled coefficients

When comparing estimations by model (Table 34) mean values based on the multiple birth instrument are consistenly smaller across models for both number of siblings and aggregate sibling exposure. Multivariate mean values are always negative while fixed-effect mean values display positive averages for both instruments.

Table 34- Means and SE by model and instrumentation

Instrumented	Model	Instrument	
		Uniform sex composition	Muliple Birth
Number of siblings	BV - 2SLS	-1.071	0.601
		(0.197)	(0.292)
	MV - 2SLS	-0.514	-0.012
		(0.188)	(0.144)
Aggregate sibling exposure	BV - 2SLS	-0.103	-0.032
		(0.022)	(0.0278)
	FE - 2SLS	0.026	0.022
		(0.041)	(0.008)
	MV - 2SLS	-0.056	-0.018
		(0.008)	(0.010)

5.2 Discussion

While bivariate OLS estimations report mixed effects, all multivariate regressions displayed negative and highly significant coefficients either for *number of siblings* or *aggregate sibling exposure*. With the exception of the ELCA 2010 sample, these numbers resemble what the literature finds for strategies that do not address endogeneity in the number of sibling estimations. These results suggest that larger families have a negative effect over children's schooling and are consistent with the theoretical predictions according to Becker's model.

The attempts to mitigate an omitted variable bias through exogenous variation of number of siblings by employing multiple births result in coefficients losing statistical significance and reporting larger magnitudes when compared to OLS estimates. In these estimations, large coefficients are associatied to large standard errors. These results are in line with other findings regarding inconsistency of coefficients between OLS models and 2SLS models, in particular by statistical significance, suggesting that there is not a clear and systematic effect between family size and educational attainment. However, sample size limitations might be affecting the precision of our estimations due to the scarcity of multiple births. In estimations were fist stages reported strong effects over number of siblings, 2SLS coefficients tend to be smaller and more precise to those associated to weaker first stages in the multivariate models. 2SLS estimations for number of siblings using uniform sex composition as an instrument share some similarities to multiple birth instrumentation; larger measures tend to be less precise and sign does not follow a clear trend. Instrumentation using uniform sex composition as source of exogenous variation for number of siblings delivers similar conclusions; coefficients vary in magnitude and direction. Two coefficients for ELCA samples reported some statistical significance in the bivariate version of the model that disappears when controls are included.

When sample restrictions become less rigid, mean values for number of siblings using both instruments are negative but their magnitude widely differ. Nonetheless, the mean value for multivariate estimations (-0.012) implies a small and negative effect. The coefficients estimated by (Marteleto, 2012) (0,064 (0,076)) and (Black, Devereux, & Salvanes, 2005) 0,038 (0,047)) display positive, small and non-significant results. In this sense, the present studies diverges with current literature estimates but reaches similar conclusions regarding the presence of an important, consistent and strong effect of a trade-off between family size and education.

Results based on the uniform sex composition of the first births don't provide evidence that suggests a systematic mechanism negatively relating family size and education in spite of not suffering from the sample issues found in the twin birth instrument and having a clear and strong first stage. Although different instruments will produce different results, none of the estimations for exogenous variation in number of siblings support through systematic and precise results the theoretical predictions stating a trade-off between quantity and quality.

A second new approach to the effect of family size on education tried to capture the intensity of this effect by analysis the exposure to siblings. The initial behavior of the variable through OLS estimations seemed to follow the results of number of siblings; negative and highly significant. Estimations using instrumental variables had similar results, all the coefficients, when controlling for a set of covariates, turned non-significant and sign as well as magnitude displayed divergence across samples. Multivariate 2SLS estimations produced negative and small coefficients for all ENCV samples when using the uniform sex composition instrument. Nonetheless, fixed effects models reported similar coefficients for the same instrument but some positive and significant results for the multiple birth instrument. These are the only statistically significant results after instrumentation for a model controlling for a set of covariates. Additionally, these positive coefficients are consisten with what the robustness check finds for both instruments.

In the context of the research question, *what is the short term effect of the total number of siblings on educational attainment in Colombia?* The quantitative analysis does not provide consistent results across multiple specifications and samples that could support one particular effect. Based on multivariate models' mean values in Table 34, it seems tempting to propose a a negative relationship but more in detail considerations do not necesarilly support this conclusion. Additionally, the only significant results of a instrumented estimation based on controlling for household fixed effects, suggest the opposite and are consistent with more robust analysis. However, this estimations do not exactly asses the effect of the number of siblings but the exposure to siblings.

With this in mind, based on the present estimates it is not possible to determine what is the short term effect of the total number of siblings, if there is any, on education in Colombia during the covered time span. Moreover, these results do not provide evidence that supports the predictions of the demand for children model. In particular, this predictions are quite specific in that they are proposed within an exogenous change framework:

"If p_c, π_z, and I were held constant, an exogenous increase in n would raise the shadow price of q, $\pi_q (= np_c)$, and thereby would reduce the demand for q" (Becker, 1973)

In this sense these findings are in line with the findings of other authors developing similar studies in developed as well as in developing countries; there is no evidence that supports a quantity-quality trade-off or a systematic relationship in this direction. There is not, howerever, strong evidence that disproves it.

6 Conclusion

In terms of the stated research objectives the study performed a quantitative analysis based on the available data that suited topic of analysis. The identification strategy, which was based on previous research, aimed at providing estimations that could answer the stated research question tackling endogeneity problems while preservating internal validity as much as it was possible. The study also explored complementary exposure measures that can aid the understanding of family effects on children's outcomes which allowed the extension of the methodological tools in order to control for household fixed effects.

Although the analysis cannot provide a consice and unique answer to the research question '*what is the short term effect of the total number of siblings on educational attainment in Colombia?'* it provides evidence of what the current strategies can say about this subject with the available data and about data limitations in addresing this issue. It also acts as a falsification excersise to the testing of theoretical preditions about how do individuals behave regarding parental decisions.

The current methodology investigating the relationship by which family size affects children's educational attainment is highly dependent on sample size. Because multiple births act as a good source of exogenous variation results based on this instrument are convincing. Nonetheless, registered based information that can address this limitation is not frequent in developing countries. While this restricts the possibility of large-scale studies it also confines research based on longitudinal perspectives.

7 References

Andes, U. d. (2011). *Colombia en movimiento - Un análisis descriptivo basado en la Encuesta Longitudinal Colombiana de la Unversidad de los Andes ELCA.* Bogotá: Universidad de los Andes.

Angrist, J. D. (1998). Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size. . *The American Economic Review*, 450-477.

Angrist, J. D., Lavy, V., & Schlosser, A. (2010). Multiple Experiments for the Causal Link between the Quantity and Quality of Children. *Journal of Labor Economics*, 773-824.

Becker, G. S. (1973). *A treatise on the family.* Cambridge: Hardvar University Press.

Black, S. E., Devereux, P. J., & Salvanes, K. G. (2005). The more the merrier? The effect of family size and birth order on children's education. *The Quarterly Journal of Economics*.

Black, S. E., Devereux, P. J., & Salvanes, K. G. (2010). Small Family, Smart Family? Family Size and the IQ Scores of Young Men. *The Journal of Human Resources, Vol. 45, No. 1 (Winter, 2010)*, 33-58.

Blake, J. (1981). *Family size and achievement.* . Berkeley: University of California Press.

Bongaarts, J. (2008). *Fertility Transitions in Developing Countries: Progress or Stagnation?* . New York: Population Council.

Booth, A. L., & Joo Kee, H. (2009). Birth order matters: the effect of family size and birth order on educational attainment. *Journal of Population Economics*, 367–397.

Bureau of Labor Statistics. (2016, May 28). *Bureau of Labor Statistics*. Retrieved from Employment proyections: http://www.bls.gov/emp/ep_chart_001.htm

Caceres, J. (2006). The impacts of family size on investment in child quality. *Journal of Human Resources 41*, 722–37.

Central Bank of Colombia (Banco de la República). (2016, May 23). *Banco de la República*. Retrieved from Minimu wage in Colombia (Salario mínimo legal en Colombia): http://obiee.banrep.gov.co/analytics/saw.dll?Go&Path=/shared/Consulta%20Series%20Estadisticas%20desde%20Excel/1.%20Salarios/1.1%20Salario%20minimo%20legal%20en%20Colombia/1.1.1%20Serie%20historica&Options=rdf&NQUser=salarios&NQPassword=salarios&lang=es

Cleland, J., & Wilson, C. (1987). Demand Theories of the Fertility Transition: an Iconoclastic View. *Population Studies, Vol. 41, No. 1*, 5-30.

Flórez, C. E., & Sánchez, L. M. (2013). *Fecundidad y familia en Colombia: ¿Hacia una una segunda transición demográfica?* Bogotá: Profamlia.

Galor, O. (2005). From Stagnation to Growth: Unified Growth Theory∗. In P. Aghion, & S. Durlauf, *Handbook of Economic Growth* (pp. 171–293). Elsevier.

Giménez, G. (2005). The Humn Capital Endowmen of Latin America and the Caribbean. *CEPAL Review 86*, 97 - 116.

Guinnane, T. W. (2011). The Historical Fertility Transition: A Guide for Economists. *Journal of Economic Literature, Vol. XLIX*, 589-614.

Hauser, R., & Sewell, W. (1985). Birth order and educational attainment in full sibships. *American Educational Research Journal, 22*, 1–23.

Leibowitz, A. (1974). Home Investments in Children. *Journal of Political Economy 82*, S111–S131.

Lesthaeghe, R. (2010). The Unfolding Story of the Second Demographic Transition. *Population Studies Center Research Report* , 10-696.

Malthus, R. (1826). *An essay on the principle of population.* Cambridge: Cambridge University Press.

Marteleto, L. J. (2012). The Changing Impact of Family Size on Adolescents' Schooling: Assessing the Exogenous Variation in Fertility Using Twins in Brazil. *Population Association of America*, 1453–1477.

Martinez, C. (2013). *Descenso de la fecundidad, participación laboral de la mujer y reducción de la pobreza en Colombia, 1990-2010.* Bogotá: Profamilia.

Ministry of National Education. (2014). *Sistema nacional de indicadores educativos para los niveles de preescolar, básica y media en Colombia.* Ministry of National Education (Ministerio de Educación).

National Administrative Department of Statistics (DANE). (2014). *COLOMBIA - Encuesta Nacional de Calidad de Vida - ENCV 2010.*

Nicholson, W., & Snyder, C. (2008). *Microeconomic theory.* Mason: Thomson.

Parra-Peña, R. I., Ordóñez, L. A., & Acosta, C. A. (2013). Pobreza, brechas y ruralidad en Colombia. *Coyuntura económica, Vol. XLIII*, 15-36.

Pineda, J. A., & Acosta, C. E. (2009). *Mercado de trabajo, género y distribución del ingreso en Colombia.* Bogotá: OIT.

Pollak, R. A. (2002). *Gary Becker's Contributions to Family and Households Economics.* St. Louis: Washington University.

Prada, E., Singh, S., Remez, L., & Villarreal, C. (2011). *Embarazo no deseado y aborto inducido en Colombia: causas y consecuencias.* New York: Guttmacher Institute.

Robinson, J. A. (2013). Colombia: Another 100 Years of Solitude? *Current history*.

Rosenzweig, M. R., & Wolpin, K. I. (1980). Testing the Quantity-Quality Fertility Model: The Use of Twins as a Natural Experiment,". *Econometrica XLVIII* , 227–240.

Saad, P. M. (2009). *Demographic trends in Latin America and the Caribbean.* Washington DC: Population Division of the United Nations Economic Commission for Latin America and the Caribbean.

Schultz, T. P. (2005). Effects of Fertility Decline on Family Well Being: Opportunities for Evaluating Population Programs. . *NBER Working Paper No. 14266*.

Universidad de los Andes. (2011). *Colombia en movimiento - Un análisis descriptivo basado en la Encuesta Longitudinal Colombiana de la Universidad de los Andes ELCA.* Bogotá: Universidad de los Andes.

Vogl, T. S. (2016). Differential Fertility, Human Capital, and Development. *Review of Economic Studies 83, ,* 365–401.

World Bank. (2016, May 18). *About LSMS*. Retrieved from The World Bank: http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/EXTLSMS/0,,contentMDK:23506656~pagePK:64168445~piPK:64168309~theSitePK:3358997,00.html

World Bank. (2016, May 28). *El Banco Mundial*. Retrieved from Datos - Colombia: http://datos.bancomundial.org/pais/colombia

Appendix A

Table 35 - Summary of estimations

	ENCV 2010	ENCV 2011	ENCV 2012	ENCV 2013	ELCA (2010 Outcome)	ELCA (2013 Outcome)	ELCA (2013 Outcome, CSS)
Bivariate OLS - Number of siblings	-0.144***	-0.131***	-0.0262	0.0102	-0.0370*	-0.122***	-0.168***
	(0.0356)	(0.0265)	(0.0316)	(0.0340)	(0.0216)	(0.0209)	(0.0365)
Bivariate OLS - Aggregate sibling exposure	0.0297***	0.0285***	0.0385***	0.0403***	0.0309***	0.0250***	0.0392***
	(0.00366)	(0.00274)	(0.00317)	(0.00332)	(0.00231)	(0.00226)	(0.00382)
Multivariate OLS 1 - Number of siblings	-0.125***	-0.139***	-0.107***	-0.103***	-0.0313**	-0.100***	-0.128***
	(0.0176)	(0.0131)	(0.0151)	(0.0156)	(0.0141)	(0.0117)	(0.0198)
Multivariate OLS 2 - Number of siblings	-0.186***	-0.202***	-0.144***	-0.146***	-0.0448**	-0.144***	-0.155***
	(0.0246)	(0.0188)	(0.0221)	(0.0229)	(0.0204)	(0.0168)	(0.0246)
Multivariate OLS 3 - Number of siblings	-0.173***	-0.173***	-0.123***	-0.123***	-0.0401**	-0.130***	-0.149***
	(0.0227)	(0.0170)	(0.0200)	(0.0208)	(0.0186)	(0.0153)	(0.0233)
Multivariate OLS 1 - Sibling exposure	-0.0150***	-0.0161***	-0.0111***	-0.0112***	-0.00407***	-0.0102***	-0.0121***
	(0.00188)	(0.00142)	(0.00160)	(0.00160)	(0.00157)	(0.00132)	(0.00218)
Multivariate OLS 2 - Sibling exposure	-0.0321***	-0.0344***	-0.0213***	-0.0260***	-0.00866***	-0.0197***	-0.0207***
	(0.00316)	(0.00250)	(0.00294)	(0.00307)	(0.00275)	(0.00232)	(0.00349)
Multivariate OLS 3 - Sibling exposure	-0.0266***	-0.0259***	-0.0154***	-0.0187***	-0.00766***	-0.0168***	-0.0188***
	(0.00278)	(0.00214)	(0.00248)	(0.00256)	(0.00239)	(0.00201)	(0.00302)
Bivariate First Stage - Number of siblings explained by multiple births	0.663***	0.104	0.692**	0.952***	0.521	0.667*	0.721*
	(0.244)	(0.255)	(0.342)	(0.298)	(0.343)	(0.377)	(0.400)
Bivariate 2SLS - Multiple birth as instrument on number of siblings	-0.327	-8.837	-2.757	-1.198	2.089	0.812	3.029
	(0.850)	(21.97)	(1.772)	(0.845)	(1.872)	(1.169)	(2.213)
Multivariate First Stage - Number of siblings explained by multiple births	0.522***	0.222	0.574**	0.654***	0.211	0.0588	0.109
	(0.176)	(0.184)	(0.244)	(0.210)	(0.241)	(0.267)	(0.318)
Multivariate 2SLS - Multiple birth as instrument on number of siblings	0.533	-0.819	-0.368	-0.757	2.000	-2.562	2.389
	(0.549)	(1.273)	(0.630)	(0.515)	(2.957)	(12.58)	(8.613)
	5.943**	1.288	2.563	8.887***	7.551**	11.23***	14.52***

Bivariate First Stage - Aggregate sibling exposure explained by multiple births	(2.358)	(2.459)	(3.366)	(3.008)	(3.177)	(3.468)	(3.772)
Bivariate 2SLS - Multiple birth as instrument on aggregate sibling exposure	-0.0364	-0.716	-0.744	-0.128	0.144	0.0482	0.150**
	(0.0975)	(1.484)	(1.072)	(0.0981)	(0.0973)	(0.0620)	(0.0712)
Multivariate First Stage - Aggregate sibling exposure explained by multiple births	4.232***	3.149**	2.529	5.044***	3.788**	3.537*	4.614*
	(1.436)	(1.454)	(1.961)	(1.700)	(1.862)	(2.025)	(2.428)
Multivariate 2SLS - Multiple birth as instrument on Aggregate sibling exposure	0.0658	-0.0576	-0.0835	-0.0982	0.112	-0.0426	0.0563
	(0.0683)	(0.0823)	(0.151)	(0.0669)	(0.117)	(0.102)	(0.111)
First stage (Probit) - Probability of having a third child explained by uniform sex composition of births 1 and 2	0.276***	0.241***	0.168***	0.287***	0.339***	0.316***	0.0139
	(0.0565)	(0.0450)	(0.0513)	(0.0522)	(0.0416)	(0.0425)	(0.0557)
First stage (Probit) - Probability of having a third child explained by uniform sex composition of births 1 and 2 by sex — Male	0.282***	0.316***	0.115	0.288***	0.361***	0.334***	0.0155
	(0.0737)	(0.0588)	(0.0711)	(0.0715)	(0.0551)	(0.0561)	(0.0709)
First stage (Probit) - Probability of having a third child explained by uniform sex composition of births 1 and 2 by sex — Female	0.271***	0.177***	0.208***	0.286***	0.321***	0.300***	0.0125
	(0.0702)	(0.0557)	(0.0623)	(0.0634)	(0.0517)	(0.0530)	(0.0674)
Bivariate 2SLS - Uniform sex composition of births 1 and 2 as instrument on number of siblings	-2.206	-0.415	-3.077	0.552	-1.001*	-0.745**	0.0321
	(1.846)	(0.355)	(3.381)	(0.997)	(0.550)	(0.328)	(0.185)
Multivariate 2SLS - Uniform sex composition of births 1 and 2 as instrument on number of siblings	-2.408	-0.201	-0.476	1.668	-1.164	-0.516	0.0856
	(2.757)	(0.384)	(0.562)	(1.892)	(1.110)	(0.419)	(0.174)
Bivariate 2SLS - Uniform sex composition of births 1 and 2 as instrument of aggregate sibling exposure	-0.207	-0.0481	-0.307	0.0884	-0.0990*	-0.0857**	0.00422
	(0.181)	(0.0428)	(0.360)	(0.157)	(0.0571)	(0.0407)	(0.0243)
Multivariate 2SLS - Uniform sex composition of births 1 and 2 as instrument of aggregate sibling exposure	-0.0729	-0.0162	-0.0279	-0.143	-0.0495	-0.0377	0.0154
	(0.0471)	(0.0309)	(0.0320)	(0.125)	(0.0394)	(0.0296)	(0.0314)
FE,2SLS - Multiple births as instrument of aggregate sibling exposure	0.00348	0.0421***	0.00260	0.0160*	-0.00317	0.0111	0.0156**
	(0.00966)	(0.00872)	(0.00795)	(0.00848)	(0.0102)	(0.00787)	(0.00792)
FE,2SLS - Uniform sex composition of births 1 and 2 as instrument of aggregate sibling exposure	0.0175	-0.0758	-0.211	-0.00466	0.114	-0.0446	-0.0836
	(0.221)	(0.339)	(0.228)	(0.132)	(0.281)	(0.194)	(0.229)
Observations	4,169	6,723	5,392	5,053	7,638	7,759	4,291
Observations (First Stage Probit)	2,518	3,957	3,299	3,198	4,782	4,565	2,220
Observations (Fixed effects)	2,518	3,957	3,299	3,198	4,782	4,565	2,220

Appendix B

Schooling construction procedure

There are four standard variables that provide data on education: i) last approved grade ii) coursing grade iii) highest completed level of schooling and iv) highest reached level of schooling. The first two variables range from grades 0 to 13. Grade 0, or *transition* grade, is the first formal grade in the educational system and the only formal grade in the *pre-school* educational level. The next educational level is *primary* education and encompasses the grades 1 to 5. The level *secondary* education includes grades 6 to 11. Two additional grades (12 and 13) are optional for individuals pursing a career as teachers. After secondary education individual can choose to continue their education by engaging in technical, technological or undergraduate programs. Technical and technological education normaly lasts two years and, according to regulations, undergraduate programs must last at least four years. Graduate education groups specialization, masters and doctorate programs.

Non-missing information in grade specific variables provided the needed information with the highest available precision. If this was not the case then variables about educational level could offer some information though with a lower degree of accuracy. *Highest completed level of schooling* was interpreted as having as much years of completed schooling as they are contained in a specific educational level and all previous levels. For example, an individual with complete secondary has at least approved all the grades in secondary education, from 6^{th} grade to 11^{th} grade or six years of complete schooling, plus all the other minimum requirements of primary and pre-school, from grade 0 to 5^{th} grade or 7 years, for a total of 13 years.

In 2010, %65.32 of the sample had missing data on the variable last approved grade which is the variable that provided the most direct information on schooling. In 2013 it was %46.4. Because these numbers are relatively high and this variable constitutes a core part of the empirical exercise, the following imputation strategy was proposed. If the variable containing the information on last approved grade is missing, then the variable reporting the current grade provides information from which the last approved grade can be inferred for coursing a particular grade implies the approval of the previous one. If this information is also missing, the 2013 wave provides retrospective information on what was the last approved grade in 2010. This step only affects imputation of schooling in 2010 and, because it involves recognition and all its associated difficulties, current grade was perceived as displaying a higher degree of confidence.

The next step, if missing information persists, involves obtaining a lower boundary for schooling through educational level information. If an individual reports a certain completed level this means it has completed all the grades it contains therefore providing a minimum quantity of approved grades. This inference was divided into two main steps and a third step in 2010 because there was a question in the 2013 wave involving educational level recognition. Current educational level can provide two outcomes if it contains non missing information; if some individual is coursing high-school it must have finished primary school and all the grades within (6 grades) and if some individual is coursing any form of technical and superior education it must have approved high-school and primary school (12 grades). Completed educational levels, the next step, provide a similar process of inference. Finally, retrospective

information on completed educational levels on 2010 serves as the last stage in estimating schooling.