ASOCIACION ARGENTINA
DE ECONOMIA POLITICA

ANALES | ASOCIACION ARGENTINA DE ECONOMIA POLITICA

LIII Reunión Anual

Noviembre de 2018

ISSN 1852-0022
ISBN 978-987-28590-6-0

Exploring peer effects in education in Latin America and the Caribbean

Di Capua, Laura
Izaguirre, Alejandro

Exploring peer effects in education in Latin America and the Caribbean.

Working paper

Laura Di Capua[*] Alejandro Izaguirre[†]

30th August 2018

Abstract

This paper assesses peer group influence on academic performance of primary school students in Latin America and the Caribbean. Based on TERCE data set, we investigate peer effects in mathematics and language tests outcomes among sixth grade students. We apply the model proposed in Lee (2007), which allows to identify endogenous and exogenous peer effects while controlling for group-level unobservables. The estimates suggest the existence of endogenous peer effects both, in mathematics and language tests scores, implying that peer's outcomes do influence student's academic results.

Key words: *Peer effects, group interaction, school performance.*

JEL classification: ***C31, I21***

[*] Doctoral student at Universidad Nacional de Rosario (UNR). Teacher and Researcher of Instituto de Investigaciones Económicas, Facultad de Ciencias Económicas y Estadística, Universidad Nacional de Rosario. E-mail:ldicapua@fcecon.unr.edu.ar

[†] Doctoral student at Universidad de San Andrés, Buenos Aires, Argentina. Teacher and Researcher of Instituto de Investigaciones Económicas, Facultad de Ciencias Económicas y Estadística, Universidad Nacional de Rosario. E-mail: izaguirre.ale@gmail.com

1 Introduction

Social scientists have long been interested in peer effects because of their far reaching implications at the individual and collective level. These non-market interactions represent how an individual's decision or outcome is directly influenced by his peer's outcome or characteristics. The sociological literature has placed great emphasis on the importance of social interactions arguing that they play an important role in determining behavioral and economic outcomes. In fact, a number of theoretical approaches such as collective socialization theories, contagion-based or epidemic theories and information asymmetries and network theories (Andrews et al., 2002) have been developed to account for contextual influence on individual's outcomes and behaviors regarding diverse aspects of life (such as criminal activity, use of public services, labor markets outcomes, etc.).

Among the various spheres in which peer effects may manifest themselves, the school context is especially important considering the vital role educational attainments have on future living conditions of individuals. Human capital accumulation has intertemporal repercussions given the proven relationship between years of schooling and labor incomes (Mincer, 1974; Becker, 1994). The analysis of peer effects in education has received considerable attention, notably since the publication of the Coleman report (Coleman et al., 1966). A common hypothesis is that student outcomes are higher in the presence of favourable peer groups, conditional on individual characteristics and family background (McEwan, 2003).

Evaluating peer effects in academic achievement is important for parents, teachers and schools; but crucially from a public policy perspective. A major question in the economic literature is whether or not interactions among students lead to large social multipliers (Epple & Romano, 1996). Depending on the nature of peer effects, there may be social gains from their existence (Hoxby, 2000). Furthermore, many researchers have studied the relative importance of peer effects in students academic performance versus the influence of other factors such as school infrastructure and teachers qualifications (Hanushek et al., 1998; Greene et al., 1999). As a matter of fact, peer effects have played a prominent role in educational policy debates concerning ability grouping, racial integration and school vouchers (Sacerdote, 2001; Gaviria & Raphael, 2001; Lin, 2005).

In this paper we analyse the possible existence of peer effects in educational achievements among sixth grade students participating in the Third Regional Comparative and Explanatory Study

(TERCE) conducted by United Nations Educational, Scientific and Cultural Organization (UNESCO). Since this survey focuses on primary school students, TERCE data provides a unique opportunity to explore peer effects in education in its early stages. Given the fact that primary education is a phase in which public policy can make a difference for students coming from vulnerable contexts, a better understanding of the educational production function shall improve equity in the education system. The latter has much relevance taking into consideration early education's welfare implications for future living standards of individuals and their families. Therefore, a deep understanding of the nature and characteristics of peer effects in education is not only central for educational policies but also for general policies targeting at social inequality.

One important difficulty in dealing with peer effects is that they are hard to identify with observational data since it is not easy to distinguish between the impacts that actually result from social interactions from the choices of with whom to interact with [1] and the existence of a common environment among group members (Manski, 1993). For this reason, disparities in educational attainments may actually reflect children and families with similar characteristics sorting together at the school level or facing similar exogenous factors. Consequently, divergence in academic performance of students could in fact reflect broader inequalities in the economy and thus policy implications differ greatly. To deal with these problems, recent developments in network literature allow to study outcomes of social interactions taking into consideration the problems caused by endogenous association of members within a group and cofounding factors (Moffitt et al., 2001; Bramoullé et al., 2009; Lee, 2007).

With this research we expect to contribute to the recent empirical literature on peer effects in education. Besides, this paper should specifically add to the scarce existing evidence on the magnitude and characteristics of peer effects in education in Latin America and the Caribbean. The article will explore personal, family and contextual factors associated with mathematics and language learning achievements for sixth grade students of those countries participating in TERCE. As this survey was applied in fifteen countries in the region, the data provides a general perspective of this subject in Latin American and Caribbean countries.

The paper is organized as follows. Section 2 reviews the existing literature on peer effects in education. Section 3 presents the methodological approach and econometric model used for estimations.

[1] This refers to selection into peer groups based on common unobserved characteristics (homophily).

The following section describes the data and variables used in the analysis, and explains how we dealt with missing observations. Finally, section 5 shows estimation results while conclusions are provided in Section 6. The article also includes an annex where an in depth analysis of the magnitude and incidence of the missing data problem is addressed.

2 Literature review

The problem of heterogeneity of results of the education process, that manifests itself in significant differences in academic performance or achievements of students, has long attracted considerable attention in the economic literature (Hanushek, 1979; Burgess, 2016). In this line of research, the influence of peers on educational outcomes has been extensively studied. The milestone in this field is the 1966's Equality of Educational Opportunity Report (Coleman et al., 1966), known as Coleman report for its director. This report pushed peer effects into the limelight when concluding *'finally, it appears that a pupil's achievement is strongly related to the educational backgrounds and aspirations of the other students in the school'* (Coleman et al., 1966, pg. 22). Since this research, the empirical literature on peer effects has grown (Sacerdote, 2001; Hanushek et al., 2003; Angrist & Lang, 2004; Stinebrickner & Stinebrickner, 2006; Ammermueller & Pischke, 2009). However, the evidence regarding the magnitude of peer effects on student's achievement is far from conclusive.

The aforementioned lack of consensus partly reflects various econometric issues that any empirical study on peer effects must address. Trying to explain the common observation that people belonging to the same group tend to behave similarly, in a pioneer study Manski (1993) differentiates three kinds of social effects: *endogenous effects*, wherein the propensity of an individual to behave in some way varies with the behaviour of the group; *exogenous (contextual) effects*, wherein the propensity of an individual to behave in some way varies with the exogenous characteristics of the group; *correlated effects*, wherein individuals in the same group tend to behave similarly because they have similar individual characteristics or face similar institutional environments.

Distinguishing between endogenous and exogenous effects is important because they have different implications for policy interventions. Endogenous effects may give rise to bidirectional influences and consequently to the possibility of social multipliers, while the repercussion of exogenous effects does not necessarily imply amplified responses to exogenous shocks (Gaviria & Raphael, 2001). As regards correlated effects, they arise when students in the same reference group achieve similar educational outcomes because they share a common set of unobserved characteristics. In this case it could imply that families send their children to the same schools according to their willingness and ability to pay for better peer influences, which are unobservable (Gaviria & Raphael, 2001).

Researchers have used various approaches to solve these issues, but there is no simple methodological answer to face the existing challenges (Calvó-Armengol et al., 2009). Manski (1993) shows that endogenous and exogenous effects cannot be separately identify in a linear-in-means model[2] due to the *reflection problem*. Thus by using this kind of econometric models only aggregate parameters are estimated (Sacerdote, 2001; Ammermueller & Pischke, 2009). Many empirical studies have addressed this issue imposing alternative structures or excluding effects on the original model. As another strategy, some use instruments to obtain consistent estimates of the endogenous peer effect (Evans et al., 1992; Gaviria & Raphael, 2001; Atkinson et al., 2008). The key here is the suitable choice of those variables which are correlated with the endogenous peer effect but not correlated with the error terms in the model.

With respect to correlated effects, some studies explicitly account for this source of bias. Researchers have used three main strategies to handle this problem. They have either exploited data where group members are randomly or quasi-randomly assigned within their groups (Angrist & Lavy, 1999; Boozer & Cacciola, 2001; Sacerdote, 2001; Zimmerman, 2003; Kang et al., 2007), they have used an instrumental variable strategy (Evans et al., 1992; Rivkin, 2001), or a family fixed effect strategy (Aaronson, 1998; Plotnick & Hoffman, 1999). Bramoullé et al. (2009) consider an extended version of the linear-in-mean model where interactions are structured through a social network allowing the existence of correlated effects. By doing so they provide necessary and sufficient conditions for identification; such conditions generalize a number of previous results due to Manski (1993), Moffitt et al. (2001) and Lee (2007).

In Lee et al. (2010) the original model is extended to consider network structures and correlated disturbances among connected individuals. The possible endogeneity of the network is a particular concern in settings where peer effects hypothetically raise from networks that are formed by individuals

[2] In the linear in means model, the outcome of each individual depends linearly on his own characteristics, on the mean outcome of his reference group and on its mean characteristics.

making choices to establish links, because such endogenity may bias estimates. Goldsmith-Pinkham & Imbens (2013) and Hsieh & Lee (2016) propose correcting this selection bias by modelling the endogenous network formation process.

Considering the fact that the model specified in Lee (2007) adequately deals with the above mentioned difficulties, it has been used as reference in various empirical researches (Lin, 2010; Lee et al., 2010; Boucher et al., 2014), especially when studying peer influences in the school context. Therefore, unlike various strategies proposed to address the basic issues affecting peer effects estimations, the one developed by Lee (2007) has the advantage of fully identifying peer effects not requiring panel data or strong assumptions that are difficult to motivate and may not hold in practice (Boucher et al., 2014).

Finally, another source of bias in empirical research comes from the determination of reference groups. The choice of reference groups is often severely constrained by the availability of data. Consequently many studies of peer effects in education focus either on the grade-within-school level (Hoxby, 2000; Hanushek et al., 2003; Angrist & Lang, 2004), or analyse peer effects at the classroom level (Kang et al., 2007; Burke & Sass, 2008; Atkinson et al., 2008; Ammermueller & Pischke, 2009). The data set used in this research does not provide information on students social networks, but allows estimations at the classroom level.

This paper advances the literature on peer effects in education in Latin America and the Caribbean, providing the first application based on Lee (2007). Although there are a few other works that analyses peer effects in the region (McEwan, 2003; Dieye et al., 2014; De Melo, 2014; Mariño Fages, 2015), they do not use the same methodological approach. This social interaction model proposed in Lee (2007) considers group interaction and the existence of the three effects mentioned above (e.i. endogenous, exogenous and correlated effects).

3 Methodological approach and Econometric model

As mentioned previously, the model considered in this paper is the one proposed in Lee (2007), such a model relies in two key assumptions. First, individuals interact in groups that are known for the modeler. Under our setting these groups are formed by classmates, the students are affected by all others in their groups (classrooms) but by none outside it. Second, individual outcome is determined by a linear-in-means model with group fixed effects. Thus, the test score of a student is affected by his characteristics and by the average test score and characteristics in his group of peers. In addition, it may be affected by any kind of correlated group-level unobservables.

Suppose there are R groups and there are m_r units in the rth group. At group level, the structural model is given by

$$Y_r \quad = \quad \lambda_0 W_r Y_r + X_{r1}\beta_{r1} + W_r X_{r2} + I_{m_r}\alpha_r + e_r, \quad r = 1, ..., R,$$

with $W_r = \dfrac{1}{m_r - 1}(l_{m_r} l'_{m_r} - I_{m_r})$ where l_{m_r} is the m_r-dimensional vector of ones, and I_{m_r} is the m_r-dimensional identity matrix. Y_r, X_{r1}, X_{r2} are the vector and matrices of the m_r observations in the rth group, or equivalently in term of each unit i in a group r.

$$y_{ri} \quad = \quad \lambda_0\Big(\frac{1}{m_r - 1}\sum_{j=1,j\neq i}^{m_r} y_{rj}\Big) + x_{ri,1}\beta_{10} + \Big(\frac{1}{m_r - 1}\sum_{j=1,j\neq i}^{m_r} x_{rj,2}\Big)\beta_{20} + \alpha_r + e_{ri},$$

with $i = 1, ..., m_r$, and $r = 1, ..., R$, where y_{ri} is the ith individual in the rth group, $x_{ri,1}$ and $x_{rj,2}$ are, respectively, k_1 and k_2-dimensional row vectors of exogenous variables, and e_{ri} are i.i.d $N(0, \sigma_0)$. In the model, the outcome of the unit i may be influenced by outcomes of other units, which effect is captured by parameter λ_0. The α_r represent the unobservables of the r-th group. As those unonbservables may correlate with exogenous variables, they are treated as fixed effects. The vector of all exogenous variables x_{ri}'s must vary across individuals in a group, as any group invariant variables will be captured in α_r. As we can see in the summations, a student is not assumed to be one of his own peer, this creates individual variations in average peer attributes. These variations survive the elimination of common unobservables.

In a general setting, $x_{ri,1}$ and $x_{rj,2}$ are subvector of x_{ri}, which may or may not have common elements. The introduced variables $\sum_{j=1,j\neq i}^{m_r} x_{rj,2}$ allow social interaction effects through observed neighborhood characteristics. Lee (2007) proposes two ways to estimate the model, generalized two-stage least squares (G2SLS) and conditional maximum likelihood (CML), and shows that the identification of endogenous and exogenous effects is possible if there are sufficient variation in group size in the sample [3]. The identification, however, can be weak if the size of all groups are large. The model assumes that W_r is exogenous conditional on the unobserved effect α_r[4], i.e. $E(e_{ri}|x_{ri}, W_r, \alpha_r)$. This assumption can accommodate many situations where W_r is endogenous, suppose W_r depends on unobserved common characteristics of the student's group (i.e. their preferences for sports, for physical infrastructure, and so on), the model admits this kind of correlation. This assumption fails to holds, for instance, if some unobserved characteristics affect both the likelihood to be in the group (classroom) and the outcome, and *differs* among individuals in the same group.

4 Data

4.1 Third Regional Comparative and Explanatory Study (TERCE)

In recent years, quantitative research on students outcomes in Latin America and the Caribbean has benefited a lot from the growing availability of international comparable data. The Third Regional Comparative and Explanatory Study (TERCE) is an example of this kind of data source. Implemented in 2013 by UNESCO, TERCE is a large scale study of learning achievements carried out in 15 countries: Argentina, Brazil, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru and Uruguay, as well as in the Mexican state of Nuevo León. Its main goals are to provide information for the discussion on educational quality in the region and to orientate decision making in public policies. TERCE is the third study of its kind in primary education conducted by UNESCO Regional Bureau of Education for Latin America and the Caribbean, preceded in 1997 by the First Regional Comparative and Explanatory Study (PERCE) and in 2006 by the Second Regional Comparative and Explanatory Study (SERCE).

TERCE assessed the performance of pupils in third and sixth grades primary school in Mathematics and Language (reading and writing skills); and students achievements in Natural Sciences, in the case of sixth grade. In order to measure learning achievements, the study applied tests regarding

[3] In this first version of the article we only estimate models using CML method, but expect to use both approaches in a future version of this research.

[4] Under group interaction assumption all students in a classroom are peers, so the conditional exogeneity of W_r is equivalent to the conditional exogeneity of the group size, m_r.

common elements of the school curricula in the region. To assure cultural adaptation to each country and to prevent from imposing foreign standards, the design and implementation of the study was done following a collaborative process with participating countries (Flotts et al., 2015).

In addition to students academic performance, context questionnaires aiming to collect information on associated factors that influence student's learning achievements were also implemented. Among the variables considered in these questionnaires, importance was given to socio-economic context, family life and personal issues, as well as educational policies and school processes. Therefore, the study also collected data on the characteristics of students and their families, teachers, the school and its principal.

The TERCE data base consists of $N_T = 67,582$ observations on students which are grouped in $R_T = 3,115$ classrooms along the 15 countries and the state of Nuevo León[5].

In table (1) we present the total number of classrooms and the quartiles of the classroom sizes distribution by country.

Table 1: Classrooms and sizes. Original data.

Country	Number of classrooms	Quartiles of classroom sizes		
		Quartile 1	Quartile 2	Quartile 3
Argentina	207	14	20	26
Brasil	126	21	29	34
Chile	197	20	28	35
Colombia	149	23	31	36
Costa Rica	197	12	19	24
Dominicana	170	13	22	30
Ecuador	210	16	26	35
Guatemala	232	14	22	31
Honduras	203	10	18	28
Mexico	168	14	23	30
Nicaragua	180	9	22	31
Panama	187	15	20	26
Paraguay	205	10	17	25
Peru	285	7	16	25
Uruguay	238	12	19	24
Nuevo Leon	161	21	27	35

[5] For an in depth description of TERCE's sample design and survey's contents refer to (Flotts et al., 2015).

4.2 Variables

To analyse students learning achievements, dependent variables used are individual results on students mathematics and language tests[6]:

Score_math: irt standardized mathematics score.

Score_lang: irt standardized language score.

As regards explanatory variables, individual characteristics, family background and peer's influence were taken into account. Following the literature (Sacerdote, 2001; Gaviria & Raphael, 2001; Lin, 2005, 2010; Lee et al., 2010; Boucher et al., 2014), we consider these variables:

Isecf: standardized index of the economic, material and sociocultural condition of the student's household. This index is directly estimated by UNESCO, and to construct it information on the mother's education level and occupation, as well as household income and good and services available at the house is collected.

Mothereduc: highest education level of the mother. This is a categorical variable using UNESCO's International Standardized Education Classificator (CINE-P, for its acronym in Spanish), that takes the following values:

- 1 Without studies
- 2 Primary school/Low secondary school [cine-p 1-2]
- 3 High secondary school [cine-p 3]
- 4 Post secondary education/Tertiary education [cine-p 4-5]
- 5 University [cine-p 6]
- 6 Master degree/Ph.D. [cine-p 7-8]

Age: student age measured in years.

Gender: dummy variable taking value one if the student is male and zero if female.

[6] Estimated as the standardized score following the Item Response Theory (see Flotts et al. (2015) for a thorough explanation on how this scores are calculated).

Indigenous: dummy variable taking value one if at least one of these conditions is met and zero otherwise:

- the mother or father self-define themselves as indigenous
- at least one of the parents speaks an indigenous language
- parents speak in an indigenous language to the student
- the student self-defines him or herself as indigenous
- the student speaks in an indigenous language

Contextual effects: average values of all the explanatory variables over the student's classmates.

Endogenous effects: average result in tests of the student's classmates.

The following Table shows basic statistical measures for all the variables considered above.

Table 2: Descriptive statistics.

Variable	Mean	S.D.
Score_math	712.3	105.3
Score_lang	711.3	103.0
Age	12.41	0.940
Gender	0.503	0.499
Indigenous	0.234	0.423
Mothereduc	2.898	1.223
Isecf	0.142	1.047

4.3 Missing data treatment

As it happens in most surveys, many observations present missing data in some variables. The percentage of missing values in the total sample of sixth grade students is 5% for language score, 4% for math score, 15% for isecf index and 23% for mothereduc indicator.

There are several methods to deal with missing data in the literature (Little, 1992; Pigott, 2001; Enders, 2010). Their accuracy depends crucially on the assumptions about the missing data mechanisms generating it. Here we combine two commonly used methods to face this problem. For missing values in explained variables we apply complete cases method, which consists of only using observations for which we have the value for the explained variable (language or math scores). In the case of missing data in explanatory variables, we impute it using random imputation.

With respect to only using observations for which the explained variable is defined, while this is an accurate method when the mechanism generating the missing data is random, in models where the explained variable is also used as explanatory variable it is not advisable to use it[7]. Given the fact that in this case overall missing data in explained variables is relatively small (5 and 4 %), we expect that any bias that could be introduced in estimates by using complete observations (complete on explained variables) shall be negligible.

As regards missing values in explanatory variables (isecf index and mothereduc indicator), as mentioned above, we impute missing data using random imputation. Supposing that the missing problem is confined to a single variable, y, and that we observe a set of variables X for all units, then the method consists in estimating a regression model based on observed data. As we know all X, we impute the missing y using the estimated model.

Let y^o and y^u be the observed and unobserved y respectively, we estimate $y^o = \beta X^o + e^o$, where $e \sim N(0, \sigma_e)$, and then we impute the missing y by $\hat{y}^u = \hat{\beta} X^u + \hat{e}^u$ (consider we completely observe X). It is worth noting that we add an error term, $\hat{e}^u$, to the imputed values $\hat{y}^u$ (hence the name *random imputation*), which is generated by simulating their distribution, $\hat{e}^u \sim N(0, \hat{\sigma}_e)$.

The model we use to impute isecf index and mothereduc indicator when these variables show missing data has the following structure,

$$y_{ir} = \beta_1 \overline{y}_{-ir} + \beta_2 x_{2,ir} + \beta_3 x_{3,ir} + e_{ir}$$

where y_{ir} is the y value (isecf index or mothereduc indicator) for the $i-th$ student in $r-th$ classroom, $\overline{y}_{-ir}$ is the mean of y in classroom r without considering y_i, $x_{2,ir}$ is the mean of isecf index and it is present in both model, whereas $x_{3,ir}$ is the kind of school (public or private) in the model for isecf index, and the level of neighbour violence in the model for mothereduc indicator, $e_{ir} \sim N(0, \sigma_e)$ is an error term.[8] The intra-classroom autocorrelation of isecf index and mothereduc indicator is relatively high. That is why we use $\overline{y}_{-ir}$ as explanatory variable. However, this triggers another issue because the variable $\overline{y}_{-ir}$ is a classroom mean of the partially observed variable y, so it is also partially observed. We ignore this fact because the goal here is not causal inference but simply

[7] The model proposed in this paper has an important link with spatial economnetric models, in particular with the Saptial Lag Model. The treatment of missing data under this model has some particular issues, see Wang & Lee (2013), LeSage & Pace (2004) and Kelejian & Prucha (2010) for details.

[8] We have selected the explanatory variables in order to maximize the R^2.

accurate prediction. Therefore it is acceptable to use any input in the imputation model to achieve this goal, and given $\overline{y}_{-ir}$ is helpful for explaining y, we consider it in the model.

As mentioned above, to estimate the models we only use cases in which we observe the explained variable. Besides, we also dismiss all observations from classrooms where the percentage of missing values in any variable (explained or explanatory) exceeds 50%; and those cases where there is only one student in the classroom. Furthermore, as the neighbourhood violence level is one of the variable we use to impute the mothereduc indicator and as this variable has missing values for a few classrooms, we drop the observations from such classrooms.

Finally, as observations with missing values in math score differ from those with missing values in language score, the final data base used for each subject differs. The final data bases for both mathematics and language consist in nearly 90% of the students and the classrooms from the original sample[9].

[9] See annex 8.1 and 8.2 for more details on missing data.

5 Empirical results

Table 3: Maximum Likelihood Estimation. 6th grade

	Mathematics	Language
Endogenous Effects	0.329***	0.104*
	(0.045)	(0.054)
Individual Effects		
Isecf	13.01***	12.88***
	(0.870)	(0.873)
Age	-7.89***	-9.36***
	(0.600)	(0.681)
Mothereduc	3.13***	5.50***
	(0.721)	(0.631)
Gender	12.39***	-8.62***
	(0.955)	(1.044)
Indigenous	-2.73**	-7.27***
	(1.169)	(1.203)
Contextual Effects		
Isecf	15.03	- 11.77
	(14.21)	(14.83)
Age	12.06	13.86
	(9.060)	(11.73)
Mothereduc	-21.52	-10.26
	(13.85)	(11.04)
Gender	24.11	20.51
	(15.62)	(18.38)
Indigenous	29.83	-26.53
	(18.70)	(21.68)
Corr($\hat{y}, y$)	0.344	0.368

Notes: Standard Errors in parenthesis.
*** indicates 1% significance level.
** indicates 5% significance level.
* indicates 10% significance level.

Table (3) displays estimates of the proposed model for student's mathematics and language academic outcomes. As explained above, this methodological approach allows to account for the incidence of endogenous effects (i.e., the influence of peer outcomes), student's individual characteristics and contextual or exogenous effects (i.e., the influence of exogenous peer characteristics); while filtering fixed effects at the group level. These fixed effects do not only include observable characteristics of the group (such as country of residence, school infrastructure, teacher's qualifications, etc.) but also unobservables, as well as, common shocks faced by the group. Endogenous peer effect estimates are listed at the top of table (3) and contextual effects at the bottom.

Before analysing the social-interaction effects, a brief discussion of the performance of the control variables is necessary. Concerning personal background controls, i.e. student's age, gender and ethnicity, they are all statistically significant in determining academic performance. Student's age is negatively related to academic achievements, both in mathematics and language outcomes. This variable may be an indirect indicator of grade repetition among students and consequently could be reflecting individual difficulties in school performance.

As regards gender, results found are consistent with the empirical literature (Hyde et al., 1990). Female students tend to outperform males in language, while males students achieve better results than females in math tests. Turning to ethnicity, results indicate that students with indigenous influence in general achieve poorer academic results than the rest, which is in line with previous research (Verdisco et al., 2009). Furthermore, this disadvantage seems stronger when it comes to language outcomes possibly indicating that indigenous children suffer from idiomatic limitations that condition their academic achievements (Flotts et al., 2015). Finally, family sociocultural and economic condition as well as mother education, both have positive significant influence in student's academic achievements, reinforcing existing findings on these topics (Davis-Kean, 2005).

With respect to those effects that surge from social interaction, exogenous peer characteristics do not significantly influence student's academic outcomes, while endogenous peer effects do. Endogenous peer effects in math scores are highly significant and somehow stronger than in language. Nevertheless, peer outcomes also impact language tests' results at 10% significance. It is worth to note that the model does not explain much of the variability of the data[10] suggesting the existence of other factors that may explain student's academic performance besides those explicitly considered here. Even so, it is clear that classmates academic outcomes do affect student's performance at school and therefore attention should be paid to these findings.

6 Concluding remarks

The results found in this research add empirical evidence that supports the hypothesis of peer effects in education, affecting in this particular application primary school attendants in Latin America and the Caribbean. Hopefully, this paper has contributed to a better visualization of the impacts of social interactions in human capital accumulation. We show that peer influence plays a significant role in early education academic achievements, mainly through endogenous effects. Furthermore it seems that these peer effects have different magnitudes according to the subject, being more important in mathematics than language. These findings may add new inputs to be considered in the educational policy agenda of the region. Undoubtedly, the issues regarding the accumulation of human capital are sure to remain a fertile ground for future research. In fact,

[10] $\text{Corr}(\hat{y}, y)$ are both less than 0.4.

we expect to extend this research to third grade pupils also evaluated in TERCE so as to achieve a more precise picture of peer effects influence on students academic performance in the region.

7 References

References

Aaronson, D. (1998). Using sibling data to estimate the impact of neighborhoods on children's educational outcomes. *Journal of Human Resources*, (pp. 915–946).

Ammermueller, A. & Pischke, J.-S. (2009). Peer effects in european primary schools: Evidence from the progress in international reading literacy study. *Journal of Labor Economics*, 27(3), 315–348.

Andrews, D., Green, C., & Mangan, J. (2002). Neighbourhood effects and community spillovers in the australian youth labour market. *LSAY Research Reports*, (pp.2̃8).

Angrist, J. D. & Lang, K. (2004). Does school integration generate peer effects? evidence from boston's metco program. *American Economic Review*, 94(5), 1613–1634.

Angrist, J. D. & Lavy, V. (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, 114(2), 533–575.

Atkinson, A., Burgess, S., Gregg, P., Propper, C., Proud, S., et al. (2008). The impact of classroom peer groups on pupil gcse results. *Centre for Market and Public Organiziation Working Paper*, 8, 187.

Becker, G. S. (1994). Human capital revisited. In *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education (3rd Edition)* (pp. 15–28). The university of Chicago press.

Boozer, M. & Cacciola, S. E. (2001). Inside the'black box'of project star: Estimation of peer effects using experimental data.

Boucher, V., Bramoullé, Y., Djebbari, H., & Fortin, B. (2014). Do peers affect student achievement? evidence from canada using group size variation. *Journal of applied econometrics*, 29(1), 91–109.

Bramoullé, Y., Djebbari, H., & Fortin, B. (2009). Identification of peer effects through social networks. *Journal of econometrics*, 150(1), 41–55.

Burgess, S. M. (2016). Human capital and education: The state of the art in the economics of education.

Burke, M. & Sass, T. (2008). Vclassroom peer effects and student achieve'mentv. *Federal Reserve Bank of Boston Working Paper*, (08), 5.

Calvó-Armengol, A., Patacchini, E., & Zenou, Y. (2009). Peer effects and social networks in education. *The Review of Economic Studies*, 76(4), 1239–1267.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & Robert, L. (1966). York. 1966. *Equality of educational opportunity*, 2.

Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: the indirect role of parental expectations and the home environment. *Journal of family psychology*, 19(2), 294.

De Melo, G. (2014). *Peer effects identified through social networks: Evidence from Uruguayan schools.* Technical report, Working Papers, Banco de México.

Dieye, R., Djebbari, H., & Barrera-Osorio, F. (2014). Accounting for peer effects in treatment response.

Enders, C. K. (2010). *Applied missing data analysis.* Guilford press.

Epple, D. & Romano, R. E. (1996). Public provision of private goods. *Journal of political Economy*, 104(1), 57–84.

Evans, W. N., Oates, W. E., & Schwab, R. M. (1992). Measuring peer group effects: A study of teenage behavior. *Journal of Political Economy*, 100(5), 966–991.

Flotts, M. P., Manzi, J., Jiménez, D., Abarzúa, A., Cayuman, C., & García, M. J. (2015). *Informe de resultados TERCE: logros de aprendizaje.* UNESCO Publishing.

Gaviria, A. & Raphael, S. (2001). School-based peer effects and juvenile behavior. *Review of Economics and Statistics*, 83(2), 257–268.

Goldsmith-Pinkham, P. & Imbens, G. W. (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics*, 31(3), 253–264.

Greene, J. P., Peterson, P. E., & Du, J. (1999). Effectiveness of school choice: The milwaukee experiment. *Education and Urban Society*, 31(2), 190–213.

Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *Journal of human Resources*, (pp. 351–388).

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of applied econometrics*, 18(5), 527–544.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (1998). *Does special education raise academic achievement for students with disabilities?* Technical report, National Bureau of Economic Research.

Hoxby, C. (2000). *Peer effects in the classroom: Learning from gender and race variation.* Technical report, National Bureau of Economic Research.

Hsieh, C.-S. & Lee, L. F. (2016). A social interactions model with endogenous friendship formation and selectivity. *Journal of Applied Econometrics*, 31(2), 301–319.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: a meta-analysis. *Psychological bulletin*, 107(2), 139.

Kang, C. et al. (2007). Does money matter? the effect of private educational expenditures on academic performance. *National University of Singapore. Department of Economics Working Paper*, 704.

Kelejian, H. H. & Prucha, I. R. (2010). Spatial models with spatially lagged dependent variables and incomplete data. *Journal of geographical systems*, 12(3), 241–257.

Lee, L.-f. (2007). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *Journal of Econometrics*, 140(2), 333–374.

Lee, L.-f., Liu, X., & Lin, X. (2010). Specification and estimation of social interaction models with network structures. *The Econometrics Journal*, 13(2), 145–176.

LeSage, J. P. & Pace, R. K. (2004). Models for spatially dependent missing data. *The Journal of Real Estate Finance and Economics*, 29(2), 233–254.

Lin, X. (2005). Peer effects and student academic achievement: an application of spatial autoregressive model with group unobservables. *Unpublished manuscript, Ohio State University.*

Lin, X. (2010). Identifying peer effects in student academic achievement by spatial autoregressive models with group unobservables. *Journal of Labor Economics*, 28(4), 825–860.

Little, R. J. (1992). Regression with missing x's: a review. *Journal of the American Statistical Association*, 87(420), 1227–1237.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3), 531–542.

Mariño Fages, D. (2015). Efecto de pares en el desempeño académico de alumnos de primaria y secundaria. *L Reunión Anual Asoción Argentina de Economía Política.*

McEwan, P. J. (2003). Peer effects on student achievement: Evidence from chile. *Economics of education review*, 22(2), 131–141.

Mincer, J. (1974). Schooling, experience, and earnings. human behavior & social institutions no. 2.

Moffitt, R. A. et al. (2001). Policy interventions, low-level equilibria, and social interactions. *Social dynamics*, 4(45-82), 6–17.

Pigott, T. D. (2001). A review of methods for missing data. *Educational research and evaluation*, 7(4), 353–383.

Plotnick, R. D. & Hoffman, S. D. (1999). The effect of neighborhood characteristics on young adult outcomes: Alternative estimates. *Social Science Quarterly*, (pp. 1–18).

Rivkin, S. G. (2001). Tiebout sorting, aggregation and the estimation of peer group effects. *Economics of Education Review*, 20(3), 201–209.

Sacerdote, B. (2001). Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly journal of economics*, 116(2), 681–704.

Stinebrickner, R. & Stinebrickner, T. R. (2006). What can be learned about peer effects using college roommates? evidence from new survey data and students from disadvantaged backgrounds. *Journal of public Economics*, 90(8-9), 1435–1454.

Verdisco, A., Cueto, S., Thompson, J., Engle, P., Neuschmidt, O., Meyer, S., González, E., Oré, B., Hepworth, K., & Miranda, A. (2009). Urgency and possibility results of pridi a first initiative to create regionally comparative data on child development in four latin american countries technical annex.Ť. *Technical Annex. Inter-American Development Bank, Washington DC.*

Wang, W. & Lee, L.-F. (2013). Estimation of spatial autoregressive models with randomly missing data in the dependent variable. *The Econometrics Journal*, 16(1), 73–102.

Zimmerman, D. J. (2003). Peer effects in academic outcomes: Evidence from a natural experiment. *Review of Economics and statistics*, 85(1), 9–23.

8 Annex

8.1 Missing data descriptions

As we mentioned, the percentage of missing values in the total sample of sixth grade students is 5% for language score, 4% for math score, 15% for isecf index and 23% for mothereduc indicator. The overall percentage of *classrooms* with at least one missing value in language score, math score, isecf index and mothereduc indicator are 44%, 36%, 60% and 86% respectively. Despite the fact that the number of classrooms with missing data is high (specially for explanatory variables), the percentage of missing values within classrooms is considerable lower. In fact, the 80% of classrooms with missing data in language and math scores do not have more than 8% and 6% of missing values respectively; whereas the 80% of classrooms with missing values of isecf index and mothereduc indicator do not exceed 20% and 33% of missing values respectively.

The aforementioned information on missing data concerns the sample as a whole, but the proportion of missing values varies considerably between countries, classrooms and variables. To get some insights in the distribution of missing values we calculate both, the percentage of missing values by country and the distribution of the percentage of missing values by classroom in each country. We report the 8th quantile of such distributions.

Table 4: Missing data by country and classrooms in explained and explanatory variables.

Country	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Argentina	64	17	66	16	89	38	95	46
Brazil	95	25	97	30	98	40	99	53
Chile	50	6	51	5	76	22	90	29
Colombia	46	5	34	3	47	8	91	16
Costa Rica	13	0	23	3	23	4	76	15
Dominicana	17	0	36	5	43	7	92	25
Ecuador	20	2	16	0	43	9	80	20
Guatemala	23	3	41	8	70	100	88	100
Honduras	16	0	38	6	44	11	80	28
Mexico	39	4	45	7	60	15	82	22
Nicaragua	36	5	68	15	68	20	86	33
Panama	63	16	62	12	76	26	92	37
Paraguay	37	7	46	10	65	22	91	40
Peru	14	0	24	4	38	10	73	22
Uruguay	32	5	36	6	70	100	86	100
Nuevo Leon	42	5	53	6	66	13	89	18

(1) % Classrooms with missing data in Math score. (2) % Missing data in Math score by classrooms, Q_8 . (3) % Classrooms with missing data in Language score. (4) % Missing data in Language score by classrooms, Q_8. (5) % Classrooms with missing data in isecf index. (6) % Missing data in isecf index by classrooms, Q_8. (7) % Classrooms with missing data in mothereduc indicator . (8)% Missing data in mothereduc indicator by classrooms, Q_8.

The percentage of classrooms with missing data in explained variables shows wide variability when measured by country. Regarding math score it ranges from 16 to 95%, while for language score it goes from

16 to 97 %. In almost every country the percentage of classrooms with missing values in language score is slightly grater than the percentage of classrooms with missing data in math score. Something worth noting is that, regardless the number of classrooms with missing data, the percentage of missing values by classroom is relatively low. For instance, the 1st and 3rd column of table (4) show that Brazil has missing values in almost every classroom both, in math and language scores, but 80% of such classrooms do not have more than 25 or 30% of missing data in those variables respectively.

As regards explanatory variables, they show more classrooms with missing data as well as a higher number of missing information by classroom.

8.2 Final data

As mentioned previously, to estimate the models we dismiss some observations due to missing data problems. The observations with missing values in math score differ from those with missing values in language score, so the final data bases used for each subject differ.

The mathematics data base consists of $N_m = 58,817$ observations (students) which are grouped in $R = 2,736$ classrooms. That is, we work with the 87% of observations and with the 88% of classrooms from the original data. The overall percentage of missing values in isecf index and mothereduc indicator is 7 and 15% respectively. Whereas the overall percentage of *classrooms* with some missing value in isecf index and mothereduc indicator is 53 and 84% respectively. The 80% of classrooms with missing data of isecf index and mothereduc indicator do not have more than 12 and 24% of missing values respectively.

The language data base consists of $N_l = 58,224$ observations (students) which are grouped in $R = 2,730$ classrooms. Consequently, we work with the 86% of the observations and with the 88% of classrooms from the original data. The overall percentage of missing values in isecf index and mothereduc indicator is 5 and 13 % respectively. The total percentage of *classrooms* with some missing value in isecf index and mothereduc indicator is 44 and 83% respectively. The 80% of classrooms with missing data in isecf index and mothereduc indicator do not have more than 10 and 22% of missing values respectively.

Given that the percentage of missing data varies across countries, the *missing filtering process* impacts differently on each country data. In the following lines we present some measures on missing data distribution by country and by subject.

Table 5: Classrooms, sizes and missing data distribution. Reduced sample.

Country	Mathematics data						Language data					
	(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)	(4)	(5)	(6)
Argentina	80.2	19.0	83.1	25.0	93.4	33.3	79.7	19.0	80.6	25.0	92.1	33.3
Brazil	71.4	27.0	80.0	20.0	96.7	33.3	71.4	25.5	71.1	16.7	95.6	29.5
Chile	91.4	28.0	72.8	17.3	90.0	24.4	91.4	28.0	72.2	17.4	90.0	23.9
Colombia	97.3	30.0	42.8	6.3	90.3	15.4	97.3	31.0	38.6	6.2	89.7	15.4
Costa Rica	98.5	19.0	21.6	3.5	74.7	15.0	98.5	19.0	19.6	0.0	74.7	15.0
Dominicana	94.7	22.0	41.6	6.2	92.5	24.3	94.7	22.0	23.6	3.0	91.9	23.5
Ecuador	88.6	22.5	59.8	10.9	83.3	21.1	75.0	22.0	32.8	4.6	76.4	16.7
Honduras	91.6	19.0	42.5	8.3	79.0	23.1	91.1	18.0	30.3	4.0	76.8	20.1
Mexico	91.7	23.0	55.2	13.6	82.5	20.3	92.3	23.0	45.8	8.7	76.8	18.2
Nicaragua	91.1	22.0	68.3	15.5	86.0	25.3	90.0	21.0	37.0	5.0	83.3	20.0
Panama	89.3	18.0	67.1	16.7	91.0	30.3	88.8	19.0	60.8	16.0	88.0	26.7
Paraguay	86.3	16.0	55.4	14.3	88.7	30.7	85.9	16.0	43.8	9.1	88.6	28.6
Peru	95.1	17.0	37.3	8.3	72.7	20.0	95.1	16.0	32.1	6.7	71.6	20.0
Uruguay	68.5	18.0	49.7	13.5	77.3	23.1	68.1	18.0	48.8	12.5	78.4	23.0
Nuevo Leon	98.1	27.0	64.6	12.0	88.0	17.5	98.1	27.0	55.1	10.6	86.1	15.9

(1) % of classrooms from the complete sample. (2) Median of classroom size. (3) % classrooms with missing values of isecf index. (4) % missing data in isecf index by classrooms, Q_8. (5) % classrooms with missing data in edumother indicator. (6) % missing data in edumother indicator by classrooms, Q_8.