9-4-2020 11:00 AM

A Language Barrier To Human Capital Development: The Case Of Guatemalan Students

Fidel Pérez Macal, *The University of Western Ontario*

Supervisor: Navarro, Salvador, *The University of Western Ontario*
: Sicular, Terry, *The University of Western Ontario*
: Lochner, Lance, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Economics

Abstract

Not being proficient in a school's predominant language of instruction can represent a language barrier for students' human capital development. In Guatemala, 24 languages are spoken apart from Spanish, which is the language of instruction in the majority of schools, and about 40 percent of the total population has a non-Spanish language as a mother tongue. National standardized tests show that non-Spanish mother tongue (non-SMT) students are outperformed by SMT students in elementary and secondary schools.

My thesis analyzes whether non-SMT students face a language barrier and traces its source. Two main findings emerge. First, non-SMT students are not yet proficient in Spanish while at school. I find evidence of this language barrier in elementary and secondary schools through a model of latent variables, local instrumental variables, and first difference-instrumental variables. Second, I find that other parents' mother tongue influences what school a parent will choose for their child. I analyze parents' enrollment decisions for schools through the lens of a model of demand as is common in the industrial organization literature. The model also features spillover effects as seen in the literature for residential sorting or social interactions.

Keywords: Language, Education, Guatemala, Human capital.

Summary for Lay Audience

Not being proficient in a school's predominant language of instruction can represent a language barrier for students' human capital development. In Guatemala, 24 languages are spoken apart from Spanish, which is the language of instruction in the majority of schools, and about 40 percent of the total population has a non-Spanish language as a mother tongue. National standardized tests show that non-Spanish mother tongue (non-SMT) students are outperformed by SMT students in elementary and secondary schools.

My thesis analyzes whether non-SMT students face a language barrier and traces its source. Two main findings emerge. First, non-SMT students are not yet proficient in Spanish while at school. Second, I find that other parents' mother tongue influences what school a parent will choose for their child.

Statement of Co-Authorship

This thesis contains material which is co-authored with Salvador Navarro Lozano. Both authors are equally responsible for the work which appears in Chapter 2 of this thesis.

Acknowledgements

I would like to thank my dissertation committee members Salvador Navarro, Terry Sicular and Lance Lochner for their guidance in conducting my research. I also like to extend my gratitude to my fellow graduate student Samantha Goertz for her unconditional support, and to Yvette Bolaños, Edgar Montúfar and Dirección General de Evaluación e Investigación Educativa de Guatemala for their cooperation in developing my research. The completion of this dissertation would not have been possible without the generous financial support of the Department of Economics at the University of Western Ontario and Salvador Navarro's research assistance for which I am grateful. Last, I am forever grateful to my family for their endless support.

To My Family

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1

Introduction

Not being proficient in a schools' predominant language of instruction can represent a language barrier for students' human capital development. Furthermore, if demographic characteristics affect the language learning process, and these characteristics are not taken into account in the education process, then the existence of such a language barrier can hinder not only a student's human capital accumulation, but also an ethnic group's long term development. In 2009, the World Bank reported that Guatemala has one of the highest inequality rates and the worst poverty in rural and indigenous areas in Latin American countries.

Knowing whether students face a language barrier at school is relevant for Guatemala due to its multilingual society. In Guatemala, 24 languages are spoken besides Spanish, and about 40 percent of the total population has a non-Spanish language as a mother tongue. The educational attainment distribution in Guatemala shows that the Spanish mother tongue (SMT) population has more education than non-SMT people. Non-SMT females have the worst educational outcomes. About 70 percent of the indigenous population attends school only until third grade. Given the consequences of growing up poor, the Guatemalan government committed to increase the level and quality of education of the population, particularly for indigenous people. However, non-SMT students' performance on national tests remains poor. In this context, my thesis investigates whether students who do not speak the predominant language as a mother tongue face an obstacle, or language barrier, in acquiring skills while at school in Guatemala since Spanish is the language of instruction after the third grade. Specifically, my thesis identifies the existence and source of the language barrier facing non-SMT students.

The second chapter of my thesis is co-authored with Salvador Navarro. We analyze third grade students' performance on national tests, since non-SMT students are persistently outperformed by SMT students. On average, non-SMT students' math and reading test scores are 0.42 and 0.54 standard deviations lower, respectively, than SMT students' test scores. This chapter analyzes the third grade students' educational achievement by estimating a

model of latent variables that feed into education production functions. We take advantage of a unique dataset that includes questions to measure non-SMT students' Spanish comprehension and effort. The results show that non-SMT students are not yet proficient in Spanish. Non-SMT students' lack of Spanish comprehension can explain the entirety of the observed test score gaps between non-SMT and SMT students.

The third grade students' lack of Spanish comprehension raises the question of whether non-SMT students in secondary schools are not yet proficient in Spanish, since their achievement on national tests is still poor. On average, non-SMT secondary students' math and reading test scores are 0.4 and 0.6 standard deviations lower, respectively, than SMT students' test scores. In the third chapter, I analyze how students' linguistic sorting across schools affects non-SMT students' Spanish comprehension and their non-Spanish language use, and therefore their educational achievement. To explain test score gaps between non-SMT and SMT students, I employ a representative and longitudinal dataset of sixth grade students in 2010. To account for the endogeneity of students' Spanish comprehension, I use different estimation approaches such as two stage least squares, first differences, and local instrumental variables. The results show that, first, students' linguistic sorting across schools is a prominent factor for non-SMT students' Spanish comprehension. Second, non-SMT students are not yet proficient in Spanish by the time they attend secondary school. Last, students' Spanish comprehension explains all of the math test score gap between non-SMT and SMT students.

Given the finding that linguistic segregation across schools is partially to blame for non-SMT students' lack of Spanish comprehension, in the last chapter, I analyze parents' enrollment decisions for their children. This chapter identifies how school attributes, children's non-Spanish language use, and spatial segregation of groups determine parents' school choice, and hence shape students' linguistic sorting across schools. I estimate a model of demand for junior high schools in Guatemala in which parents consider schools as differentiated products. In particular, I allow for the degree of differentiation across schools to depend on characteristics of other parents who select the school in equilibrium. The results show that non-SMT parents value schools in which their child is likely to speak and learn Spanish. However, non-SMT parents also prefer to sort their child into schools where other parents have a similar mother tongue. This latter preference dominates the former as we move away from the Guatemalan capital city, which leads to both spatial and linguistic segregation at school.

My thesis makes two main contributions to the Guatemalan literature. First, there is no consensus on factors that drive test scores gaps between non-SMT and SMT students in the current Guatemalan literature. My findings suggest that Spanish comprehension is an important factor explaining test score gaps between non-SMT and SMT students at elementary and secondary school. Second, parents' linguistic preferences are shaping linguistic composition at school, which is a determinant factor of Spanish comprehension. In the next three chapters, I explain each of these findings.

Chapter 2

Are non-Spanish mother tongue students in third grade facing a language barrier in the Guatemalan education system?

2.1 Introduction

In 2009, the World Bank reported that Guatemala has one of the highest inequality rates in Latin American countries, and the worst poverty in rural and indigenous areas. In Guatemala, indigenous people account for 40 percent of the total population.[1] Roughly 70 percent of the indigenous population attended school only until third grade. The root of such an ethnic inequality may be the result of an educational exclusion of indigenous population that existed before the Guatemalan civil war and that was amplified by war (Chamarbagwala and Morán, 2011).

As part of the Guatemalan civil war peace agreement in 1996, the Guatemalan government committed to increase the level and quality of education of the population, particularly the indigenous population. One particular challenge for the indigenous population is language, since the indigenous population speaks a variety of indigenous languages, and the formal education in Guatemala is mainly conducted in Spanish. To increase the educational attainment of the non-Spanish mother tongue (non-SMT) indigenous population, the Guatemalan government institutionalized a bilingual program in schools.[2] The program aims to gradually increase non-SMT students' Spanish comprehension and also to provide educational instruction, lectures and books, in the students' mother tongue during the first

[1] The indigenous classification represents people's self-classification into this category, which may or may not depend on their mother tongue.

[2] In Guatemala 24 languages are spoken besides Spanish.

six years of education. In practice, however, the program is available only in the first three elementary grades, which implies that from the fourth grade on, the language of instruction is Spanish in the majority of schools (Toledo and Guantá, 2009). The expectation, therefore, is that by the fourth grade non-SMT students should have a good understanding of Spanish.

Table 2.1 shows both math and reading test scores for third grade students by academic year. Since students are not frequently asked their mother tongue, we also classify them by their self-identification as indigenous, which is highly correlated with mother tongue. Test scores are standardized to have a mean of zero and standard deviation of one. The numbers clearly reveal significant test score gaps that have been persistent over time. In 2006, non-SMT students score 0.33 standard deviations below the mean, while SMT students are above the mean by 0.095 standard deviations. In 2014, on average, non-SMT students in third grade are still scoring 0.31 standard deviations below the mean and outperformed by SMT students. Similarly, reading test score gaps indicate that SMT students outperform non-SMT students. On average, non-SMT students score -0.422 and -0.39 standard deviations below the mean for 2006 and 2014 respectively. Test score gaps by students' self-identification as indigenous indicate these gaps have been increasing over time.

The quality of education in Guatemala has emerged as an important research topic due to non-SMT students' poor performance at school, even when the Guatemalan government has produced positive outcomes in increasing the level of education of the non-SMT population (Rubio, 2004; Enge and Chesterfield, 1996). A number of papers have previously studied the educational achievement of students whose mother tongue is not Spanish in Guatemala. On the one hand, Marshall (2011) and Marshall and Sorto (2012) find that speaking a Mayan language at home does not statistically affect rural third grade students' performance on national tests. On the other hand, Meade (2011) indicates that having a non-Spanish language as a mother tongue reduces third grade students' achievement on math national tests by 0.191 standard deviations, equivalent to an unexplained math test score gap of 53 percent. Marshall (2009) finds that rural third grade students in 2001 who speak a non-Spanish language are outperformed by -0.14 standard deviations by their counterparts. McEwan and Trowbridge (2007) find that being a non-SMT student negatively affect test scores by 0.13 standard deviations, equivalent to an unexplained math test score gap of 24 percent. Furthermore, the authors argue that differences in school quality as measured by school fixed effects can account for 69 percent of math test score gaps for third grade students in 2001. Hernandez-Zavala et al. (2006) find that 45 percent of the math gap between non-SMT and SMT students remains unexplained after accounting for differences in household or school inputs.

In this chapter, we investigate whether these gaps are a consequence of students facing a language barrier, i.e., whether the negative effect of being a non-SMT student on test scores may be the result of a low level of Spanish comprehension. The analysis relies on a cross-section of data in 2006 for third grade students. The econometric approach consists of estimating students' latent variables (Spanish comprehension, effort and other unobserved

traits) that feed into education production functions. Since we allow for the distribution of these latent variables to be different for non-SMT students, we can investigate whether the non-SMT gap is mediated by the level of Spanish comprehension of students. We also control for other variables commonly used in the Guatemalan literature like student and parental characteristics. The dataset does not include questions about school characteristics, but it still allows us to include school fixed effects to control for such omitted variables. Then, following McEwan and Trowbridge (2007), we decompose test score gaps into three portions that represent the effect of being a non-SMT student, the effect of Spanish comprehension, and the effect of exerting effort. As opposed to other findings in the Guatemalan literature, by including latent Spanish comprehension in the model, we are able to identify that students whose mother tongue is not Spanish are not yet proficient in Spanish, and that non-SMT students' lack of Spanish comprehension can entirely explain the math and reading test score gaps.

This chapter proceeds as follows. In the next section, we provide an overview of the Guatemalan educational system and some relevant evidence of non-SMT students' poor performance at school. In Sections 2.3 and 2.4, we describe the data and the model respectively. We then discuss the results, which include the production function estimates, the test score gap decomposition, and the determinants of Spanish comprehension, and effort. Last, we conclude.

2.2 A brief review of the Guatemalan educational system and some relevant evidence of non-SMT students' poor performance at school

The educational system in Guatemala consists of 3 stages: elementary, junior high, and high school. The elementary stage consists of 6 grades and usually starts in the year the child turns 7, as the school year begins in January. In this first stage, schools teach basic knowledge of math, history, science, and Spanish grammar, among other courses. The second stage consists of three grades starting at the age of 13 and provides a deeper understanding of the same courses. The high school stage, also consisting of three grades, is different than the previous stages in the sense that students may choose from a variety of specialized programs in addition to the core courses mandated by the Ministry of Education. For example, the most popular program, Bachillerato, includes only the core courses.[3] A different program is Perito, which in addition to the core courses, provides students with vocational courses such as mechanics. It is geared to students who may not continue on to college. At any stage, students have to score above 60% in all subjects in order to progress onto the next grade.

[3]This program consists of two grades.

The educational attainment distribution and population in Guatemala can be divided in two main groups of people, people who have Spanish as a mother tongue and those whose mother tongue is one of the more than 20 indigenous languages spoken in Guatemala. Table 2.2 shows the educational attainment of these two groups. As the numbers reveal, the educational attainment of non-SMT speakers is substantially lower than their SMT counterparts regardless of gender. The table reveals that around 71.1 and 85.7 percent of the non-SMT male and female population, respectively, has less than or equal to 3 years of education.

While the origin of these gaps is subject to debate, the Guatemalan civil war arose in part as a consequence of the perception of the disparate treatment of the indigenous Guatemalan population, e.g., excluding the indigenous population from the educational system by not providing schooling assistance programs in their mother tongue (Toledo and Guantá, 2009). The Guatemalan civil war lasted for 36 years, with the final peace agreement being signed in 1996. The war itself contributed to a further widening of these gaps. As documented by Chamarbagwala and Morán (2011), Mayan females saw their educational attainment reduced from 3% (before the worst period of the war), to 12% (during the worst period of the war), to 30% (after the worst period of the war) with respect to a 3.83 years of schooling baseline. For Mayan males, the reduction was 0.12%, 0.47% and 1.17% years of schooling relative to a baseline of 4.66 years of schooling.

Following the Guatemalan civil war peace agreement, the Guatemalan government began a transformation of the Guatemalan educational system. The stated goals that the Ministry of Education wants to achieve are: to improve the quality of schools, to reduce gender gaps in education, and to ensure schools can teach effectively in a multicultural society. This last goal was to be met by the introduction of a bilingual program designed to provide non-SMT students with instruction, lectures and books, in their own mother tongue. The introduction of the program was backed up by the positive outcomes of an experimental bilingual education project implemented in 40 schools between 1980 and 1984.[4] Enge and Chesterfield (1996) argue that the experimental bilingual education project had a positive effect in reducing grade retention and drop-out rates, and also an improvement on standardized test scores for non-SMT students. A key feature of the bilingual program is the introduction of a curriculum designed to allow for non-SMT students' smooth transition from their mother tongue to the language of instruction at school: Spanish. Using this curriculum, the transition is supposed to be achieved during the first six years of education. In practice, however, the program is targeted only toward the first three elementary grades, which implies that from the fourth grade on, the language of instruction is Spanish at school (Toledo and Guantá, 2009). Furthermore, Patrinos and Velez (2009) estimate that the government saved about $5 million due to the bilingual program (e.g., less grade retention), enough to cover the education of 100,000 students in Guatemala.

[4]In addition to the bilingual program, other programs or school types have been implemented. For instance, multigrade schools were created to provide education to highly isolated communities. In these schools students of different ages/levels are grouped together to take classes.

With the establishment of the Millennium Development Goals (MDG) program, Guatemala's government instituted the requirement of an internationally known system for measuring student performance (Fortín, 2013; Anderson, 2001). In line with the MDG program, in 2005, the Guatemalan government established a law that mandates the Ministry of Education to test all students in math and reading at the end of their high school studies, with the objective of monitoring the quality of education. The Ministry of Education occasionally tests first, third and sixth grade students in elementary schools as well.

National standardized tests shows that non-SMT students are consistently performing poorly at school. Table 2.1 shows math and reading test scores for third grade students for the 2006-2014 academic years. Test scores are separately shown by students' mother tongue and by self-identification as indigenous, since third grade students are not frequently asked their mother tongue by the Ministry of Education. As evidenced from Table 2.1, non-SMT students are outperformed by their counterparts regardless of the subject and students' classification, i.e., mother tongue or indigenous.

The current Guatemalan educational literature is scarce and findings about the determinants of test score gaps are mixed. On the one hand, some studies find that commonly observed household background variables have the largest impact when explaining test score gaps between non-SMT and SMT students. Hernandez-Zavala et al. (2006) analyze student performance at school for Mexico, Peru and Guatemala. By estimating education production functions and using the Oaxaca-Blinder methodology, the authors decompose observed test score gaps between students who grew up or not speaking an indigenous language at home. The authors find that about 33 and 22 percent of the observed math gap is the result of differences in household background and in observable school attributes, respectively. Furthermore, the authors argue that the 45 percent of the math gap is unexplained. Meade (2011) analyzes the sources of disparities in Guatemalan primary schools for first, third and sixth grades in the 2006-07 school year. The author finds, first, that students' self-identification as indigenous does not statistically explain third grade student achievement on national tests after controlling for students' mother tongue. Second, that having a non-Spanish language as a mother tongue reduces third grade student achievement on math national tests by 0.191 standard deviations, equivalent to an unexplained math test score gap of 53 percent. Furthermore, the author shows that school quality as measured by school types does not explain student achievement in math test scores.[5]

On the other hand, other researchers suggest that school quality or community characteristics contribute more to explain test score gaps. McEwan and Trowbridge (2007) analyze the educational achievement of rural third grade students in 2001. Their estimates imply that students who speak a Mayan language at home underperform on national tests by 0.13 and 0.30 standard deviations for math and reading respectively, even after controlling for

[5]School types in Meade (2011) stand for different educational programs that the Ministry of Education has implemented to ensure schools can teach effectively in a multicultural society and to expand the school coverage across Guatemala. The bilingual educational program is one of these school types.

school fixed effects. Decomposing test score gaps between these two groups of students by differences in school quality and family variables, the authors find that differences in school quality can account for 69 and 65 percent of math and reading test score gaps, respectively.[6] Differences in family background between these two groups of students explain 8 and 6 percent of math and reading test score gaps, respectively. The unexplained test score gap (the dummy variable for whether students speak a Mayan language at home) accounts for 24 and 29 percent for math and reading respectively. Marshall (2009) estimates an achievement production function using data for rural third grade students in 2001. The author finds that characteristics of the community where students reside, as measured by departmental dummy variables, are partly responsible for the observed test score gaps between students who do and do not speak a Mayan language. The author also finds that students who speak a non-Spanish language are outperformed in math by -0.14 standard deviations by their counterparts.[7] However, the author finds that rural students attending schools with indigenous teachers perform better at math tests.

Marshall (2011) study school quality and attendance for Guatemalan rural students in 2001. The author finds that speaking a Mayan language at home does not statistically affect third grade students' educational achievement on national tests.[8] However, the author finds that if students who speak a Mayan language at home are matched with teachers who teach speaking a Mayan language, these students are less likely to fail a grade or dropout. This finding suggests that school quality as measured by teachers' ability to instruct in students' mother tongue plays an important role for non-SMT students' grade progression. Last, Marshall and Sorto (2012) study the effects of teacher mathematics knowledge and pedagogy on Guatemalan rural students in 2001. The authors also find that third grade rural students' achievement on math test scores is not statistically influenced by speaking a Mayan language at home. However, the influential factor for student's achievement is teachers' mathematics knowledge for teaching.

As the Guatemalan literature shows, non-SMT students are in a disadvantaged position either by poor family or school conditions. Researches have provided economic interpretations for non-SMT students poor performance at school. However, the lack of consensus about the main determinants for test score gaps suggests that unobserved heterogeneity is an important concern. An omitted variable in the Guatemalan literature is students' Spanish comprehension. Then, not only estimates may be biased by not controlling for Spanish comprehension, but also current findings may be misleading the Ministry of Education' policies to improve the non-SMT students' educational development. Facing a learning barrier at school may affect not only non-SMT students' school performance, but also influence poverty in the long run.

[6]School quality is measured as school fixed effects.

[7]This negative effect is not statistically significant in this paper.

[8]In Marshall (2011), Mayan languages stand for Q'eqchi or Kaqchikel.

2.3 Data

The empirical analysis relies on data from the Ministry of Education in 2006. The Ministry of Education monitors the quality of education by testing students' math and reading skills through national standardized tests. These tests are mainly targeted to high school students, but elementary students are also tested in some years. National tests are completely designed by Ministry of Education's staff and marked based on Item Response Theory (Rasch correction), which involves grading not only by the number of correct answers, but also by the degree of difficulty of each question. By the time students take the national tests, students also self-report their home, parents and own characteristics. The student survey is also designed by Ministry of Education's staff. The 2006 dataset is a representative sample of students attending public schools. The sample consists of 656 schools and includes 20,965 students. From this dataset, 35 and 22 percent of students self-reported being indigenous and having a non-Spanish language as a mother tongue respectively.

The 2006 dataset is advantageous for studying the effects of Spanish comprehension on students' human capital accumulation because this dataset includes data about non-SMT students' self-evaluation regarding Spanish acquisition. Later datasets for third grade students do not contain this information. Furthermore, the dataset includes questions that can be considered as directly related to students' effort when studying such as whether students do their homework without any help. As explained in Section 2.4, we use these measures as indicators of the latent variables Spanish comprehension and effort.

Table 2.3 shows the variables we use to infer the the latent variable for Spanish comprehension ($\theta_{c,i}$). The first column displays the variable names, and the second and third columns indicate the proportion of non-SMT and SMT students who reported this characteristic, respectively. The last column shows the gap between these two groups of students. To infer the latent variable of Spanish comprehension, we use students' responses to the question of whether they already think in Spanish when speaking, writing, and subtracting numbers. We also include whether students speak Spanish with their siblings, with at least one parent, and during school recess. For instance, while 80.2% of non-SMT students reported Spanish thinking when calculating mathematical operations, 90.8% of SMT students reported Spanish thinking. Similar gaps hold for Spanish thinking when writing and speaking Spanish. An equivalent pattern is found in the context of Spanish language use. Non-SMT students do not speak Spanish as intensively as SMT students.

Students also reported whether they speak a non-Spanish language with their siblings, parents, and peers. As evidenced from the table, around 76% of non-SMT students reported speaking a non-Spanish language with their siblings and parents, while less than 6.4% of SMT students reported non-Spanish language use with their family members: siblings or parents. The last column shows the gap for non-Spanish language use between the groups. On average, there is a 72.3, 78.3 and 27.2 percentage-point difference between non-SMT and

SMT students use of a non-Spanish language with family members, at least one parent, and at school recess. As expected, non-SMT students are speaking a non-Spanish language much more frequently than SMT students.

Table 2.4 displays the variables used to infer the latent variable of students' effort, $\theta_{e,i}$. Students' effort is inferred from three variables: hours of study at home, no help at home when doing homework, and always do homework. The first pattern is that non-SMT students devote less time to study. Around 75.9% of non-SMT students study less than one hour as opposed to 68.2% of SMT students. In terms of receiving help at home to do their homework, non-SMT students are 3.6 percentage-points more likely to receive help than SMT students. There is no statistical difference between groups regarding to whether they always do their homework.

Table 2.5 shows statistics about the main variables that are commonly used in empirical studies for Guatemala, and which we also include in our analysis. As mentioned in Section 2.2, non-SMT students are outperformed by SMT students by 0.428 and 0.543 standard deviations in national math and reading tests respectively. Non-SMT students show a lower enrollment rate in pre-elementary education (or pre-school) of 1.4 percentage-points. Non-SMT students are more likely to speak an indigenous language with teachers and to attend a school that follows the bilingual program curriculum. Non-SMT students are also less likely to self-report having SMT parents. In terms of parents' educational attainment, non-SMT parents show a lower educational attainment, especially for non-SMT mothers. About 28.7% of non-SMT students have fathers with a higher educational attainment than elementary school as opposed to 43.9% percent of SMT students.[9] In the case of mothers, 21.1 and 38.6 % of students have mothers with a higher educational attainment than elementary school for non-SMT and SMT students, respectively. Although, there exists statistical difference in parents' educational attainment between these two groups, empirical evidence for Guatemala suggests that school inputs or being a non-SMT student could be important in explaining test score gaps between non-SMT and SMT students, rather than family inputs such as parental education.

2.4 The model

In what follows we specify a simple education production function that depends both on observable characteristics of the student, the family, and the school; and also on various latent unobservable traits and decisions made by the student. Let y_i^* be the (potentially latent) knowledge of a subject possessed by student, i. Let W_i denote the vector of all individual, family, and school inputs and characteristics that determine the individual's knowledge. The knowledge production function, f_y, is thus given by:

[9] 0.287=0.196+0.036+0.055

$$y_i^* = f_y(W_i). \tag{2.1}$$

In general, no dataset contains all variables contained in W_i required to estimate this production function under such a general specification. In particular, an element of W_i is whether the student is a non-SMT speaker. At this level of generality, it is not possible to distinguish between the possible mechanisms behind the effect that being a non-SMT student has on knowledge production. Therefore, some assumptions are needed. We begin by separating W_i into inputs and characteristics observed by the econometrician, and those that are not directly observed. We let X_i stand for student and family observable inputs for student i, while the vector, X_s, represents observable inputs from school s that the student i attends.

As discussed in Section 2.3, one key advantage of the dataset we use is that it includes measures that we can use to infer some of the unobservable variables in W_i. In particular, we model the unobserved inputs using a correlated factor model as follows. First, the education literature emphasizes that knowledge acquired by students at school is, in large part, determined by their effort, $\theta_{e,i}$ (e.g., Stinebrickner Ralph and Stinebrickner Todd R. (2008)). Student effort is the result of an optimization problem in which students decide on input intensity based on their prior knowledge of the subject. Therefore, student effort encapsulates all the inputs, current and lagged, that the student uses when choosing effort. Second, given our particular interest in studying the possibility of a language barrier being present, we also separately consider the student's Spanish comprehension, $\theta_{c,i}$, as part of the unobservable variables in W_i. Third, to control for other inputs that may not be fully accounted for in effort or Spanish comprehension, we also control for the remaining unobserved traits of student i by $\theta_{ut,i}$, which may include what we generically call "ability". Under these assumptions, we can rewrite the knowledge production function, f_y, as:

$$y_i^* = f_y(X_i, X_s, \theta_{e,i}, \theta_{c,i}, \theta_{ut,i}). \tag{2.2}$$

Our factor model for the unobserved inputs or latent variables works as follows. In the following discussion, we let $d \in \{e, c\}$ index the unobserved input of interest. Let $M_{i,j}^d$, $j = 1, \ldots, J^d$, be a j^{th} noisy measure of the latent variable $\theta_{d,i}$. The measures we observe are categorical variables, so we model their relation to the latent variable of interest as an ordered discrete choice factor model. Let L_j^d be the number of categories included in measure j for latent variable $\theta_{d,i}$. For each of these measures we define constants (cutoffs) $\left\{\kappa_{j,\ell}^d\right\}_{\ell=1}^{L_j^d}$ such that $\kappa_{j,\ell-1}^d < \kappa_{j,\ell}^d$, $\kappa_{j,0}^d = -\infty$ and $\kappa_{j,L_j}^d = \infty$, and write:

$$\text{if } M_{i,j}^d = \ell \Rightarrow \kappa_{j,\ell-1}^d < \lambda_{0,j}^d + \theta_{d,i}\lambda_{1,j}^d + \epsilon_{i,j}^d < \kappa_{j,\ell}^d. \tag{2.3}$$

In this model, $\theta_{d,i}$ is the factor; $\lambda_{1,j}^d$ is called the "loading" and measures the strength of the relation between the measure $M_{i,j}^d$ and the factor $\theta_{d,i}$; and $\epsilon_{i,j}^d$, called the "uniqueness", captures other elements such as measurement errors. We impose the following assumptions. First, we assume that $\theta_{d,i} \perp\!\!\!\perp \left\{\epsilon_{i,j}^d\right\}_{j=1}$ and $\epsilon_{i,j}^d \perp\!\!\!\perp \epsilon_{i,j'}^d$ for $j, j' \in \left\{1, \ldots, J^d\right\}$, $j \neq j'$. Second, we assume that

$$\begin{aligned} \theta_{d,i} &\sim N\left(Q_{i,t}'\beta_d, \sigma_{\theta_d}^2\right), \\ \epsilon_{i,j}^d &\sim N\left(0, \sigma_{\epsilon_j^d}^2\right), \end{aligned} \tag{2.4}$$

where $Q_{i,t}$ depends on student, parental and school characteristics.[10] Third, we assume that $\theta_i = (\theta_{e,i}, \theta_{c,i}, \theta_{ut,i})$ are jointly normal with variance covariance matrix given by Σ_θ.

Under these assumptions, identification of all the elements of the factor model just described follows from the analysis in Carneiro et al. (2003), Cunha et al. (2010) and Fruehwirth et al. (2016). In fact the identification results in these papers are established under weaker conditions. In particular, the normality assumption in equation 2.4 is not needed. The advantage of observing multiple measures, so that we can recover the distribution of each $\theta_{d,i}$, is that we can allow $Q_{i,t}$ to include an indicator for whether the student is a non-SMT speaker. In other words, it allows us to establish the extent to which non-SMT students exert more or less effort than their SMT counterparts; and, more importantly for our purposes, whether Spanish comprehension differs by mother tongue.

An additional concern arises from the fact that the test scores observed by the econometrician are unlikely to perfectly measure the student knowledge, y_i^*. As a consequence, we impose the common assumption that the student's observed test score, y_i^g, is a noisy measure of $y_i^{g,*}$ such that $y_i^g = y_i^{g,*} + \epsilon_i^g$, for $g \in \{\text{math, reading}\}$. We further assume that the knowledge production function is linear in parameters so that

$$y_i^g = \alpha_0 + \theta_{e,i}\alpha_{e,g} + \theta_{c,i}\alpha_{c,g} + \theta_{ut,i}\alpha_{ut,g} + X_i'\alpha_{x,g} + \epsilon_i^g. \tag{2.5}$$

Notice that we control for other unobserved traits for student i with the factor $\theta_{ut,i}$, which may include ability. We also account for school specific inputs, X_s, by having school fixed

[10] Specifically, we control for school fixed effects.

effects in the means of our latent variables. Therefore, we assume that $\epsilon_i^g \perp\!\!\!\perp (\theta_i, X_i, \epsilon_{i,j})$, that $\theta_i \perp\!\!\!\perp X_i$ for all $j \in \{1, \ldots, J^d\}$, and that $\epsilon_i^{math} \perp\!\!\!\perp \epsilon_i^{reading}$. The parameter $\alpha_{ut,g}$ measures the strength of the relation between the measure y_i^g and the factor $\theta_{ut,i}$, and ϵ_i^g captures other elements such as measurement errors. Our specification in equation 2.5 contains, as part of X_i, an indicator for whether the student is a non-SMT speaker, which concurrently with equation 2.4 allows us to establish how much of the effect of being a non-SMT student is a direct effect as measured by $\alpha_{x,g}$ (and that the current Guatemalan literature cannot explain, see Section 2.2), and how much arises from differences in θ_i.

To determine what is driving educational achievement gaps between non-SMT and SMT students, we use a simple decomposition that separates educational achievement into differences in mother tongue, Spanish comprehension and effort as follows:

$$E\left(y_i^g \mid NonSMT\right) - E\left(y_i^g \mid SMT\right) \quad = \quad \alpha_{d,g}\left[E\left(\theta_{d,i} \mid NonSMT\right) - E\left(\theta_{d,i} \mid SMT\right)\right] \quad (2.6)$$

We use this decomposition to compare the effect of being a non-SMT student on test score gaps in this chapter with the findings in the current Guatemalan literature. To calculate the expected value of each latent variable in equation 2.6, we use the estimates of the factor models to predict the most likely values of the latent variables given the data we observed. Specifically, we employ Bayes rule to predict the latent variables distribution conditional on the data, and then use this distribution to predict the latent variables expected values.

2.5 Results

Our results indicate that Spanish comprehension and effort play a key role in third grade students' performance on national tests. First, we look at the effects of non-Spanish mother tongue, Spanish comprehension, and effort on students' performance on the tests. Then, we present how differences in mother tongue, Spanish comprehension, and effort between non-SMT and SMT students contribute to the observed test score gaps. Last, we discuss the determinants of Spanish comprehension, and effort, i.e., estimates from the factor model.

2.5.1 The effects of Spanish comprehension and effort on educational achievement

In this section we discuss how Spanish comprehension and effort influence students' performance on math and reading tests. Given our particular interest in studying the possibility

of a language barrier being present, in the estimation of the production functions we control for student Spanish comprehension, but also for variables that the international educational literature has shown as influential: effort and other unobserved traits such as ability. Furthermore, we also control for whether students are non-SMT speakers, as is common in the Guatemalan literature. Controlling for being a non-SMT student both in the production function and in the latent variable of Spanish comprehension allows us to observe whether the unexplained effect of being a non-SMT student on the production function is the result of the omitted variable of students' lack of Spanish comprehension in the current Guatemalan literature.

We structure the discussion that follows by two main streams in the current Guatemalan literature. First, we compare the findings in this chapter against those findings in McEwan and Trowbridge (2007) and Hernandez-Zavala et al. (2006). These papers find evidence that either family or school inputs are important drivers of performance at school. Tables 2.6 and 2.7 show results for math and reading, respectively, in this context. Then, our results are comparable to those results in Meade (2011), Marshall (2009), and Marshall (2011), in the sense that we look for evidence of whether non-SMT students better learn a subject by speaking their mother tongue with their teachers. Tables 2.8 and 2.9 show math and reading results, respectively, in this context. With the exception of Marshall (2011), the above mentioned papers find evidence of poor performance at school by mother tongue. To make the discussion of this section clearer, we first discuss with some detail the math results. Then, we just highlight main differences for reading results.

All tables for the production functions share the same structure, and all models control for the same variables, except for those stated in the tables. These tables are divided in three main rows. The upper row of the table shows the estimates of variables commonly used in the Guatemalan literature such as being a non-SMT student. The middle row displays the estimates associated with the latent variables: Spanish comprehension, effort and other unobserved traits; while the bottom row displays whether models include other controls besides those presented in the first and second main rows. Columns in the tables are also divided in two main groups. In the first group, from second to the sixth column, we estimate models by ordinary least squares (OLS) to make a better comparison to the findings in Meade (2011), Marshall (2011), Marshall (2009), McEwan and Trowbridge (2007), and Hernandez-Zavala et al. (2006). In the second group, the last column, we show the estimates from the correlated factor model.

In short, three main findings emerge in this section. First, we find that students' Spanish comprehension and effort are the most important determinants of performance on the tests. In terms of the estimated magnitudes, the most influential attribute on test scores is student effort. The impacts of Spanish comprehension and effort on test scores are positive, significant and robust to other unobserved traits that students may have. Second, non-SMT students learn a subject better if they are matched with teachers who can speak the students' mother tongue. Third, our results indicate that the unexplained and direct effect of being

a non-SMT student on math and reading test scores decreases when we control for Spanish comprehension in models estimated by OLS, but this negative effect is still significant in some model specifications. However, this negative effect is always insignificant when the production functions are jointly estimated with the correlated factor model.

We begin by comparing our results with those papers that find evidence that either family or school inputs are important drivers of performance at school: the first stream in the current Guatemalan literature. First, recall that in Section 2.2 we discuss that the current Guatemalan literature has analyzed elementary students' educational performance. This literature finds that from 24 to 53 percent of the math gap is left unexplained after accounting for household characteristics and school inputs.

Table 2.6 displays the results for math. In columns 2-6 we show models estimated by OLS. All models control for students' pre-elementary education (or pre-school), parents' mother tongue and parents' educational attainment (variables commonly used in the Guatemalan literature).[11] In the simple model specification, see OLS (1) in the second column, when we control for only those variables commonly used in the Guatemalan literature, results indicate that non-SMT students who reside in rural areas are outperformed by 0.37 standard deviations by their counterparts: either SMT students or non-SMT students who reside in urban areas. Furthermore, students who speak a non-Spanish language with their teachers score 0.258 standard deviations lower than those students who speak only Spanish. In the second model specification, OLS (2) in the table, we control for school fixed effects. The main difference relative to OLS (1) is that OLS (2) indicates that non-SMT students have 0.12 standard deviations lower scores than SMT students regardless where non-SMT students reside. This finding is comparable to the findings in Marshall (2009) and McEwan and Trowbridge (2007) where the authors find a negative effect of -0.13 and -0.14 respectively.

In the OLS model specifications (3), (4), and (5) in Table 2.6, we look for evidence of whether either effort measures, Spanish comprehension measures or reading test scores (as a proxy of students' innate ability) can account for the negative effect of being a non-SMT student on test scores. First, the results from the model OLS (3) indicates that effort is not the driving force of non-SMT students' poor performance at school. However, results in the model OLS (4) provides preliminary evidence that non-SMT students are not yet proficient in Spanish. After controlling for the Spanish comprehension measures, the negative direct effect of being a non-SMT student on the national test score is no longer significant; however, this estimate is not robust to different model specifications. Last, when controlling for reading test scores as a proxy of students' innate ability, the preliminary finding that the Spanish

[11] Household's educational attainment stands for the maximum educational attainment of either the mother or the father. The variables students' pre-elementary education (or pre-school), parents' mother tongue and educational attainment are dummy variables. Students' pre-elementary education takes a value of one if students attended pre-school. Parents' mother tongue takes a value of one if parents are SMT speakers. Household's educational attainment takes a value of one if at least one parent completed elementary school for instance.

comprehension measures can account for the negative direct effect of being a non-SMT student on the national test score still holds.

The last column in Table 2.6 shows the result from the correlated factor model of latent variables. Recall that we allow the means of all latent variables to depend on student, family and school inputs, but the Table 2.6 displays only the controls that directly affect the production function, not the ones that affect it indirectly through the the latent Spanish comprehension, effort and unobserved traits.[12] Results from the correlated factor model indicate that, first, Spanish comprehension and effort have the expected positive signs and are statistically significant. Second, similar to the OLS (5) model, the latent variable for Spanish comprehension reduces the negative direct effect of being a non-SMT student on the production function, with an estimate of 0.023.

The second stream in the current Guatemalan literature looks for evidence of whether non-SMT students better learn a subject by speaking their mother tongue with their teachers: Marshall (2009), and Marshall (2011). Specifically, we now also include variables in the production function that control for whether non-SMT students speak their mother tongue with teachers, and whether non-SMT students attend schools that follow the bilingual program curriculum. Our results are also comparable to the finding in Meade (2011) about the negative effect of being a non-SMT student on test scores when not controlling for Spanish comprehension.

Table 2.8 shows the results for the math production function. The first model estimated by OLS controls only for those variables commonly used in the Guatemalan literature. Results suggest that rural non-SMT students perform poorly at school. However, if non-SMT students are able to attend schools where the teacher can speak a non-Spanish language, non-SMT students better learn math. When controlling for school fixed effects, see OLS (2) in the table, the main difference is that non-SMT students, regardless of where they reside, are outperformed by 0.177 standard deviations by their SMT counterparts. Still, non-SMT students are better at learning when speaking their mother tongue with teachers or when attending a school with the bilingual program.

In models OLS (3), (4) and (5), we control for the effort measures, Spanish comprehension measures, and reading test scores as a proxy of students' innate ability, respectively. Three main findings emerge. First, by controlling for effort, the negative effect of being a non-SMT student on the math test score, see model OLS (3), does not change its magnitude and is still significant. This finding still holds even when controlling for the Spanish comprehension measures and reading test scores, see models OLS (4) and (5); however, the estimates in these models for being a non-SMT student are smaller than the estimate in OLS (3). Second, models OLS (4) and (5) also suggest that non-SMT students are better at learning math when speaking their mother tongue with their teachers. Third, comparing the estimates

[12]The identification of the reading factor loading for the latent variable of other unobserved traits requires variables that only affect the mean of the latent variable and not directly the production.

for being a non-SMT student by OLS in Tables 2.6 and 2.8 (the two Guatemalan literature streams), we observe that by controlling for Spanish comprehension measures, the direct effect of being a non-SMT student is at least reduced, but in some cases is still significant.

Finally, in last column of Table 2.8, we show the production function estimates of the correlated factor model. One main new finding emerges. The direct effect of being a non-SMT student on math test scores is not significant, and the estimate is positive. This finding suggests that the negative relationship of being a non-SMT student and performance at school from previous papers is the result of non-SMT students' lack of Spanish comprehension. Different from OLS estimates in Tables 2.6 and 2.8, the factor model estimates of the direct effect of being a non-SMT student are always statistically equal to zero, and robust to different model specifications.

Furthermore, results in Table 2.8 still indicate that effort and Spanish comprehension positively affect students' performance at school, and that student effort is the most influential factor on math performance as reflected in the estimate magnitudes. The fact that Spanish comprehension can account for the negative effect of being a non-SMT student on test scores is suggestive evidence of non-SMT students' lack of Spanish comprehension. The estimation of the correlated factor model allow us to look for a more direct evidence. We discuss how much Spanish comprehension explains test score gaps in Section 2.5.2, and the main drivers of Spanish comprehension in Section 2.5.3.

In the context of reading test scores, we just highlight main differences relative to the math case. Table 2.7 displays the results for the reading production function under the first stream of the Guatemalan literature. Results are similar to the math case. Estimates from OLS models for being a non-SMT student are negative and significant if Spanish comprehension measures are not included as controls. However, this finding does not hold when we jointly estimate the correlated factor model and the system of production functions (see last column).

In the context of the second stream, Table 2.9 shows the estimates of the reading production function. In brief, the finding that the negative relationship between being a non-SMT student and poor performance on national tests still holds when models are estimated by OLS, even when controlling for Spanish comprehension measures. However, when the factor model is estimated, such a negative relationship between being a non-SMT student and poor performance on tests is not longer significant.

To conclude this section, results show that both Spanish comprehension and effort play an important role in learning math and reading at school for non-SMT students. Since the bilingual education program at school only targets the first three grades in the Guatemalan educational system, one would expect these third grade non-SMT students to be close to proficient in Spanish. The findings, however, suggest that non-SMT students in third grade are not yet proficient in Spanish, and that their educational development may be at risk.

2.5.2 Decomposition of students' math and reading test scores by mother tongue, Spanish comprehension and effort

In this section, we focus the discussion specifically on the model specification that allows non-SMT students to better learn a subject by speaking their mother tongue with their teachers. The discussion is structured as follows. First, we define our benchmark specification by restating the findings in Guatemalan literature. Then, to compare the findings in this chapter with the previous literature for Guatemalan students, we use equation 2.6 to show how differences in mother tongue, Spanish comprehension, and effort between non-SMT and SMT students contribute to the observed test score gaps.

We use as a benchmark specification the findings in Meade (2011), Marshall (2009), McEwan and Trowbridge (2007), and Hernandez-Zavala et al. (2006) which indicate that the negative effect of being a non-SMT student on school performance ranges from 24 to 53 percent of the math gap. Marshall (2009) and McEwan and Trowbridge (2007) label this negative effect as the unexplainable portion of the gap that cannot be accounted for by differences in student, parental or school inputs. Furthermore, McEwan and Trowbridge (2007) find that differences in school qualities explain 69 percent of math test score gaps.

Table 2.10 displays how differences in mother tongue, Spanish comprehension, and effort between non-SMT and SMT students contribute to the observed test score gaps. For each test score, we show the results of three model specifications. The first two OLS models stand for the OLS models (2) and (4) at Tables 2.8 and 2.9 for math and reading respectively. The OLS model (2) is comparable to the current Guatemalan literature in the sense that we control for student, parental and school characteristics.[13] On top of the OLS model (2) specification, in OLS model (4) we also control for the effort and Spanish comprehension measures observed in the dataset. The results of the factor model are in the last column.

In brief, non-SMT students are not yet proficient in Spanish. The unexplained effect in the current Guatemalan literature that students perform poorly at school due to their non-Spanish mother tongue decreases from 41.4 and 31.5 percent to 24.8 and 14.9 percent, for math and reading, respectively, when we include effort and Spanish comprehension measures as controls in OLS models. The results from the factor model indicate that students' math and reading performance at school does not differ by mother tongue once we account for differences in Spanish comprehension and effort.

The poor performance of non-SMT students in math is due to their lack of Spanish comprehension. Table 2.10 displays this finding. First, the result for the OLS model (2) indicates that having a different language other than Spanish as mother tongue reduces students performance on the national test by 0.177 standard deviations, which accounts for 41.4 percent of the observed total gap of 0.428. This negative effect is in line with the current

[13]Specifically, we control for school fixed effects.

Guatemalan literature. Second, when we estimate the OLS model with effort and Spanish comprehension measures as noisy measures for the true effort and Spanish comprehension, see OLS (4), important findings emerge. Spanish comprehension seems to play a key role in acquiring knowledge for non-SMT students. Differences in Spanish comprehension between students by mother tongue explain about 44.9 percent of the observed gap. The negative effect of being a non-SMT student drops by 16.6 percent, and the effort measures only explain 1.6 percent of the gap.

Third, the joint estimation of the factor model and system of test scores allows us to integrate out students' Spanish comprehension, and thus find an unbiased estimate of Spanish comprehension on test scores. As mentioned above, by controlling for noisy measures of effort and Spanish comprehension, the negative effect of being a non-SMT student gets smaller in absolute value. Therefore, it is expected that if we had the real Spanish comprehension of students, this negative effect would tend to zero or even to a positive effect.

The results for the factor model in Table 2.8 indicates that, first, the direct effect of being a non-SMT student on test scores is no longer statistically different than zero. In other words, students' performance at school does not differ by mother tongue when we account for differences in Spanish comprehension and effort. Second, non-SMT students' poor performance at school is due to their lack of Spanish comprehension. Differences in Spanish comprehension by students' mother tongue entirely explain the math test score gap. Non-SMT students are outperformed by 0.532 standard deviations by their SMT counterparts. Third, the factor model also indicates that non-SMT students exert more effort than their counterparts. In order to progress into higher grades the Ministry of Education mandates students to score above 60 percent in all subjects. Therefore, if non-SMT students are not able to learn as much as their native speaker counterparts by attending classes, it is most likely that non-SMT students may utilize other resources to learn by themselves to offset the learning deficit as a result of their lack of Spanish comprehension. For instance, Fu and Mehta (2018) indicate that students who have performed poorly at school are more likely to receive help from their parents.

In the context of reading, we briefly discuss the results here, since the findings are quite similar. First, the results from the OLS model (2), without controlling for effort or Spanish comprehension, indicate that students perform poorly at school because of their mother tongue. They have test scores that are 31.5 percent lower relative to an SMT student. However, when we control for effort and Spanish comprehension measures, OLS model (4), the test score reduction is only about 14.9 percent. In the case of the factor model, students' performance at school does not differ by mother tongue anymore.

The findings in this section reveal an educational challenge for non-SMT students in the Guatemalan educational system, since the Ministry of Education requires students to score above 60 percent in all subjects to progress into higher grades.[14] We show that non-SMT

[14]The national tests do not affect students' school progression.

students are not yet proficient in Spanish, which affects their performance at school and puts at risk their educational development. Non-SMT students' lack of Spanish proficiency in third grade raises the question of whether non-SMT students' poor performance at higher grades is also the result of their lack of Spanish proficiency as well, a topic I discuss in Chapter 3.

2.5.3 Determinants of Spanish comprehension and effort

This section shows the estimates for students' Spanish comprehension and effort from model 2.3, and also other unobserved traits students may have from model 2.5. We assume that the means of latent variables can be influenced by student, parental and school characteristics. We include as student characteristics whether students are non-SMT speakers, live in a rural area, and if they attended pre-elementary education. Among household characteristics, we include whether parents have Spanish as a mother tongue, and parental educational attainment.[15] Last, we control for school fixed effects to account for any school attributes that can influence these latent variables.

Table 2.11 shows the estimates for students' Spanish comprehension, effort and other unobserved traits. In short, this table shows that non-SMT students are not yet proficient in Spanish. Third grade non-SMT students, on average, have a lower Spanish comprehension than their SMT counterparts by 1.62 standard deviations.[16] This lack of Spanish comprehension stands for 50.2 percent of the observed math gap, which amounts to a reduction of 0.215 standard deviations in math. However, if non-SMT students have at least one SMT parent, their Spanish comprehension improves. In the context of effort, the main driver of student effort is parental educational attainment. All the above mentioned estimates are statistically different than zero.

Now, we discuss the results of Table 2.11. This table is divided in three main blocks. The upper block contains the factor loadings of the measurement system. The middle block shows the estimates of latent variables (means). The bottom block displays the variance-covariance matrix of these latent variables. Columns in the table display the results for Spanish comprehension, effort and other unobserved traits, respectively.

The upper block of Table 2.11 indicates that the latent variables, $\theta_{c,i}$, $\theta_{e,i}$, and $\theta_{ut,i}$ are statistically significant to explain students' self-reported measures for Spanish comprehension, effort, and test scores.[17] See Tables 2.3 and 2.4 to recall the names of the factor loadings. In the context of Spanish comprehension, first column, factor loadings ($\lambda_.$) measure the

[15] Parental educational attainment stands for the maximum educational attainment of either the mother or the father.

[16] $1.618 = \frac{0.631}{0.152^{0.5}}$

[17] Recall from Section 2.4 that we employ test scores to identify the distribution of students' unobserved traits, see equation 2.5.

strength between the unobserved Spanish comprehension, $\theta_{c,i}$, and student's self-reported Spanish acquisition. For instance, the positive factor loading λ_4 suggests that the better the Spanish comprehension of a student, the more like the student speaks Spanish at school recess. As evident from column one, all measures related to Spanish thinking or speaking have a positive relationship with students' real level of Spanish comprehension. However, if non-SMT students speak their mother tongue, the fifth, sixth and seventh factor loadings suggest a low level of Spanish comprehension for non-SMT students.

In the context of effort and relative to how many hours students study, the second column in the table indicates that students that put high levels of effort are less likely to report that they receive help at home (λ_1), and that they are more likely to report that they always do their homework (λ_2). In the third column, the factor loading for other unobserved traits is displayed. This factor loading indicates that students who perform well at math tests also perform well at reading tests.

The middle block shows the estimates of the means of the latent variables. We discuss first the determinants of Spanish comprehension. An important finding emerges. The negative effect of being a non-SMT student on test scores is the result of non-SMT students' lack of Spanish comprehension. On average, non-SMT students have a lower Spanish comprehension equivalent to 1.62 standard deviations than SMT students. This lack of Spanish comprehension reduces students' math test scores by 0.215 standard deviations, which is equivalent to 50.2 percent of the observed math test gap. However, non-SMT students' Spanish comprehension is higher if they have at least one SMT parent at home. Parental education and student pre-elementary education are not as important as having an SMT parent at home to improve Spanish comprehension.

In the context of effort, see the column for $\theta_{e,i}$, estimates indicate parental education is one of the main driving forces for student effort. Students with a parent with an educational attainment of 6 years (complete elementary education) exert 1.20 standard deviations more effort than students whose parents did not finish elementary education.[18] This is in line with the international literature in which parental educational attainment has a positive impact on students' performance at school via students' effort. For instance, Houtenville and Conway (2008) find that parental education has a positive impact on how much parents exert effort to help their children at school. Fu and Mehta (2018) model parental effort as a direct input in an education production function. The authors find a positive relationship between parental effort and performance at school. We also find that student effort, on the other hand, is not influenced by their mother tongue or parental mother tongue.

For the last latent variable, students' other unobserved traits $\theta_{ut,i}$, the estimates indicate that, students' pre-elementary education, parents' mother tongue or parents' educational attainment do not influence the unobserved traits. What the factor model results reveal is that once we control for how students learn a subject, e.g., by understanding lectures,

[18] $1.195 = \frac{0.242}{0.041^{0.5}}$

exerting effort or speaking their mother tongue with teachers, unobserved traits do not depend on family or student background.

The bottom block of Table 2.11 shows the latent variables' variance-covariance matrix. There is a positive correlation (0.418) between students' Spanish comprehension and effort.[19] After controlling for student, parental and school characteristics in the means of latent variables, the correlation implies that if students understand lectures through their listening skills, they are more likely to do their homework or studying at home without any help.

To summarize this section, researches have provided economic interpretations for non-SMT students poor performance at school. The Guatemalan literature indicates that having a mother tongue different than Spanish reduces students' performance at school. Such a negative relationship disappears when controlling for students' Spanish comprehension in the learning process of math and reading. Here, we show that the source of such a negative relationship is specifically due to non-SMT students' lack of Spanish comprehension.

2.6 Conclusion

Understanding the mechanisms that drive educational achievement at school in a multicultural country is a relevant question for policymakers. In Guatemala 24 languages are spoken besides Spanish, and after third grade, the predominant language of instruction at school is Spanish. Third grade non-SMT students have been outperformed on national math and reading standardized tests by SMT students, and without any sign of improvement over academic years. Understanding such a poor performance of non-SMT students has been challenging. A number of papers argue that non-SMT students' poor performance on national tests is the result of low quality schools where these students attend. However, other papers indicate that this poor performance is due to a low parental education for instance. This lack of consensus about the main driving forces of test score gaps between non-SMT and SMT students suggests that previous papers have been plagued by omitted variable bias. Finding out whether non-SMT students are not yet proficient in Spanish is crucial, since facing educational challenges while at school put at risk the well-being of the Guatemalan non-SMT population. This chapter contributes to the Guatemalan literature by looking for evidence that non-SMT students are not yet proficient in Spanish, which affects their performance at school.

This chapter uses data for third grade students in 2006, since recent datasets for students do not include questions to infer students' Spanish proficiency and effort. The jointly estimation of a correlated factor model and production functions allow us to estimate the effect of students' Spanish comprehension and effort on the students' learning process of math and

[19] $0.418 = \frac{0.033}{0.152^{0.5} * 0.041^{0.5}}$

reading. Relative to SMT students, we find that non-SMT students have a lower level of Spanish comprehension, which is affecting their math and reading educational achievements at school. Spanish comprehension can entirely account for the observed math and reading test score gaps between non-SMT and SMT students. This finding is robust to model specifications.

In terms of policy recommendations, the findings in this chapter suggest that the Ministry of Education should carefully and continuously monitor educational achievement in elementary schools as opposed to high school students. Growing up facing educational challenges may be shaping the low educational attainment of the Guatemalan non-SMT population. Early interventions to mitigate educational challenges such as the low Spanish language comprehension of non-SMT students may have important effects on their well-being in the long-run.

Bibliography

Maria Elena Anderson. Guatemala : the education sector. Guatemala Poverty Assessment (GUAPA) Program. Technical report, World Bank, Washington, DC, 2001. URL http://documents.worldbank.org/curated/en/584251468031751174/Guatemala-the-education-sector.

Pedro Carneiro, Karsten T. Hansen, and James Heckman. Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice. Technical report, Institute for the Study of Labor (IZA), April 2003. URL https://EconPapers.repec.org/RePEc:iza:izadps:dp767.

Rubiana Chamarbagwala and Hilcías E. Morán. The human capital consequences of civil war: Evidence from Guatemala. *Journal of Development Economics*, 94(1):41–61, January 2011. ISSN 0304-3878. doi: 10.1016/j.jdeveco.2010.01.005. URL http://www.sciencedirect.com/science/article/pii/S0304387810000076.

Flavio Cunha, James J. Heckman, and Susanne M. Schennach. Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Econometrica*, 78(3):883–931, May 2010. ISSN 0012-9682. doi: 10.3982/ECTA6551. URL https://doi.org/10.3982/ECTA6551. Publisher: John Wiley & Sons, Ltd.

Kjell I. Enge and Ray Chesterfield. Bilingual education and student performance in Guatemala. *International Journal of Educational Development*, 16(3):291–302, July 1996. ISSN 07380593. doi: 10.1016/0738-0593(95)00038-0. URL http://linkinghub.elsevier.com/retrieve/pii/0738059395000380.

Álvaro Fortín. Evaluación Educativa Estandarizada en Guatemala: Un camino recorrido, un camino por recorrer. Technical report, Ministerio de Educación, Guatemala, 2013.

Jane Cooley Fruehwirth, Salvador Navarro, and Yuya Takahashi. How the Timing of Grade Retention Affects Outcomes: Identification and Estimation of Time-Varying Treatment Effects. *Journal of Labor Economics*, 34(4):979–1021, February 2016. ISSN 0734-306X. doi: 10.1086/686262. URL https://doi.org/10.1086/686262.

Chao Fu and Nirav Mehta. Ability Tracking, School and Parental Effort, and Student Achievement: A Structural Model and Estimation. *Journal of Labor Economics*, 36(4): 923–979, October 2018. ISSN 0734-306X. doi: 10.1086/697559. URL https://doi.org/10.1086/697559. Publisher: The University of Chicago Press.

Martha Hernandez-Zavala, Harry Anthony Patrinos, Chris Sakellariou, and Joseph Shapiro. *Quality Of Schooling And Quality Of Schools For Indigenous Students In Guatemala, Mexico, And Peru*. Policy Research Working Papers. The World Bank, August 2006. doi: 10.1596/1813-9450-3982. URL http://elibrary.worldbank.org/doi/book/10.1596/1813-9450-3982.

Andrew J. Houtenville and Karen Smith Conway. Parental Effort, School Resources, and Student Achievement. *The Journal of Human Resources*, 43(2):437–453, 2008. ISSN 0022166X. URL www.jstor.org/stable/40057353. Publisher: [University of Wisconsin Press, Board of Regents of the University of Wisconsin System].

Jeffery H. Marshall. School quality and learning gains in rural Guatemala. *Economics of Education Review*, 28(2):207–216, April 2009. ISSN 0272-7757. doi: 10.1016/j.econedurev.2007.10.009. URL http://www.sciencedirect.com/science/article/pii/S0272775708000745.

Jeffery H. Marshall. School quality signals and attendance in rural Guatemala. *Special Issue: Economic Returns to Education*, 30(6):1445–1455, December 2011. ISSN 0272-7757. doi: 10.1016/j.econedurev.2011.07.011. URL http://www.sciencedirect.com/science/article/pii/S0272775711001142.

Jeffery H. Marshall and M. Alejandra Sorto. The effects of teacher mathematics knowledge and pedagogy on student achievement in rural Guatemala. *International Review of Education / Internationale Zeitschrift für Erziehungswissenschaft / Revue Internationale de l'Education*, 58(2):173–197, 2012. ISSN 00208566, 15730638. URL www.jstor.org/stable/41502402. Publisher: Springer.

Patrick J. McEwan and Marisol Trowbridge. The achievement of indigenous students in Guatemalan primary schools. *International Journal of Educational Development*, 27(1):61–76, January 2007. ISSN 0738-0593. doi: 10.1016/j.ijedudev.2006.05.004. URL http://www.sciencedirect.com/science/article/pii/S0738059306000502.

Benjamin Meade. *Examining the structural roots of achievement disparities in Guatemalan primary schools*. PhD thesis, 2011. URL https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi?url=http://search.proquest.com/docview/851547825?accountid=15115.

Harry Anthony Patrinos and Eduardo Velez. Costs and benefits of bilingual education in Guatemala: A partial analysis. *International Journal of Educational Development*, 29(6):594–598, November 2009. ISSN 0738-0593. doi: 10.1016/j.ijedudev.2009.02.001. URL http://www.sciencedirect.com/science/article/pii/S0738059309000182.

F.E. Rubio. Educación Bilingüe en Guatemala: Situación y desafios. 2004.

Stinebrickner Ralph and Stinebrickner Todd R. The Causal Effect of Studying on Academic Performance. *The B.E. Journal of Economic Analysis & Policy*, 8(1), 2008. ISSN 19351682. doi: 10.2202/1935-1682.1868. URL https://www.degruyter.com/view/j/bejeap.2008.8.1/bejeap.2008.8.1.1868/bejeap.2008.8.1.1868.xml.

María Toledo and Margarito Guantá. MODELO EDUCATIVO BILINGÜE E INTERCULTURAL. Technical report, Ministerio de Educación Guatemala, 2009. URL http://www.mineduc.gob.gt/DIGEBI/documents/modeloEBI.pdf.

2.7 Tables

Table 2.1: Third grade students' math and reading scores on standardized tests by academic year and students' mother tongue and ethnicity

Math test scores

	Spanish Mother Tongue			Indigenous		
	No	Yes	Gap	Yes	No	Gap
2006	-0.3331	0.0953	-0.428***	-0.1610	0.0841	-0.245***
2007				-0.2788	0.1621	-0.441***
2008				-0.2847	0.1865	-0.471***
2010				-0.2168	0.2332	-0.450***
2013				-0.3102	0.2625	-0.573***
2014	-0.3127	0.1717	-0.485***	-0.1999	0.3160	-0.516***

Reading test scores

	Spanish Mother Tongue			Indigenous		
	No	Yes	Gap	Yes	No	Gap
2006	-0.422	0.121	-0.543***	-0.254	0.134	-0.388***
2007				-0.428	0.248	-0.676***
2008				-0.289	0.189	-0.478***
2010				-0.247	0.267	-0.514***
2013				-0.335	0.288	-0.623***
2014	-0.390	0.214	-0.604***	-0.229	0.361	-0.590***

Note: Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. These students represent the 77.7 percent of the sample size. Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. The indigenous classification represents students' self-report of being indigenous, which may or may not depend on students' mother tongue. The Ministry of Education in Guatemala does not frequently ask students their mother tongue. Test scores have a mean of zero and standard deviation of one at each academic year. Significant levels: *** at 1%, ** at 5%, and * at 10%.
Source: Authors' calculation using the Ministry of Education's 2006-2014 data.

Table 2.2: Educational attainment distribution in Guatemala by gender and mother tongue

	Male			Female		
	(a) Non-SMT	(b) SMT	Gap (a-b)	(a) Non-SMT	(b) SMT	Gap (a-b)
Years of complete education						
0	0.448	0.193	0.255***	0.698	0.272	0.425***
1-3	0.263	0.208	0.055***	0.159	0.202	−0.043***
4-6	0.205	0.259	−0.054***	0.103	0.229	−0.126***
7-9	0.039	0.105	−0.066***	0.018	0.084	−0.067***
10-12	0.035	0.142	−0.106***	0.018	0.145	−0.127***
>12	0.009	0.093	−0.084***	0.004	0.066	−0.062***
Total	1	1		1	1	

Note: SMT stands for people's self-report of having Spanish as a first language. Non-SMT people speaks either one of the Mayan languages, or Xinka or Garífuna languages as a first language. The educational system in Guatemala consists of 3 stages: elementary, junior high, and high school. The elementary stage, consists of 6 grades and usually starts in the year the child turns 7, as the school year begins in January. The second and third stages, junior high and high school, consists of three grades. Attendance to college means having more than 12 years of complete education. Significant levels: *** at 1%, ** at 5%, and * at 10%.
Source: Authors' calculation using the 2002 Guatemalan national census.

Table 2.3: Measures for students' Spanish comprehension

	(a) Non-SMT students	(b) SMT students	Gap (a-b)
Variables for constructing Spanish comprehension ($\theta_{c,i}$)			
Spanish thinking when			
Speaking ($\lambda_0 = 1$)	0.770	0.917	−0.147***
Subtracting numbers (λ_1)	0.802	0.908	−0.106***
Writing (λ_2)	0.823	0.916	−0.093***
Spanish language use with (at)			
My siblings (λ_3)	0.266	0.903	−0.637***
School's recess (λ_4)	0.643	0.900	−0.257***
Non-Spanish language use with (at)			
My siblings (λ_5)	0.766	0.043	0.723***
At least one parent (λ_6)	0.848	0.064	0.783***
School's recess (λ_7)	0.294	0.021	0.272***

Note: Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. These students represent the 77.7 percent of the sample size. Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. Each measure for Spanish comprehension is a binary variable, taking a value of one if students report Spanish thinking when speaking for instance and zero otherwise. Therefore, numbers in this table represent proportions. The symbols in parenthesis, λ_i for i=1..7, are the associated factor loadings when estimating the latent variable for Spanish comprehension. Significant levels: *** at 1%, ** at 5%, and * at 10%.
Source: Authors' calculation using the Ministry of Education's 2006 data.

Table 2.4: Measures for students' effort

	(a) Non-SMT studens	(b) SMT students	Gap (a-b)
Variables for constructing effort ($\theta_{e,i}$)			
How many hours do you study? ($\lambda_0 = 1$)			
No	0.280	0.248	0.032***
< 1	0.479	0.434	0.045***
$\geqslant 1$ and < 2	0.150	0.186	−0.036***
$\geqslant 2$ and < 3	0.031	0.049	−0.017***
$\geqslant 3$	0.060	0.083	−0.024***
No help at home (λ_1)	0.638	0.674	−0.036***
Always do homework (λ_2)	0.793	0.791	0.002

Note: Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. These students represent the 77.7 percent of the sample size. Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. In this table, each category of each measure for effort is a binary variable, taking a value of one if students report always doing homework for instance, and zero otherwise. Therefore, numbers in this table represent proportions. The symbols in parenthesis, λ_i, are the associated factor loadings when estimating the effort latent variable. Significant levels: *** at 1%, ** at 5%, and * at 10%.
Source: Authors' calculation using the Ministry of Education's 2006 data.

Table 2.5: Variables commonly use in empirical studies for Guatemala

	(a) Non-SMT students	(b) SMT students	Gap (a-b)
Math test score	−0.333	0.095	−0.428***
Reading test score	−0.422	0.121	−0.543***
Pre-elementary education	0.691	0.706	−0.014*
Speak non-Spanish with teachers	0.448	0.179	0.269***
Bilingual program at school	0.170	0.032	0.138***
SMT mother	0.202	0.853	−0.651***
SMT father	0.250	0.867	−0.617***
Mother's educational attainment			
No	0.497	0.338	0.158***
Elementary	0.292	0.275	0.017**
Junior high school	0.130	0.194	−0.064***
High school	0.042	0.070	−0.029***
Collegue or higher	0.039	0.122	−0.083***
Father's educational attainment			
No	0.367	0.292	0.075***
Elementary	0.347	0.269	0.078***
Junior high school	0.196	0.221	−0.025***
High school	0.036	0.069	−0.033***
Collegue or higher	0.055	0.149	−0.094***

Note: Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. These students represent the 77.7 percent of the sample size. Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. In this table, each category of each variable is a binary variable (except test scores), taking a value of one if students self-report having attended pre-elementary education for instance, and zero otherwise. Therefore, numbers in this table represent proportions. Test scores have a mean of zero and standard deviation of one. Significant levels: *** at 1%, ** at 5%, and * at 10%.
Source: Authors' calculation using the Ministry of Education's 2006 data.

Table 2.6: Estimates for math production functions

	OLS (1)	OLS (2)	OLS (3)	OLS (4)	OLS (5)	Factor model(*)
Student's characteristics						
Non-SMT student	0.089	−0.120***	−0.128***	−0.056	−0.048	0.023
Non-SMT * Rural area	−0.370***	0.054	0.044	0.026	0.016	−0.011
Speak non-Spanish with teachers	−0.258***	−0.181***	−0.165***	−0.155***	−0.099***	−0.150***
Latent variables						
Spanish comprehension $E(\theta_{c,i})$						0.355***
Effort $E(\theta_{e,i})$						2.109***
Unobserved traits $E(\theta_{ut,i})$						1.000
Controls						
Student's pre-elementary education	Yes	Yes	Yes	Yes	Yes	
Parents' mother tongue	Yes	Yes	Yes	Yes	Yes	
Parents' educational attainment	Yes	Yes	Yes	Yes	Yes	
School fixed effects		Yes	Yes	Yes	Yes	
Effort measures			Yes	Yes	Yes	
Spanish comprehension measures				Yes	Yes	
Reading test scores					Yes	

Note: The dependent variable is students' math test scores. This variable has a mean of zero and standard deviation of one. Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. These students represent the 77.7 percent of the sample size. Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. Standard errors are clustered at school level. Significant levels: *** at 1%, ** at 5%, and * at 10%.
(*) The factor model and the system of test scores are jointly estimated. The means of latent variables include school fixed effects.
Source: Authors' calculation using the Ministry of Education's 2006 data.

Table 2.7: Estimates for reading production functions

	OLS (1)	OLS (2)	OLS (3)	OLS (4)	OLS (5)	Factor model(*)
Student's characteristics						
Non-SMT student	0.035	−0.118***	−0.120***	−0.033	−0.040	−0.008
Non-SMT * Rural area	−0.370***	0.030	0.015	0.001	0.029	−0.055
Speak non-Spanish with teachers	−0.289***	−0.195***	−0.179***	−0.171***	−0.104***	−0.188***
Latent variables						
Spanish comprehension $E(\theta_{c,i})$						0.382***
Effort $E(\theta_{e,i})$						1.848***
Unobserved traits $E(\theta_{ut,i})$						0.879***
Controls						
Student's pre-elementary education	Yes	Yes	Yes	Yes	Yes	
Parents' mother tongue	Yes	Yes	Yes	Yes	Yes	
Parents' educational attainment	Yes	Yes	Yes	Yes	Yes	
School fixed effects		Yes	Yes	Yes	Yes	
Effort measures			Yes	Yes	Yes	
Spanish comprehension measures				Yes	Yes	
Math test scores					Yes	

Note: The dependent variable is students' reading test scores. This variable has a mean of zero and standard deviation of one. Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. These students represent the 77.7 percent of the sample size. Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. Standard errors are clustered at school level. Significant levels: *** at 1%, ** at 5%, and * at 10%.
(*) The factor model and the system of test scores are jointly estimated. The means of latent variables include school fixed effects.
Source: Authors' calculation using the Ministry of Education's 2006 data.

Table 2.8: Estimates for math production functions when controlling for non-SMT students' mother tongue use with their teachers

	OLS (1)	OLS (2)	OLS (3)	OLS (4)	OLS (5)	Factor model(*)
Student's characteristics						
Non-SMT student	0.016	−0.177***	−0.179***	−0.106***	−0.078*	0.040
Non-SMT * Rural area	−0.361***	0.001	−0.004	−0.016	0.005	−0.016
Speak non-Spanish with teachers	−0.347***	−0.245***	−0.222***	−0.212***	−0.135***	−0.200***
Non-SMT student * Speak non-Spanish with teachers	0.268***	0.206***	0.186***	0.186***	0.116**	0.171*
Non-SMT student * Bilingual program at school	−0.163	0.211*	0.196*	0.161	0.003	0.133
Latent variables						
Spanish comprehension $E(\theta_{c,i})$						0.341***
Effort $E(\theta_{e,i})$						0.581***
Unobserved traits $E(\theta_{ut,i})$						1.000
Controls						
Student's pre-elementary education	Yes	Yes	Yes	Yes	Yes	
Parents' mother tongue	Yes	Yes	Yes	Yes	Yes	
Parents' educational attainment	Yes	Yes	Yes	Yes	Yes	
School fixed effects		Yes	Yes	Yes	Yes	
Effort measures			Yes	Yes	Yes	
Spanish comprehension measures				Yes	Yes	
Reading test scores					Yes	

Note: The dependent variable is students' math test scores. This variable has a mean of zero and standard deviation of one. Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. These students represent the 77.7 percent of the sample size. Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. Standard errors are clustered at school level. Significant levels: *** at 1%, ** at 5%, and * at 10%.
(*) The factor model and the system of test scores are jointly estimated. The means of latent variables include school fixed effects.
Source: Authors' calculation using the Ministry of Education's 2006 data.

Table 2.9: Estimates for reading production functions when controlling for non-SMT students' mother tongue use with their teachers

	OLS (1)	OLS (2)	OLS (3)	OLS (4)	OLS (5)	Factor model(*)
Student's characteristics						
Non-SMT student	−0.045	−0.171***	−0.167***	−0.081*	−0.068	−0.051
Non-SMT * Rural area	−0.370***	−0.008	−0.018	−0.028	0.002	−0.095
Speak non-Spanish with teachers	−0.381***	−0.253***	−0.231***	−0.225***	−0.135***	−0.269***
Non-SMT student * Speak non-Spanish with teachers	0.277***	0.185***	0.165***	0.173***	0.100**	0.237***
Non-SMT student * Bilingual program at school	−0.115	0.135	0.118	0.087	0.123	0.143
Latent variables						
Spanish comprehension $E(\theta_{c,i})$						0.333***
Effort $E(\theta_{e,i})$						0.485***
Unobserved traits $E(\theta_{ut,i})$						0.842***
Controls						
Student's pre-elementary education	Yes	Yes	Yes	Yes	Yes	
Parents' mother tongue	Yes	Yes	Yes	Yes	Yes	
Parents' educational attainment	Yes	Yes	Yes	Yes	Yes	
School fixed effects		Yes	Yes	Yes	Yes	
Effort measures			Yes	Yes	Yes	
Spanish comprehension measures				Yes	Yes	
Math test scores					Yes	

Note: The dependent variable is students' reading test scores. This variable has a mean of zero and standard deviation of one. Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. These students represent the 77.7 percent of the sample size. Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. Standard errors are clustered at school level. Significant levels: *** at 1%, ** at 5%, and * at 10%.
(*) The factor model and the system of test scores are jointly estimated. The means of latent variables include school fixed effects.
Source: Authors' calculation using the Ministry of Education's 2006 data.

Table 2.10: Contribution of students' mother tongue, Spanish comprehension and effort to the observed math and reading test score gaps

	Math test score gaps			Reading test score gaps		
	OLS (2)	OLS (4)	Factor model(+)	OLS (2)	OLS (4)	Factor model(+)
Differences in						
Non-SMT student	−0.177	−0.106	0.040	−0.171	−0.081	−0.051
Spanish comprehension		−0.192	−0.532		−0.208	−0.519
Effort		−0.007	0.246		−0.008	0.205
Observed test score gaps		-0.428			-0.543	

Note: test scores have a mean of zero and standard deviation of one. The contributions to the gap use the most likely values of the latent variables given the data we observed. Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. These students represent the 77.7 percent of the sample size. Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. Models include school fixed effects.
(+) The factor model and the system of test scores are jointly estimated.
Source: Authors' calculation using the Ministry of Education's 2006 data.

Table 2.11: Determinants of students' Spanish comprehension, effort and other unobserved traits

	θ_c	θ_e	θ_{ut}
λ_1	0.807***	−0.889***	0.842***
λ_2	0.848***	0.262***	
λ_3	3.192***		
λ_4	1.178***		
λ_5	−2.541***		
λ_6	−2.253***		
λ_7	−1.537***		
Means			
Student characteristics			
Non-SMT student	−0.631***	0.065	
Non-SMT student * Rural area	0.077**	0.025	
Pre-elementary education	0.034***	0.085*	0.020
Parental characteristics[(a)]			
SMT mother	0.247***	0.140	−0.169
SMT father	0.414***	0.133	−0.137
Two SMT parents	0.213***	−0.047	0.069
Elementary school	0.071***	0.242***	0.036
Junior high school	0.075***	0.237***	−0.068
High school	0.011	0.273***	−0.051
More than high school	0.059***	0.294***	−0.014
Variance-Covariance matrix			
Spanish comprehension	0.152		
Effort	0.033***	0.041	
Unobserved traits	−0.025**	0.044	0.139

Note: The variables θ_c, θ_e and θ_{ut} stand for the latent variables for Spanish comprehension, effort and other unobserved traits for student i. See Tables 2.3 and 2.4 for the factor loadings' names. Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. These students represent the 77.7 percent of the sample size. Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. Standard errors are clustered at school level. Significant levels: *** at 1 % , ** at 5%, and * at 10%. Besides variables shown in the table, the means of latent variables include school fixed effects.
(a) Parental educational attainment stands for the maximum educational attainment reached by either the mother or father.
Source: Authors' calculation using the Ministry of Education's 2006 data.

Chapter 3

Are non-Spanish mother tongue students at secondary school still facing a language barrier in the Guatemalan education system?

3.1 Introduction

Not being proficient in a school's predominant language of instruction can represent a language barrier for students' human capital development. Furthermore, if demographic characteristics affect the language learning process, and these characteristics are not taken into account in the education process, then the existence of such a language barrier may not only hinder student's human capital accumulation, but also a community's long-term development (Lundberg and Startz, 1998). In the context of Guatemala, Spanish comprehension plays a key role for human capital development in elementary schools (see Chapter 2). In this third chapter, I investigate whether non-Spanish mother tongue (non-SMT) students still face an obstacle or a language barrier in acquiring skills while at secondary school in Guatemala.

Knowing whether students face a language barrier is relevant for Guatemala due to its multilingual society. In Guatemala, 25 languages are spoken, and Spanish is the language of instruction in the majority of secondary schools. The Ministry of Education (Mineduc) carried out national math and reading tests in 2010, 2013 and 2015 to assess students' achievement in the last grades of elementary, junior high and high school, respectively, which represent grades 6, 9, and 12 in the Guatemalan education system. Those students whose mother tongue is Spanish represented 62% of the student population in grade 6 in

2010, while the remaining students spoke a non-Spanish language as a mother tongue.[1] The academic performance of students on the national tests is displayed in Table 3.1 for the last grades of elementary, junior high and high school. Test scores are standardized to have a mean of zero and a standard deviation of one at each grade. Scores on the national tests for SMT students are higher than for non-SMT students. Furthermore, the low performance of non-SMT students was persistent from lower to higher grades, and without any significant improvement between elementary and high school. This achievement gap between non-SMT and SMT students raises the question of whether non-SMT students still face a language barrier at school. If true, it is crucial to identify the extent of the damage in terms of school performance and the root of the language barrier.

The analysis in this chapter exploits unique data from the Guatemalan Ministry of Education to build a panel dataset. This new dataset represents the first panel data available for Guatemalan secondary students. It starts with students who attended the last grade of elementary school in 2010. Both students and principals were surveyed to measure family background and school quality, respectively, in different years. By combining longitudinal student survey data with school survey data, this chapter represents, to my knowledge, the first Guatemalan longitudinal study for secondary students. The construction of this dataset allows me to model the acquisition of human capital as a cumulative process in which not only school and family inputs matter, but also the students' innate ability that affects their performance at school (Todd and Wolpin, 2003, 2007; Hanushek et al., 2009; Heckman and Raut, 2016).

A limitation in this chapter is that the dataset does not include questions to estimate students' latent variables for Spanish comprehension and effort like the ones used in Chapter 2, where both latent variables play a key role for human capital development. However, the dataset in this third chapter includes non-SMT students' self-evaluation of not being proficient in Spanish, and of their mother tongue use.[2] To determine whether non-SMT students face a language barrier at secondary school, I hypothesize that students' lack of Spanish comprehension negatively affects learning at school, or in other words, students who self-evaluate as not being proficient in Spanish perform worse on national tests than their counterparts. I test this hypothesis in the empirical section. This lack of Spanish comprehension would put non-SMT students in a disadvantaged position because Spanish is the language of instruction in the majority of secondary schools.

The lack of information that can be used to pin-down students' effort may lead to overestimation of the language barrier effect at school. The psychology literature has emphasized the importance of non-cognitive skills such as motivation and persistence on students' performance while at school. These non-cognitive skills determine students' effort and, as a

[1] Non-Spanish mother tongue students are those who speak either a Mayan, Xinca or Garífuna language.

[2] Students did not report to whom they speak to. Chapter 2 also shows that when elementary non-SMT students are still learning Spanish, the intensity of their mother tongue use will negatively affect their Spanish comprehension.

result, test scores (Aiken, 1971; Wolfe and Johnson, 1995; Duckworth and Seligman, 2005; Heckman et al., 2006). Almlund et al. (2011) indicate that students with low abilities can offset their poor performance at school by exerting more effort, and Beattie et al. (2017) show that students who poorly perform on tests frequently report that they are more often stressed than their classmates. Therefore, if non-SMT students do poorly on a test as a result of their lack of Spanish comprehension, they can choose a strategy to overcome their low grade.[3] Non-SMT students, for example, can decide to learn from teachers by using their mother tongue or by spending more time reading books by themselves. Such strategies can reduce non-SMT students' interaction time with students proficient in Spanish, which may affect non-SMT students' Spanish learning process. In other words, students' effort can display a positive relationship with their lack of Spanish comprehension (see Chapter 2), thus causing biased estimates of the language barrier effect on test scores at school.

Due to the lack of information that is needed to separate the independent impact of students' lack of Spanish comprehension from other factors driving student achievement such as effort or innate ability, I follow three strategies to investigate the sensitivity of the estimated impacts to potential confounding factors. First, I estimate OLS and IV models to determine the effect of the lack of Spanish comprehension on tests scores with extensive sets of control variables, including student and family background, general and subject-specific school inputs, and regional dummy variables. I refer to these OLS models as baseline models. These OLS models are comparable to and produce similar results to those in the existing literature on Guatemalan education.

Second, I use fixed-effect models to exploit differences in the performance of students across time. This student fixed-effects analysis allows me to identify the effect of a lack of Spanish comprehension using only variation between subjects, thereby directly controlling for unobserved student-specific characteristics that similarly affect math and reading performance (e.g., innate ability or family background). I also estimate OLS and IV models when controlling for student fixed effects.

Third, both the baseline specification and the fixed-effect models can still provide biased estimates of the language barrier effect or the effect of a lack of Spanish comprehension on test scores if, for instance, the effect of student effort on test scores is influenced by students' Spanish comprehension (essential heterogeneity). Non-SMT students' unobserved effort may display a positive relationship with their lack of Spanish comprehension. If so, the specification of a model for non-SMT students' lack of Spanish comprehension is essential for interpreting the IV estimates. Following the local instrumental variable literature (Basu et al., 2007; Carneiro et al., 2010), I estimate the average treatment effect, ATE, to provide supporting evidence.

While identification of causal effects of non-SMT students' lack of Spanish comprehension

[3]The Guatemalan education system requires students to score 60% in all subjects to progress to the next grade.

on test scores is clearly difficult, all empirical strategies consistently indicate a robust negative relationship between non-SMT students' self-evaluation of not being proficient in Spanish and their achievement on national tests. This robust negative relationship paired with the hypothesis that non-SMT students are not yet proficient in Spanish at secondary school provides evidence that non-SMT students' human capital development is at risk. In the baseline model estimated by OLS, I do not find evidence that reporting not being proficient in Spanish affects student performance on national tests. However, the baseline models estimated by IV indicate that reporting not being proficient in Spanish is associated with a lower math achievement ranging from -1.523 to -0.523 standard deviations. This finding also holds when controlling for student fixed effects and students' essential heterogeneity. To put these estimates into perspective, they imply that the math test score gap is entirely explained by non-SMT students' lack of Spanish comprehension. The consistency of estimated impacts across alternative estimation approaches supports the finding that non-SMT students' lack of Spanish comprehension negatively affect their learning at school.

The first stage in the two stage least squares provides an understanding of non-SMT students' lack of Spanish comprehension. Non-SMT students' lack of Spanish comprehension could have originated from their lack of interaction with people proficient in Spanish such as SMT students at school. Therefore, a potential instrument for Spanish comprehension at time t can be the linguistic composition at school in period $t-1$, where non-SMT students attended.[4] I find that linguistic composition at school in time $t-1$ influences both non-SMT students' lack of Spanish comprehension and mother tongue use in time t, which affects the students' Spanish learning process. This finding is robust even after controlling for both distance to main cities where Spanish is mainly spoken (main municipal cities) and for municipal and rural area fixed effects.[5] In other words, SMT students' sorting among schools within a local area can explain both non-SMT students' lack of Spanish comprehension and their mother tongue use. This finding implies that the regional distribution of the non-SMT population across Guatemala is not the root of non-SMT students' language barrier as I discuss in Chapter 4.

Due to the negative effect of non-SMT students' lack of Spanish comprehension on their learning at school, this chapter also explores a policy scenario by which the government might improve non-SMT students' Spanish comprehension and, therefore, their performance. If a language barrier exists at school, keeping non-SMT students segregated may not increase their Spanish comprehension, especially for students attending schools in rural areas. I follow Heckman and Vytlacil (2005) to simulate the effect of what can be understood as a second language program. The policy intervention in this chapter simulates a linguistic integration at time $t-1$ where SMT and non-SMT students are artificially mixed in the

[4] In this chapter a difference of one period in the model means a difference of three years in the data.

[5] Guatemala is geographically divided into eight regions. Each region is divided into departments, and departments into municipalities. In total Guatemala has 22 departments and 337 municipalities. Each municipality consists of rural and urban areas, and every one of them has its own main city, many communities. Municipalities are the smallest geographical division observed in this data set.

same school in rural areas. The artificial linguistic integration at school may change the distribution of students proficient in Spanish at time t. This policy simulation serves as a thought experiment that sheds some light on one possible solution given that parents can enroll their children at any school.

The simulation of the second language program provides evidence that linguistic integration at school may play an important role in alleviating the language barrier for non-SMT students.[6] Non-SMT students attending rural schools improve their performance under the policy simulation, but they still do not perform as good as native speakers. However, the feasibility of such a policy is beyond the scope of this research, since such a policy would heavily depend on a deep understanding of how parents select schools.

The next section gives a brief description of the Guatemalan educational literature. The following section gives a description of both the Guatemalan education system and the data used in this chapter. Section 3.4 explains the empirical approach. Section 3.5 discusses the results for the math and reading production functions. Section 3.6 presents the policy simulation. The concluding section provides further discussion of the results.

3.2 Literature review

The Guatemalan educational literature shows mixed findings when trying to understand why non-SMT students do worse than SMT students at elementary school. The three hypotheses in the literature for the test gap are: differences in household background, differences in school qualities, or differences in student Spanish comprehension.

The low achievement on national tests for non-SMT students may be related to their family's socio-economic status, such as parents' low educational attainment or low family income. On the one hand, one part of the existing Guatemalan literature shows that indigenous people have low educational attainment (Chamarbagwala and Morán, 2011). Therefore, differences in educational attainment can explain the earning gap between indigenous and non-indigenous people. In this context, Meade (2011) finds that students' mother tongue and parents' level of schooling affect student achievement at Guatemalan elementary school. Hernandez-Zavala et al. (2006) analyze test score gaps for students who grew up or not speaking an indigenous language at home for Guatemala, Mexico and Peru. The authors find that the gap in test scores among the two groups is mostly explained by family factors such as parents' educational level. For instance, the Guatemalan census in 2002 shows the clear gap in the education levels of SMT and non-SMT parents, especially mothers, of two year old children (see Table 3.2).

[6]The estimation approach for this analysis follows the work of Heckman and Vytlacil (2005), Carneiro and Lee (2009), Carneiro et al. (2010) and Carneiro et al. (2010) in estimating the policy relevant treatment effect.

On the other hand, some empirical papers for Guatemala provide evidence that the elementary student achievement gap is mainly driven by school factors. Enge and Chesterfield (1996) find evidence that teachers' training and the supply of instructional material in the students' native languages have a significant impact on rural elementary student achievement in Guatemala. Rubio (2004) shows that indigenous students in the first six grades attending schools with the bilingual education program have a higher passing rate, lower dropout rate and larger number of graduates. McEwan and Trowbridge (2007a) analyze the educational achievement of rural third grade students in 2001. Their estimates imply that students who speak a Mayan language at home perform poorly on national tests due to schools' low quality (as measured by school fixed effects) they attend. Marshall (2009) finds evidence that children's school attendance influences both Spanish and math test scores, but also that community fixed effects play a key role in explaining students' achievement in national tests by ethnicity.

In contrast to the previous Guatemalan educational literature, Chapter 2 shows that Spanish comprehension plays a key role for human capital development in Guatemala. The second chapter indicates that non-SMT students with a low level of Spanish comprehension perform poorly at elementary school, but these students can offset their lack of Spanish comprehension by either speaking their mother tongue with teachers or exerting more effort. By doing so, however, non-SMT students hinder their Spanish learning process.

Identifying which of the above-mentioned hypotheses influence the achievement gap for secondary students is crucial given the high degree of inequality between the indigenous and non-indigenous populations in Guatemala.

3.3 The Guatemalan education system and the data

The Guatemalan education system can be divided into four school types: public, private, cooperative, and municipal. Public schools are completely managed by Mineduc. Private schools are profit maximizing. Municipal schools are managed by local governments rather than by Mineduc. Last, in the cooperative schools, the schools' administration is formed by parents and the local government. Regardless of the school type, the curriculum and all syllabi are designed by Mineduc, and all schools are required to cover the Mineduc curriculum.

In 2010 Mineduc randomly selected and tested 18,441 grade six students in math and reading to nationally represent students in public schools.[7] By the time students take the tests, students also completed a survey asking questions about school, family inputs and house infrastructure. The students survey is also designed by Mineduc. The tests and the

[7]National tests are marked based on Item Response Theory (Rasch correction model), which involves grading students not only by the number of correct answers, but also by the degree of difficulty of each question.

survey were not planned for use in a longitudinal analysis, but the creation of the longitudinal data set is made possible by the fact that all junior high and high school students were tested in 2013 and 2015, so that students who took the 2010 tests can be followed in the later grades using their names.

In addition to the student-level data, school-level data are available for all schools in 2013 and 2015. Each year principals provide information in the following areas: 1) the principal's own education, 2) school infrastructure, and 3) the students' preparation for the national tests. This second dataset covers all junior high and high schools. Both datasets contain a school code variable that allows matching of students to schools. This matching allows for the construction of a longitudinal student-level panel dataset representing the 2010 elementary student cohort that includes both student and school data.

The dataset in this chapter is unbalanced because grade six students who did not progress or dropped out of school are not included in the higher grades. Table 3.3 shows the relationship between students' mother tongue and both students' grade progression and their test scores from the longitudinal dataset. Urban and rural sectors are shown separately, as grade progression rates differ markedly between the two sectors. In terms of non-SMT students, the table shows that non-SMT students account for 38 percent of the sample and that 81 percent of these non-SMT students live in rural areas.[8] Non-SMT students attending rural schools have both the lowest grade progression rates and the lowest performance at school. The fact that SMT students in rural area junior and senior high schools outperform non-SMT students at rural school suggests that language of instruction at school may be a barrier for non-SMT students. For instance, conditional on attending rural schools and in terms of math scores, SMT students outperform non-SMT students by 0.218 and 0.199 standard deviations for junior high and high school respectively. Regarding students who attend urban schools, SMT students also have better performance than non-SMT students.

Are the achievement gaps between SMT and non-SMT students entirely due to socio-economic differences such as family income and parental education? Are these achievement gaps the result of differences in school attributes? Or, do they reflect a lack of Spanish comprehension? Table 3.4 gives the complete list of variables (and descriptive statistics) employed in this chapter to identify which of these three hypotheses from the Guatemalan literature explain secondary students' achievement gap. This table is divided by students' mother tongue. Furthermore, among non-SMT students, the table contrasts student characteristics between rural and urban areas. Given the large number of variables, I only discuss the main differences.

Regarding the first hypothesis, socio-economic differences by mother tongue, student-level variables from the student dataset include age, gender, family assets, remittances, parental educational, and books and newspapers read. With respect to non-SMT students'

[8]The 2002 national census indicates that about 39 percent of the population does not have Spanish as a mother tongue.

mother tongue use (see first row), the table indicates that the same proportion of students, 0.63, speak their non Spanish mother tongue between rural and urban areas. However, non-SMT students who live in rural areas are older than students who attend schools in urban areas. Non-SMT students who live in rural areas are 1.34 years older than the expected age while urban non-SMT students are just 0.6 years older. Another difference is that non-SMT students who reside in rural areas read fewer books and newspapers than non-SMT students in urban areas.

In the context of the second hypothesis, school-characteristic differences by mother tongue, school-level variables from the school dataset include teachers' training, teachers' performance, the presence of a school library, the length of classes (in minutes), and whether the school is certified as having a bilingual program. From school-level variables, three main differences emerge. First, schools located in rural areas have less library infrastructure: books, desks and a studying area. For example, 7.5 percent of schools in rural areas have libraries with all this infrastructure, while this number more than doubles for schools in urban areas. Second, only 32% of schools in rural areas use previous national tests to help students study, while more that 55% of schools in urban areas employ this method. Lastly, rural schools devote less time to preparing students for the national test (see last 8 variables).

In terms of the last hypothesis, whether differences in Spanish comprehension affect learning at school, non-SMT students' lack of Spanish comprehension could originate from the lack of interaction with people proficient in Spanish, for example. The peer effects literature shows that peer effects are important not only in the determination of friendships, which are especially strong intra-race, (Fruehwirth, 2013; Hanushek et al., 2009; Marmaros and Sacerdote, 2006; Mayer and Puller, 2008), but also can affect students' achievement due to the race composition in the class (Hanushek et al., 2009). Therefore, peer effects may affect the Spanish learning process of non-SMT students, suggestive evidence of which is shown in Figure 3.1.

The top graph in Figure 3.1 shows the relationship between the school's proportion of SMT students in time $t-1$ and the school's proportion of non-SMT students who reported both mother tongue use and not being proficient in Spanish in time t. This figure clearly reveals that if the school's proportion of SMT students in time $t-1$ rises, non-SMT students are more likely to report that they do not speak their mother tongue. Similarly, a school's proportion of SMT students also appears to influence non-SMT students' Spanish comprehension, but this influence is not as strong as the case of non-SMT students' mother tongue use. These negative relationships suggest that non-SMT students are better at learning in the language of instruction at school if they have the opportunity to attend a school highly populated by Spanish native speakers. Furthermore, the bottom graph in Figure 3.1 reveals a positive relationship between the proportion of SMT students at school and non-SMT students' average achievement on the national test. This positive relationship suggests that Spanish native speakers might influence non-SMT students' Spanish comprehension, which could result in higher test scores for non-SMT students.

With this in mind, community characteristics such as the degree of isolation can determine intra-race vs. inter-race peer effects. Table 3.5 shows the non-SMT students' average mother tongue use, for both junior high and high school, by municipalities' SMT student share. This table indicates that the higher the municipalities' share of SMT students, the lower the proportion of non-SMT students' mother tongue use. For example, 32% of municipalities have less than 50% of SMT students at school, and in these municipalities more than 52% of non-SMT students reported using their non-Spanish mother tongue.

Non-SMT students' low human capital may persist indefinitely without a change in community demography or governmental support. The existence of a relationship between either a community characteristic or the linguistic composition at school in time $t-1$, $\sigma_{s,t-1}^{SMT}$, and non-SMT students' Spanish comprehension may hinder students' learning at school. I refer to this relationship as a language barrier for non-SMT students.

3.4 Empirical approach

The goal of this chapter is to look for evidence on the potential existence of a language barrier at secondary school. To do so, I hypothesize that students' lack of Spanish comprehension negatively affects learning at school.

The empirical approach relies on estimating value-added production functions for math and reading. I follow two different estimation approaches to investigate the sensitivity of the estimated impacts of not being proficient in Spanish to other confounding factors driving student achievement. The baseline specification is based on the Chapter 2 findings, but modified to incorporate both the data limitations in this chapter and the value-added production function specification. The second approach takes as reference the baseline specification, but controls for student fixed effects.

3.4.1 The baseline specification

Based on the results in Chapter 2, I assume that the non-SMT student i, $nsmt_i$, who attends school s_i can learn a subject, $y_{i,t}$, in three ways. First, the student can better learn a subject, the better their Spanish comprehension. To control for the student's understanding of Spanish, I employ the non-SMT student's self-evaluation of not being proficient in Spanish, $q_{i,t}nsmt_i$. This variable, $q_{i,t}nsmt_i$, takes a value of one only if non-SMT students reported not being proficient in Spanish and zero otherwise. Second, the student can ask questions to their teachers employing their mother tongue to learn a subject. To control for this student-teacher interaction, I employ the non-SMT student's mother tongue use, $m_{i,t}nsmt_i$. This variable, $m_{i,t}nsmt_i$, takes a value of one if non-SMT students reported intensive use of their

mother tongue and zero otherwise. Third, the student i can exert effort to learn, $e_{i,t}$. I also assume that the student's innate ability, θ_i, influences how the student learns. Therefore, the value-added production function is defined as:

$$y_{i,t} = \alpha_0 + \alpha_1 y_{i,t-1} + \alpha_2 nsmt_i + \alpha_3 q_{i,t} nsmt_i + \alpha_4 m_{i,t} nsmt_i + \alpha_5 e_{i,t} + \alpha_6 \theta_i + \epsilon_{i,t}, \quad (3.1)$$

where $\epsilon_{i,t}$ stands for an error term.[9]

A drawback in this chapter is that students were not asked for their level of effort when attending schools. To partially control for students' effort, I include controls that influence students' effort in the value-added production function like the ones shown in Chapter 2. I assume that the student's effort, $e_{i,t}$, is defined as:

$$e_{i,t} = X_{i,t}\rho + \varepsilon_{i,t}, \quad (3.2)$$

where $X_{i,t}$ is a row vector that controls for parent, student, regional, and schools characteristics. I include in $X_{i,t}$ parents' educational attainment, family assets and remittances (to control for family income).[10]

As mentioned in Section 3.2, community characteristics, the bilingual program at school and other school characteristics may play a role in explaining test score gaps by ethnicity. I include municipal and rural area fixed effects in $X_{i,t}$ to control for unobserved community characteristics that can influence students' performance at school. For instance, municipal and rural area fixed effects may control for the population's linguistic distribution across Guatemala. I also include year fixed effects. The variable $X_{i,t}$ also includes all school attributes displayed in Table 3.4, which includes a dummy variable for whether student i is taught following the bilingual educational program at school s, bp_{s_i}. Furthermore, I also include in $X_{i,t}$ the linguistic composition at school s that student i attends, $\sigma^{SMT}_{s_i,t}$, to control for linguistic peer effects (Hanushek et al., 2009). I substitute equation 3.2 into equation 3.1 and simplify terms to get:

$$y_{i,t} = \alpha_0 + \alpha_1 y_{i,t-1} + \alpha_2 nsmt_i + \alpha_3 q_{i,t} nsmt_i + \alpha_4 m_{i,t} nsmt_i + X_{i,t}\rho\alpha_5 + \alpha_6 \theta_i + \alpha_5 \varepsilon_{i,t} + \epsilon_{i,t},$$

[9] By including previous test scores in the production function, I assume that all previous student, family and school inputs do not affect test scores (Todd and Wolpin, 2003).

[10] Chapter 2 also shows that after controlling for student effort which is influenced by parental attributes, parental attributes does not play an important role explaining test scores through other students' unobserved traits.

or in terms of the reduced form specification:

$$y_{i,t} = \beta_0 + \beta_1 y_{i,t-1} + \beta_2 nsmt_i + \beta_3 q_{i,t} nsmt_i + \beta_4 m_{i,t} nsmt_i + X_{i,t}\beta_5 + \xi_{i,t}, \quad (3.3)$$

$$\xi_{i,t} = \beta_6 \theta_i + \beta_7 \varepsilon_{i,t} + \epsilon_{i,t}, \quad (3.4)$$

where $\beta_i = \alpha_i$ for $i \in \{0, 1, 2, 3, 4\}$, and the error term $\xi_{i,t}$ consists of the students' innate ability, θ_i, the unobserved part of students' effort, $\varepsilon_{i,t}$, and an individual and time specific term, $\epsilon_{i,t}$.

Under the baseline equation 3.3, a non-negative value for β_3 indicates that non-SMT students who self-evaluate as not being proficient in Spanish are learning, through their listening skills, as much as their counterparts. However, a negative β_3 implies that non-SMT students are performing worse than their counterparts at school due to their lack of Spanish comprehension. Furthermore, the magnitude of β_3 also provides a direct measure of the effect of students' lack of Spanish comprehension or students' language barrier on their test scores. The rejection of the null hypothesis

$$H_o : \beta_3 \geqslant 0 \quad (3.5)$$

would thus support the findings in Chapter 2 that show that elementary students are not yet proficient in Spanish.

Under the assumption that the expected value of the error term in equation 3.3 conditional on observables is zero, running OLS on this baseline specification provides unbiased estimates. However, the identification of causal effects of not being proficient in Spanish on test scores in equation 3.3 is complicated due to various potential problems that can bias estimates: students' attrition in grade progression, students' innate ability, and variables' endogeneity such as Spanish comprehension.

3.4.2 Students' attrition in grade progression

The first problem is selection due to students' attrition in grade progression, which can bias estimates such as $\widehat{\beta}_3$. The selection in school progression requires some additional

explanation of the Guatemalan education system. Mineduc establishes the curriculum of all subjects, and schools are required to cover this curriculum. Furthermore, Mineduc requires that students must be promoted to the next grade if their marks are above 60 percent in all classes they take each academic year.[11] How students are tested and the frequency of subject tests, however, are determined by principals, which may create heterogeneity of school thresholds for progression. Top schools can design more challenging tests than their counterparts. Therefore, students are observed in period t only if they were above the school-specific threshold in period $t-1$.

In order to control for selection due to students' attrition, the probability of the students' grade progression is included as a control function when estimating the production functions.[12] The students' grade progression probability is estimated by a logit model and included as a polynomial of second order in the production function. The right hand variables in the logit specification for student grade progression represent the difference between the student's score on the national test for both math and reading and the average score at the student's school at time $t-1$. This model specification is based on Table 3.6, which shows national test scores by grade progression. The main message of this table is that those students who progressed to the next grade had scores on the national test, for both math and reading, that were above the national mean, as well as above the average for their schools.[13] I assume that a school's threshold is given by or highly correlated with the students' average scores (see appendix for the results). I include the probability of the students' grade progression in $X_{i,t}$.

3.4.3 Students' innate ability

The second problem in finding evidence of a language barrier at secondary school under equation 3.3 is students' innate ability because this unobserved variable can bias estimates such as $\widehat{\beta}_3$. There are two approaches to deal with students' innate ability.

First, the standard approach to deal with students' innate ability is to difference equation 3.3, which eliminate any fixed effects at student, school or regional level from the production function. By taking a first difference to equation 3.3, the value added model is given now by:

$$\Delta y_{i,t} = \beta_1 \Delta y_{i,t-1} + \beta_3 \Delta q_{i,t} nsmt_i + \beta_4 \Delta m_{i,t} nsmt_i + \Delta X_{i,t}\beta_5 + \Delta \xi_{i,t}, \tag{3.6}$$

[11] The national tests do not affect students' school progression.

[12] I include the probability of the students' grade progression in all model specifications. See James Heckman and Salvador Navarro-Lozano (2004) for details.

[13] The schools' average scores were calculated using all students from the dataset. Students who I can follow over grades were included.

$$\Delta\xi_{i,t} = \beta_7\Delta\varepsilon_{i,t} + \Delta\epsilon_{i,t}. \tag{3.7}$$

To find evidence of the existence of a language barrier, I build a longitudinal dataset that consists of at most 3 observations per student: last grade in elementary, junior high and high school. However, by taking the first difference, the sample size will be determined by those students who attended high school, since one of these observations is lost by modeling the production function as an accumulative process. Therefore, if a language barrier exists at school, it is likely that students affected the most may have not reached high school; consequently, the estimate of $\widehat{\beta}_3$ given by equation 3.6 may be less negative or equal than the estimate of $\widehat{\beta}_3$ under equation 3.3 if in this last equation I am able to control for student's ability and to include junior and high school students in the estimation.

The second and less common approach is to control for students' innate ability in equation 3.3. Adding the superscript r to parameters, error term and test scores in equation 3.3 to represent the reading production function, I can solve for students' innate ability from the reading production function. Similar, I add the superscript m to parameters, error term and test scores in equation 3.3 to represent the math production function. With the students' innate ability at hand, I can substitute it into the math production function to get:

$$y_{i,t} = \pi_0 + \pi_1 y_{i,t-1} + \pi_2 nsmt_i + \pi_3 q_{i,t} nsmt_i + \pi_4 m_{i,t} nsmt_i + X_{i,t}\pi_5 + \pi_6 y^r_{i,t} + \pi_7 y^r_{i,t-1} + \psi_{i,t} \tag{3.8}$$

where $\psi_{i,t} = \pi_8\varepsilon_{i,t} + \pi_9\epsilon^r_{i,t} + \epsilon^m_{i,t}$, and $\pi_j = \beta^m_j - \beta^r_j\frac{\beta^m_6}{\beta^r_6}$ for $j = 0, 2, 3, 4, 5, 8$.[14]. The new error term $\psi_{i,t}$ includes the unobserved part of student effort, and the error terms of reading and math production functions, respectively. Therefore, under equation 3.8, to find evidence of a language barrier at school, $\pi_3 = \beta^m_3$, it is necessary that Spanish comprehension does not affect reading performance at school, $\beta^r_3 = 0$.[15] I show that this is the case for secondary students in Section 3.5.

Given the fact that using reading test scores to control for students' ability in equation 3.8 is not commonly used, as a robustness check, I difference equation 3.8 to get:

$$\Delta y^m_{i,t} = \pi_1\Delta y^m_{i,t-1} + \pi_3\Delta q_{i,t} nsmt_i + \pi_4\Delta m_{i,t} nsmt_i + \Delta X_{i,t}\pi_5 + \pi_6\Delta y^r_{i,t} + \pi_7\Delta y^r_{i,t-1} + \Delta\psi_{i,t} \tag{3.9}$$

[14]The remaining parameters are given by $\pi_1 = \beta_1$, $\pi_6 = \frac{\beta^m_6}{\beta^r_6}$, $\pi_7 = -\beta^r_1\frac{\beta^m_6}{\beta^r_6}$, $\pi_7 = \frac{\beta^m_6}{\beta^r_6}$, and $\pi_9 = -\frac{\beta^m_5}{\beta^r_5}$.

[15]I show that Spanish comprehension and mother tongue use are statistically equal to zero when explaining reading test scores. Then, in section 3.5.2 I use the notation $\pi_3 = \beta_3$ for Spanish comprehension and $\pi_4 = \beta_4$ for mother tongue use for equation 3.8.

$$\Delta\xi^{m}_{i,t} \quad = \quad \pi_8 \Delta\varepsilon_{i,t} + \pi_9 \Delta\epsilon^{r}_{i,t} + \Delta\epsilon^{m}_{i,t}. \tag{3.10}$$

I compare the estimates from equation 3.9 with the estimates from equation 3.6 to see whether the correlation of reading test scores with the error term of the reading production function, $\epsilon^{r}_{i,t}$, bias estimates in equation 3.9. I show that this is not the case in Section 3.5. Therefore, I am reasonable certain that estimates from equation 3.8 will not be biased by the correlation of reading test scores and the reading error term, $\epsilon^{r}_{i,t}$.

3.4.4 Endogeneity

The third problem is that non-SMT students' lack of Spanish comprehension may be endogenous if it is correlated with the unobserved part of students' effort in any model specification, the baseline or first differenced, which I would expect to bias estimates such as $\widehat{\beta}_3$.

To understand why the unobserved part of students' effort may lead to an overestimate of the β_3, remember that the Guatemalan education system requires that students score 60 percent in all subjects to progress to the next grade, and top ranking schools likely demand more effort from students to score 60 percent. As a consequence, if a student faces some probability of not progressing to the next grade, the student's effort or study time management may depend on both their traits as well as on the quality of the school they attend. Given the probability of not passing to the next grade, if non-SMT students are not learning due to their lack of Spanish comprehension, it is likely that they exert more effort by devoting more time to study by themselves rather than spending time learning Spanish. When this is the case, the correlation between students' lack of Spanish comprehension, $q_{i,t}$, and the unobserved part of students' effort, $\varepsilon_{i,t}$, may be equal to or larger than zero (see Chapter 2). This means that the $\widehat{\beta}_3$ would be overestimated, which works against me finding a language barrier (i.e., a negative $\widehat{\beta}_3$).

Non-SMT students' mother tongue use may be endogenous if it is also correlated with the unobserved part of students' effort in the production function, which I would also expect to bias the $\widehat{\beta}_4$ estimate upward. Similar to the second problem, non-SMT students may offset their Spanish comprehension deficit by asking for help using their mother tongue rather than spending time learning Spanish.

The students' sorting across schools, $bp_{s_i,t}$ or $\sigma^{SMT}_{s_i,t}$ can also bias estimates if parents' enrollment decision are based on school quality, which influence students' effort. The Guatemalan literature about the impact of the bilingual program on test scores shows positive effects, at least for the first three elementary grades. Therefore, any unobserved characteristic of this program that parents consider when selecting a school can bias the results.

The identification of a causal effect of a language barrier at school on test scores is clearly difficult not only due to the endogeneity of students' lack of Spanish comprehension with the unobserved part of students' effort, $\epsilon_{i,t}$, but also due to the endogeneity of other variables resulting from an equilibrium process that affects students' performance at school.

To reliably separate the independent impact of students' lack of Spanish comprehension from other factors driving student achievement, I employ an instrumental variable approach. The variables I treat as endogenous are students' self-report of being non-SMT speakers, $nsmt_i$, non-SMT students' self-evaluation of not being proficient in Spanish, $q_{i,t}nsmt_i$, and mother tongue use, $m_{i,t}nsmt_i$, students' attendance at a school with the bilingual program, $bp_{s_i,t}$, and SMT students' sorting across schools, $\sigma^{SMT}_{s_i,t}$. In what follows, I discuss instruments by assuming what instruments influence each endogenous variable. Of course, in a joint 2SLS procedure all instruments may influence all endogenous variables.

I discuss the instruments for students' report of being non-SMT speakers, $nsmt_i$. The instruments are whether the mother, father or both parents have a non-Spanish language as a mother tongue. I assume that parents' mother tongue does not have a direct effect on students' test scores, once I have controlled for parents' educational attainment and family assets.[16] For instance, the findings in Chapter 2 indicate that parents' mother tongue only influences students' Spanish comprehension. Therefore, parents' educational attainment and family assets may non-parametrically control for any omitted variable that parents may consider to improve their child's performance at school. I also control for municipal and rural area fixed effects which may control for the Guatemalan linguistic distribution.

Regarding the instruments for non-SMT students' self-evaluation of not being proficient in Spanish, $q_{i,t}nsmt_i$, I employ as an instrument the linguistic composition at school s where student i attended in time $t-1$, $\sigma^{SMT}_{s_i,t-1}$.[17] Non-SMT students' Spanish comprehension may be affected by schoolmates. The higher the number of SMT students at school, the higher non-SMT students' Spanish proficiency might be. In the Guatemalan context, then, students' linguistic sorting across schools in time $t-1$, $\sigma^{SMT}_{s_i,t-1}$, might not have a direct impact on students' performance in time t, but still may be correlated with the non-SMT students' Spanish proficiency in time t. The variable students' linguistic sorting across schools, $\sigma^{SMT}_{s_i,\cdot}$, is constructed to represent deviations from a uniform sorting of SMT students across schools. So, a positive number stands for an excess of SMT students in a school relative to the uniform sorting of SMT students across schools.

Now, I discuss the instrument for whether non-SMT students intensively speak their mother tongue. The instrument is the proximity from the school where the student attended

[16]The family asset variable can be a measure of the parents' education because education determines income and thus asset ownership.

[17]A difference of one period in the model in this chapter means a difference of three academic years in the data. Todd and Wolpin (2003) suggest that student inputs can be used as instrumental variables if these inputs were applied at a time sufficiently prior to the earliest observation used and the students' innate ability is controlled for.

to the nearest main municipal city in the periods t and $t-1$ (school-main municipal city proximity). In the context of a language learning process, the more students can interact with Spanish native speakers (not related to schoolmates), the faster they become proficient in Spanish. This is the effect that school-main municipal city proximity is meant to capture. However, this also can be correlated with school quality. For example, the further the school is from a main or developed city, the less likely that well trained teachers will work at these schools. I assume that school-main municipal city proximity only influence students' mother tongue use after controlling for municipal and rural area fixed effects, and also after controlling for the extensive school attributes contained in the dataset such as teachers' training, schools' infrastructure, and principals' educational attainment.

The instrument for students sorting across schools with the bilingual program is the students' distance from home to the nearest school with the bilingual program.[18] This distance may represent students' accessibility to attend a school with a bilingual education program. Even if student i did not attend a school with the bilingual program, I can identify the nearest school with such a program. I also assume that after controlling for the extensive school attributes contained in the dataset, the instrument students' distance from home to the nearest school with the bilingual program does not directly influence students' performance at school.

The last set of the instruments may influence students' linguistic sorting across schools in time t, $\sigma^{SMT}_{s_i,t}$. I employ two groups of instruments. The first group is the students' proximity from home to the school where students attended. I still assume that this instrument does not directly influence students' performance after controlling for fixed effects and school attributes. The second group of instruments is the SMT students' sorting that surrounds the school s where student i attended, $\sigma^{SMT}_{-s_i,t}$. I follow the assumption behind the industrial organization literature. The linguistic sorting in school s, $\sigma^{SMT}_{s_i,t}$, may be correlated with the linguistic sorting of other schools due to common marginal costs, but $\sigma^{SMT}_{-s_i,t}$ will not have a direct effect on students' performance at school after controlling for schools' attributes, parents' educational attainment, family assets, students' characteristics and regional fixed effects. The variable $\sigma^{SMT}_{-s_i,t}$ represents a weighted average of the student linguistic sorting across schools without the school s_i that student i attended.[19]

I show results for equations 3.6, 3.8 and 3.9 for math production functions by assuming that the expected value of the error term is zero (OLS), but also when assuming that the unobserved part of students' effort is correlated with students' lack of Spanish comprehension for instance.

To summarize Section 3.4, the main objective of this chapter is to look for evidence of the existence of a language barrier at secondary school, by separating the independent impact

[18]The dataset does not indicate where students reside, but I assume that they live near the school they attend at grade six.

[19]I employ as weights the distance from school s to school $-s$ at municipal level.

of students' lack of Spanish comprehension from other factors driving student achievement such as effort or innate ability. I employ two strategies to investigate the sensitivity of the estimated impacts to potential confounding factors. First, I rely on a fixed effect strategy to control for students' innate ability. In this first strategy, the sample size does not include students who did not reach high school. In a second strategy, I employ reading test scores to control for students' innate ability in the math production function in the baseline specification, allowing me to include students who did not attend high school. I expect to find a higher language barrier effect on test scores following the second strategy than under the first strategy.

3.5 Results

This section shows the effect of not being proficient in Spanish when attending secondary school. First, I discuss the results from the math and reading production functions when controlling for student fixed effects, equation 3.6. Particularly, I show that Spanish comprehension does not play a key role when explaining students' performance on national reading tests. Second, by using reading test scores to control for students' innate ability on the math production function, see equation 3.9, I confirm that Spanish comprehension is a key factor for developing only math skills. Last, I present evidence of the damage of not being proficient in Spanish when attending secondary school by estimating equation 3.8 for math, where I control for students' innate ability by using reading test scores and I employ the full sample.[20]

The results indicate that the linguistic composition at school influences both non-SMT students' lack of Spanish comprehension and mother tongue use. Therefore, non-SMT students with a low level of Spanish comprehension perform poorly on national tests. All model specifications consistently indicate a robust negative and significant relationship between non-SMT students' lack of Spanish comprehension and their performance at school. Non-SMT students who are not yet proficient in Spanish may be outperformed at least by 0.523 standard deviations by their counterparts in terms of math scores. Students' lack of Spanish comprehension can entirely explain test score gaps by mother tongue (see Table 3.1 for students' score gaps).

[20]Both foreign students and students who work while at school are excluded in this chapter.

3.5.1 Results from the math and reading production functions when controlling for student fixed effects

Recall from Section 3.4.1 that the variables $nsmt_i$, $q_{i,t}nsmt_i$ and $m_{i,t}nsmt_i$ stand for non-Spanish mother tongue speaker, not being proficient in Spanish and intensive use of a non-Spanish language, respectively. These variables take a value of one if students reported to belong in one of the categories and zero otherwise.

All tables of results when controlling for students' fixed effects (Tables 3.7-3.9) share the following structure. In the first column I show results when assuming that the expected value of the error term, $\epsilon_{i,t}$, conditional on observables, is zero. From the second to the last columns, I relax the previous assumption and employ two stage least squares to deal with the variables I consider as endogenous. In the second column, I treat as endogenous non-SMT students' self-evaluation of not being proficient in Spanish, $q_{i,t}nsmt_i$. The third column adds as endogenous variables both the sorting of students into schools with the bilingual program, $bp_{s_i,t}$, and the students' linguistic sorting across schools, $\sigma^{SMT}_{s_i,t}$. The last column displays results when also treating the variable intensively speaking a non-Spanish language as endogenous, $m_{i,t}nsmt_i$.

I discuss results from equation 3.6, which represents the first differenced baseline specification. Tables 3.7 and 3.8 display results for math and reading, respectively. I reserve the discussion of 2SLS first stages until analyzing results with the full sample.

The math test score results from Table 3.7 indicate that non-SMT students who report as not being proficient in Spanish perform worse than their counterparts in math. When the model is estimated by OLS, the first column, results indicate that the language barrier effect on math tests is significant and close to -0.061 standard deviations. The second, third and fourth columns estimate the equation by 2SLS. Results from these columns indicate that the language barrier effect is still significant, and the effect is not negligible for non-SMT students who self-evaluated as not being proficient in Spanish. These students are going to be, at least, outperformed by 0.346 standard deviations by their counterparts, which accounts for 91 percent of the observed math test score gap. The language barrier effect of 0.346 standard deviations can be understood as an upper bound since (1) when treating other variables as endogenous only increases the gap, and (2) only students who reach high school are included in the estimation. The finding that a language barrier still exists at secondary school put at risk the human capital development of people who do not have Spanish as a mother tongue, since the Guatemalan educational system requires students to score above 60 percent to progress into higher grades.

In terms of reading test scores, results indicate that Spanish comprehension plays no role in acquiring reading skills. Table 3.8 displays this finding. Regardless of the model specification, this table shows that a non-SMT student's reading performance, at most, is affected by 0.161 standard deviations by a language barrier, but its effect is not statistically

significant. A possible interpretation for this finding is that scientific subjects such as math require more in-class time to learn and develop practical skills.

Given that Spanish comprehension does not influence reading skills, reading test scores can be used to pin-down students' innate ability. Equation 3.9 displays the value-added production function for math when controlling for students' innate ability. If the estimates of the language barrier effect on math test scores under equations 3.6 and 3.9 do not differ, I can argue both that Spanish proficiency does not develop reading skills, and that the correlation between reading test scores and its error term in the math production function does not heavily influence the language barrier effect on math.

Table 3.9 displays the results for the math production functions when controlling for students' innate ability. The finding that students who are not proficient is Spanish perform worse than their counterparts still holds. The language barrier effect on math test scores does not heavily differ when estimating equation 3.6 or 3.9. Therefore, I am reasonably certain that the estimation of equation 3.8 provides an unbiased estimate of the language barrier damage when both controlling for students' innate ability and including all students in the estimation, which I discuss next.

3.5.2 Results for the math baseline specification when controlling for students' innate ability

I now turn to the results for the math production function when both controlling for students' innate ability and including all students in the estimation. Table 3.10 shows the main results of the damage of not being proficient while at school.

In terms of this table's structure, first, the table is divided into three subtables. The upper table displays results for OLS and 2SLS second stages when estimating equation 3.8. The middle and bottom tables show the 2SLS first stages for not being proficient in Spanish and non-Spanish language use, respectively. I start discussing the first stage results, then I discuss the second stage results. Second, in terms of the table's column structure, the first two columns show results when estimating equation 3.8 by OLS, when I do not and do control for language barrier effects, respectively. Results in these columns can be comparable to the current Guatemalan literature that does not control for a language barrier at school. From the third to the sixth column, I display results when estimating equation 3.8 by 2SLS, when treating as endogenous the students' report of being non-SMT speakers, $nsmt_i$, students' self-evaluation of their lack of Spanish comprehension, $q_{i,t}nsmt_i$, the students' sorting across schools, $bp_{s_i,t}$ *and* $\sigma^{SMT}_{s_i,t}$, and students' mother tongue use, $m_{i,t}nsmt_i$, respectively.

Key results from the first stage estimation are reported at the middle and bottom of Table 3.10. First, students' linguistic sorting across schools in time $t-1$, $\sigma^{SMT}_{s_i,t-1}$, is an influential

factor for reducing non-SMT students' lack of Spanish comprehension in time t. A lag of one period in the model represents a difference of 3 years in the data. Second, results also show that linguistic segregation at school is robust across model specifications, even when controlling for municipal and rural area fixed effects. These fixed effects should account for the Guatemalan linguistic distribution. This negative relationship between students' linguistic sorting across schools and students' lack of Spanish comprehension suggests that without either students' linguistic relocation across schools or governmental support, the low human capital of non-SMT students can persist indefinitely.

Similar to Spanish comprehension, students' linguistic sorting across schools is also an influential factor for non-SMT students' mother tongue use, see the bottom part of the table. Non-SMT students report not intensively speaking their mother tongue if they attend a school highly populated by Spanish native speakers. An interesting finding is that non-SMT parents do not influence their non-SMT child's mother tongue use, suggesting that non-SMT parents may care about the Spanish learning process of their children. However, non-SMT students still report intensive use of their mother tongue.

Preliminary evidence of the root of the lack of Spanish comprehension is also shown at the bottom part of Table 3.10. The average school linguistic composition, $\sigma^{SMT}_{-s_i,t}$ that surrounds the school s_i where the non-SMT student i attends positively influences their mother tongue use at time t. This positive relationship suggests that there exists some sort of segregation at school by mother tongue, since both municipal and rural area fixed effects and school proximity to main municipal cities should account for the Guatemalan linguistic distribution.

To conclude the first stage analysis, the international literature on peer effects indicates that friendships are strong intra-race, and such friendships may be shaped by parents' school selection for their child. Intra-race friendships paired with the fact that speaking one's mother tongue hinders the learning process of a new language may be creating the language barrier at school in Guatemala. The finding in this chapter that students' linguistic sorting across schools affects Spanish comprehension seems to fit the above-mentioned fact supported by the international literature.

I now discuss the results for the math production function for OLS and 2SLS second stages (upper part of Table 3.10). The first two columns show results when estimating the model by OLS. These two model specifications indicate that non-SMT speakers, $nsmt_i$, are not statistically different than their counterparts in terms of math test scores. Furthermore, results also show that the effect of Spanish comprehension, $\widehat{\beta}_3$, on math test scores is not negative. However, non-SMT students who intensively speak their mother tongue would score 0.05 standard deviations lower grades than their counterparts.

The third column displays the result when treating students' self-report of being non-SMT speakers as endogenous. In terms of the language barrier effect on math test scores, $\widehat{\beta}_3$,

students who are not yet proficient in Spanish score 0.11 standard deviations lower grades than their counterparts, which accounts for 28.9 percent of the observed math gap. Speaking intensively a non-Spanish language still negatively affects non-SMT students' performance at school, but with a larger effect of 0.206 standard deviations. However, students who are non-SMT speakers perform better that their counterparts by 0.282 standard deviations.

In the fourth column I deal with the endogeneity of not being proficient in Spanish. If non-SMT students who are not proficient in Spanish exert more effort to offset their poor performance at school, the language barrier effect on math may be overestimated (less negative). The language barrier effect on students' math performance is negative, significant and accounts for 0.523 standard deviations, which represents 137 percent of the observed gap. Non-SMT students' mother tongue use still affects their learning of math. For the remaining columns, I just highlight key differences relative the this column.

The fifth column displays results when dealing with the endogeneity of students' sorting into the bilingual program and linguistic sorting across schools, bp_{s_i} and $\sigma^{SMT}_{s_i,t}$, respectively. Relative to the estimate of $\widehat{\beta}_3$ from previous column, the language barrier effect at school is negative and significant, but with a larger coefficient magnitude: 1.52. The overestimation in the OLS case suggests that there exists an agglomeration effect of students who are not yet proficient in Spanish at school. This agglomeration effect is also shown at the 2SLS first stages where the students' linguistic sorting influences students' Spanish learning process.

The last column shows results when also treating non-SMT students' mother tongue use as endogenous. Still, the fact that non-SMT students are not yet proficient in Spanish significantly affects their math performance at school. The language barrier effect on math still accounts for 1.084 standard deviations. However, speaking a non-Spanish language has a stronger effect on non-SMT students' math performance than the fourth column estimate: 0.725 standard deviations.

To summarize this section, I consider how not being proficient in Spanish is related to non-SMT student achievement after controlling for differences in family background, school attributes, and municipal and rural area fixed effects. While identification of causal effects is not an easy task due to the endogeneity of students' decisions to learn and students' innate ability, the consistent estimated impacts across alternative estimation approaches supports the finding that non-SMT students are performing poorly at secondary school due to their lack of Spanish comprehension.

3.6 A policy scenario: improving non-SMT students' Spanish comprehension

The estimation of the baseline and the fixed-effect models provide evidence that non-SMT students perform poorly at school due to their lack of Spanish comprehension, which is influenced by students' linguistic sorting across schools. Therefore, improving non-SMT students' Spanish listening skills at school would be important for alleviating non-SMT students' poor performance. The policy I evaluate in this section can be understood as a second language program. I artificially change the students' linguistic sorting across schools to simulate an increase of non-SMT students' interaction with native Spanish speakers, which can result in non-SMT students' Spanish comprehension improvement. Specifically, this policy simulation look for evidence of what could have happened if SMT students would have attended schools in rural areas, where the majority of non-SMT students attend.

I follow Heckman and Vytlacil (2005) to look for evidence of such a policy.[21] It is worth mentioning that by estimating the treatment effect of a second language program using as an instrument the students' linguistic sorting across schools in the previous period, $\sigma_{s_i,t-1}^{SMT}$, the estimation procedure simulates only the non-SMT students' Spanish learning process and not any SMT students' test score peer effect. The SMT students' simulated sorting across rural schools is based on the following two rules. First, ten percent of SMT students were taken from urban schools only if after taking these students, the school's share of SMT students is equal to or higher than 66 percent. This last percentage represents a uniform sorting of non-SMT students across schools. Second, these selected SMT students were uniformly distributed across rural schools only if these rural schools were located within a 15 kms. range of where the selected students were taken from.

Although the main objective of this section is to provide evidence of the effect of Spanish as a second language program on non-SMT students' math performance, it is worth mentioning that the estimation approach for treatment effect of a second language program has as a sub-product the estimation of the marginal treatment effect (MTE).[22] Having an estimate

[21] The policy relevant treatment effect (PRTE) follows the work of Heckman and Vytlacil (2005), Carneiro and Lee (2009), Carneiro et al. (2010) and Carneiro et al. (2010). The PRTE measures the test score's gain or loss given a change in the linguistic sorting at schools. In this chapter, the $PRTE$ is defined as:

$$PRTE = E\left(y_{i,Z_1}\right) - E\left(y_{i,Z_0}\right) = \int_x \int_u MTE\left(x,u\right) \cdot \left(F_{P^{z_0}|X}\left(u\right) - F_{P^{z_1}|X}\left(u\right)\right) du \cdot f(x) \cdot dx \quad (3.11)$$

where $Z_{i=\{1,0\}}$ is the students' linguistic sorting across schools under an alternative policy, Z_1, and the baseline policy, Z_0. The marginal treatment affect (MTE), the family and school inputs distribution, $f(x)$, and the distributions $F_{P^{z_1}|X}\left(u\right)$ and $F_{P^{z_0}|X}\left(u\right)$ are non parametrically estimated following Carneiro and Lee (2009).

[22] In the estimation of MTE, I do not control for those variables that I treat as endogenous: students' sorting across schools and their mother tongue use.

for the marginal treatment effect allows me to estimate the average language barrier effect following Basu et al. (2007), which controls for the students' essential heterogeneity and self-selection to not learn Spanish, which can bias the language barrier effect on test scores. Therefore, relaxing the OLS and IV assumption that the error term in the educational production function does not differ by potential gains or losses of being treated. I, first, briefly discuss this result. Then, I show the results for the treatment effect of a second language program.

The estimation of average treatment effect provides two results: the average and marginal language barrier effect on test scores for a non-SMT student who is indifferent between learning Spanish.[23] First, results for the average language barrier effect on test scores are shown in Table 3.11. Similar to Section 3.5, results indicate that non-SMT students' lack of Spanish comprehension affects their performance on national math tests. If students report as not being proficient in Spanish, their math scores would be lower by 1.428 standard deviations than their counterparts. Second, Figure 3.2 depicts the marginal language barrier effect on non-SMT students' test scores.[24] The marginal language barrier effect displays a decreasing trend. The higher the probability of non-SMT students' interaction with SMT speakers, the less likely that non-SMT students do not understand Spanish. As a result, the math test score increases. Non-SMT students who face a low probability of improving their Spanish comprehension may be outperformed almost by 2 standard deviations by their counterparts, while the those who can improve their Spanish proficiency will be around -0.5 standard deviations.

The results for the policy treatment effect on non-SMT students' test scores are presented in Table 3.12. This table shows the effect of a second language program by rural and urban area, given that SMT students were moved from urban to rural schools. The first row displays the policy effect for non-SMT students who reside in rural areas. The policy evaluation indicates that by improving students' Spanish comprehension, non-SMT students' achievement on math national tests increase, but the language barrier effect is still significant, -0.705 standard deviations. On the other hand, non-SMT students who attend schools in urban areas are not affected by the simulation of the policy (see second row). Although the total effect of the policy, third row, suggests that non-SMT students attending schools in rural areas better perform in math test scores after the simulation, further research is still needed to know how parents select schools in order to successfully implement this alternative policy.

[23]The support for local instrumental variable estimation is shown in the Figure A.1 in the appendix.

[24]Recall that $q_{i,t} = 1$ stands for whether non-SMT students' reported not being proficient in Spanish and zero otherwise. The propensity score gives the conditional probability that a non-SMT student is not proficient in Spanish, $q_{i,t} = 1$, when this student faces an exogenous shock, U. In this chapter, the higher the probability of an exogenous shock, U, the less likely non-SMT students will report not being proficient in Spanish.

3.7 Conclusion

This chapter investigates whether students still face an obstacle or barrier in acquiring skills while at secondary school in Guatemala, since 25 languages are spoken and Spanish is the language of instruction in most of schools. The achievement on the national tests for non-SMT students is lower than their counterparts, and without any significant improvement between elementary and high school. This achievement gap raises the question of whether, or to what extent, the existence of a language barrier in school is a contributing factor.

The empirical analysis relies on the estimation of education production functions that uses a value-added specification. Three estimation approaches are carried out for the math value-added production function. The first noteworthy finding is that non-SMT students' poor performance at school may be the result of their lack of Spanish proficiency. This finding is robust to alternative estimation approaches. The estimates indicate that non-SMT students' scores are at least 0.523 standard deviations lower than their counterpart scores which explains a significant portion of the math test gap.

The second finding is that even after controlling for both distance from the school to the main municipal city and for municipality fixed effects, students' linguistic sorting among schools can explain non-SMT students' lack of Spanish comprehension. This finding implies that parents' enrollment decisions other than the regional distribution of non-SMT population across Guatemala can be the root of non-SMT students' language barrier.

Last, the policy simulation, artificially mixing SMT and non-SMT students in the same schools in rural areas, provides evidence that non-SMT and SMT students' integration could help to alleviate the language barrier.

These three findings provide evidence that without government policy interventions, the low human capital of the Guatemalan non-SMT population may persist indefinitely. Identifying the root of the non-SMT students' language barrier may shed light for the needed policy intervention to eliminate the language barrier at school and to improve the non-SMT population's well-being.

Bibliography

Lewis R. Aiken. Verbal Factors and Mathematics Learning: A Review of Research. *Journal for Research in Mathematics Education*, 2(4):304–313, 1971. ISSN 00218251, 19452306. doi: 10.2307/748485. URL http://www.jstor.org/stable/748485.

Mathilde Almlund, Angela Lee Duckworth, James J. Heckman, and Tim D. Kautz. Personality Psychology and Economics. *National Bureau of Economic Research Working Paper Series*, No. 16822, 2011. doi: 10.3386/w16822. URL http://www.nber.org/papers/w16822.

Anirban Basu, James J. Heckman, Salvador Navarro-Lozano, and Sergio Urzua. Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. *Health Economics*, 16(11):1133–1157, 2007. ISSN 1099-1050. doi: 10.1002/hec.1291. URL http://dx.doi.org/10.1002/hec.1291.

Graham Beattie, Jean-William P. Laliberté, Catherine Michaud-Leclerc, and Philip Oreopoulos. What Sets College Thrivers and Divers Apart? A Contrast in Study Habits, Attitudes, and Mental Health. *National Bureau of Economic Research Working Paper Series*, No. 23588, 2017. doi: 10.3386/w23588. URL http://www.nber.org/papers/w23588.

Pedro Carneiro and Sokbae Lee. Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality. *Journal of Econometrics*, 149(2):191–208, April 2009. ISSN 0304-4076. doi: 10.1016/j.jeconom.2009.01.011. URL http://www.sciencedirect.com/science/article/pii/S0304407609000281.

Pedro Carneiro, James J. Heckman, and Edward Vytlacil. Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin. *Econometrica*, 78(1):377–394, 2010. ISSN 1468-0262. doi: 10.3982/ECTA7089. URL http://dx.doi.org/10.3982/ECTA7089.

Rubiana Chamarbagwala and Hilcías E. Morán. The human capital consequences of civil war: Evidence from Guatemala. *Journal of Development Economics*, 94(1):41–61, January 2011. ISSN 0304-3878. doi: 10.1016/j.jdeveco.2010.01.005. URL http://www.sciencedirect.com/science/article/pii/S0304387810000076.

Angela L. Duckworth and Martin E.P. Seligman. Self-Discipline Outdoes IQ in Predicting Academic Performance of Adolescents. *Psychological Science*, 16(12):939–944, December 2005. ISSN 0956-7976. doi: 10.1111/j.1467-9280.2005.01641.x. URL http://journals.sagepub.com/doi/abs/10.1111/j.1467-9280.2005.01641.x.

Kjell I. Enge and Ray Chesterfield. Bilingual education and student performance in Guatemala. *World Bank's Education Sector Review: Priorities and Strategies for Education*, 16(3):291–302, July 1996. ISSN 0738-0593. doi: 10.1016/0738-0593(95)00038-0. URL http://www.sciencedirect.com/science/article/pii/0738059395000380.

Jane Cooley Fruehwirth. Identifying peer achievement spillovers: Implications for desegregation and the achievement gap. *Quantitative Economics*, 4(1):85–124, 2013. ISSN 1759-7331. doi: 10.3982/QE93. URL http://dx.doi.org/10.3982/QE93.

Eric A. Hanushek, John F. Kain, and Steven G. Rivkin. New Evidence about Brown v. Board of Education: The Complex Effects of School Racial Composition on Achievement. *Journal of Labor Economics*, 27(3):349–383, 2009. URL https://ideas.repec.org/a/ucp/jlabec/v27y2009i3p349-383.html.

James Heckman, Jora Stixrud, and Sergio Urzua. The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. Technical report, National Bureau of Economic Research, Inc, February 2006. URL http://EconPapers.repec.org/RePEc:nbr:nberwo:12006.

James J. Heckman and Lakshmi K. Raut. Intergenerational long-term effects of preschool-structural estimates from a discrete dynamic programming model. *Journal of Econometrics*, 191(1):164–175, March 2016. ISSN 0304-4076. doi: 10.1016/j.jeconom.2015.10.001. URL http://www.sciencedirect.com/science/article/pii/S0304407615002493.

James J. Heckman and Edward Vytlacil. Structural Equations, Treatment Effects, and Econometric Policy Evaluation1. *Econometrica*, 73(3):669–738, 2005. ISSN 1468-0262. doi: 10.1111/j.1468-0262.2005.00594.x. URL http://dx.doi.org/10.1111/j.1468-0262.2005.00594.x.

Martha Hernandez-Zavala, Harry Patrinos, Christos Sakellariou, and Joseph Shapiro. Quality of schooling and quality of schools for indigenous students in Guatemala, Mexico, and Peru. Technical report, The World Bank, August 2006. URL http://EconPapers.repec.org/RePEc:wbk:wbrwps:3982.

James Heckman and Salvador Navarro-Lozano. Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models. *The Review of Economics and Statistics*, 86(1):30–57, 2004. URL https://ideas.repec.org/a/tpr/restat/v86y2004i1p30-57.html.

Shelly Lundberg and Richard Startz. On the Persistence of Racial Inequality. *Journal of Labor Economics*, 16(2):292–323, 1998. ISSN 0734306X, 15375307. doi: 10.1086/209890. URL http://www.jstor.org/stable/10.1086/209890.

David Marmaros and Bruce Sacerdote. How Do Friendships Form?*. *The Quarterly Journal of Economics*, 121(1):79–119, February 2006. ISSN 0033-5533. doi: 10.1093/qje/121.1.79. URL http://dx.doi.org/10.1093/qje/121.1.79.

Jeffery H. Marshall. School quality and learning gains in rural Guatemala. *Economics of Education Review*, 28(2):207–216, April 2009. ISSN 0272-7757. doi: 10.1016/j.econedurev.2007.10.009. URL http://www.sciencedirect.com/science/article/pii/S0272775708000745.

Adalbert Mayer and Steven L. Puller. The old boy (and girl) network: Social network formation on university campuses. *Journal of Public Economics*, 92(1):329–347, February 2008. ISSN 0047-2727. doi: 10.1016/j.jpubeco.2007.09.001. URL http://www.sciencedirect.com/science/article/pii/S0047272707001181.

Patrick J. McEwan and Marisol Trowbridge. The achievement of indigenous students in Guatemalan primary schools. *International Journal of Educational Development*, 27(1):61–76, January 2007a. ISSN 0738-0593. doi: 10.1016/j.ijedudev.2006.05.004. URL http://www.sciencedirect.com/science/article/pii/S0738059306000502.

Benjamin Meade. *Examining the structural roots of achievement disparities in Guatemalan primary schools*. PhD thesis, 2011. URL https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi?url=http://search.proquest.com/docview/851547825?accountid=15115.

F.E. Rubio. Educación Bilingüe en Guatemala: Situación y desafios. 2004.

Petra E. Todd and Kenneth I. Wolpin. On The Specification and Estimation of The Production Function for Cognitive Achievement. *Economic Journal*, 113(485):3–33, 2003. URL https://ideas.repec.org/a/ecj/econjl/v113y2003i485pf3-f33.html.

Petra E. Todd and Kenneth I. Wolpin. The Production of Cognitive Achievement in Children: Home, School, and Racial Test Score Gaps. *Journal of Human Capital*, 1(1):91–136, 2007. ISSN 19328575, 19328664. URL http://www.jstor.org/stable/10.1086/526401.

Raymond N. Wolfe and Scott D. Johnson. Personality as a Predictor of College Performance. *Educational and Psychological Measurement*, 55(2):177–185, April 1995. ISSN 0013-1644. doi: 10.1177/0013164495055002002. URL http://dx.doi.org/10.1177/0013164495055002002.

3.8 Tables and figures

Table 3.1: Scores on standardized tests by Spanish vs. non-Spanish mother tongue students

	Math test scores			Reading test scores		
	Mother tongue			Mother tongue		
	Non-Spanish (a)	Spanish (b)	Gap (a-b)	Non-Spanish (a)	Spanish (b)	Gap (a-b)
Last grade in						
Elementary school	−0.19	0.12	−0.31***	−0.40	−0.24	−0.64***
Junior high school	−0.25	0.14	−0.39***	−0.40	−0.22	−0.62***
High school	−0.32	0.11	−0.43***	−0.49	−0.17	−0.66***
Average			−0.38			−0.64

Note: Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. Test scores have a mean of zero and standard deviation of one at each grade. Test score means are calculated using all junior high and high school students in this table. Significant levels: *** at 1%, ** at 5%, and * at 10%.
Source: Author's calculation using the Ministry of Education's 2010, 2013 and 2015 data.

Table 3.2: Parents' educational attainment distribution in Guatemala by mother tongue

	Fathers		Mothers	
	SMT	Non-SMT	SMT	Non-SMT
Parental educational attainment				
Lower than elementary	21.1	43.0	30.6	69.2
Elementary (grades 1-3)	25.3	28.8	24.0	17.6
Elementary (grades 4-6)	28.5	21.4	24.1	10.4
Junior high	9.8	3.5	8.0	1.4
High school	10.2	2.7	10.0	1.2
More than high school	4.9	0.6	3.3	0.2
	100%	100%	100%	100%

Note: Spanish mother tongue (SMT) stands for people' self-report of having Spanish as a first language. Non-SMT people speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. The educational system in Guatemala consists of 3 stages: elementary, junior high, and high school. The elementary stage, consists of 6 grades and usually starts in the year the child turns 7, as the school year begins in January. The second and third stages, junior high and high school, consists of three grades. Attendance to college means having more than 12 years of complete education. Significant levels: *** at 1%, ** at 5%, and * at 10%.
Source: Authors' calculation using the 2002 Guatemalan national census.

Table 3.3: Students' grade progression from elementary to high school and their test scores

	Students' mother tongue			
	Spanish (62%$^{(a)}$)		Non-Spanish (38%)	
Location of elementary schools	Urban (34%)	Rural (66%)	Urban (19%)	Rural (81%)
Percentage of students in the elementary school data who reached last grade of:				
Junior high	48%	43.2%	67.5%	39.8%
High school	33.9%	24.3%	45.5%	21.4%
Students' test scores in last grade of:				
Junior high				
Math score	0.136	-0.034	-0.056	-0.252
Reading score	0.224	-0.063	-0.209	-0.494
High school				
Math score	0.102	-0.082	-0.094	-0.281
Reading score	0.233	-0.035	-0.205	-0.465

Note: Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. Test scores have a mean of zero and standard deviation of one at each grade.
(a) Percentages in parenthesis stand for the proportion of students who self-classify into this category. Number of students in the elementary school data is 18,441.
Source: Author's calculation using the Ministry of Education's 2010, 2013 and 2015 data.

Table 3.4: Descriptive statistics of variables

	Spanish mother tongue		Non-Spanish mother tongue School in urban area		Non-Spanish mother tongue School in rural area	
	Mean	Std.	Mean	Std.	Mean	Std.
Students' variables						
Do not understand Spanish (true=1)			0.194	0.395	0.028	.166
Intensively speak non-Spanish mother tongue (true=1)			0.635	0.481	0.634	0.482
Family assets	3.159	1.525	2.931	1.548	2.253	1.491
Remittances at home (true=1)	0.267	0.442	0.230	0.421	0.241	0.428
Age	0.681	1.244	0.639	1.191	1.346	1.401
Gender (male=1)	0.495	0.500	0.491	0.500	0.523	0.500
Mother years education	6.760	5.081	5.048	4.590	5.024	4.806
Father years education	7.121	5.330	6.160	4.811	5.972	5.331
Number of read books in the last year	0.708	1.208	1.306	1.401	0.460	1.058
Number days read newspaper	0.796	1.808	1.549	2.176	0.369	1.293
School's share with						
Teacher's reading perfomance (good)	0.745		0.758		0.750	
Teacher's reading perfomance (excellent)	0.210		0.199		0.194	
Teacher's reading training (good)	0.739		0.740		0.776	
Teacher's reading training (excellent)	0.230		0.228		0.183	
Teacher's math perfomance (good)	0.747		0.750		0.745	
Teacher's math perfomance (excellent)	0.203		0.203		0.192	
Teacher's math training (good)	0.725		0.729		0.747	
Teacher's math training (excellent)	0.241		0.237		0.199	
School's library (No)	0.477		0.436		0.635	
School's library (books)	0.309		0.334		0.250	
School's library (books-classroom)	0.063		0.056		0.039	
School's library (books-classroom-desks)	0.152		0.174		0.075	
Train teachers with previous national tests	0.422		0.427		0.377	
Give test scores to parents	0.233		0.235		0.225	
Train students with previous national tests	0.554		0.602		0.321	
Math classes per week (1-2)	0.254		0.331		0.163	
Math classes per week (3-4)	0.241		0.258		0.296	
Math classes per week (5-6)	0.475		0.384		0.518	
Math classes per week (7-8)	0.030		0.027		0.023	
Reading classes per week (1-2)	0.525		0.597		0.397	
Reading classes per week (3-4)	0.210		0.194		0.257	
Reading classes per week (5-6)	0.249		0.192		0.313	
Reading classes per week (7-8)	0.016		0.017		0.033	
Minutes per class (30-45)	0.895		0.935		0.786	
Minutes per class (46-60)	0.093		0.058		0.200	
Minutes per class (61-90)	0.012		0.007		0.014	
Students preparation for math national test (No-less one month)	0.472		0.457		0.640	
Students preparation for math national test (1-2 months)	0.308		0.309		0.243	
Students preparation for math national test (more 2 months)	0.220		0.235		0.117	
Students preparation for reading national test (No-less one month)	0.477		0.466		0.642	
Students preparation for reading national test (1-2 months)	0.302		0.302		0.247	
Students preparation for reading national test (more 2 months)	0.221		0.232		0.111	
Students preparation for math national test (hours)	1.908		2.087		1.759	
Students preparation for reading national test (hours)	1.900		1.939		1.784	
Meetings for students' performance (every 15 days)	0.354		0.309		0.423	
Meetings for students' performance (every 1 month)	0.478		0.490		0.420	
Meetings for students' performance (every 2 months)	0.165		0.197		0.153	
Reading test at school (every week)	0.357		0.298		0.503	
Reading test at school (every 1 month)	0.218		0.202		0.233	
Reading test at school (every 2 months)	0.143		0.150		0.147	
Math test at school (every week)	0.355		0.302		0.475	
Math test at school (every 1 month)	0.199		0.185		0.219	
Math test at school (every 2 months)	0.165		0.164		0.189	

Note: Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. The variable Do not you understand Spanish? stands for non-SMT student's self-evaluation of not being proficient in Spanish. This variable takes a value of one only if non-SMT students reported not being proficient in Spanish and zero otherwise.
Source: Author's calculation using the Ministry of Education's 2010, 2013 and 2015 data.

Table 3.5: Distribution of SMT students across municipalities and non-SMT students' mother tongue use[(a)]

		Non-SMT students who reported mother tongue use at	
Municipal distribution	Percentage of SMT students	Junior high school	High school
18%	0-25%	81%	75%
14%	25-50%	54%	52%
20%	50-75%	30%	31%
48%	75-100%	17%	13%

(a) Way of reading the table: in 18% of municipalities (1) between 0 and 25% of SMT students reside, and (2) 81% of non-SMT students speaks their mother tongue at junior high school.
Note: Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. Guatemala is geographically divided by more than 300 municipalities.
Source: Author's calculation using the Ministry of Education's 2010, 2013 and 2015 data.

Table 3.6: Students' selection in school progression and their test scores

	Students who attended school until the last grade in					
	Elementary school		Junior high school		High school	
	Reading	Math	Reading	Math	Reading	Math
Students' average scores for						
elementary school	-0.186	-0.123	0.016	0.001	0.367	0.250
junior high school			-0.233	-0.134	0.155	0.089
School's average scores for						
elementary school	-0.099	-0.058	-0.009	-0.027	0.194	0.128
junior high school			-0.148	-0.082	0.098	0.055

Note: Test scores have a mean of zero and standard deviation of one at each grade. The student's average scores are calculated using only those sixth grade students who can be followed over grades. The school's average scores are calculated using all junior high and high school students in this table.
Source: Author's calculation using the Ministry of Education's 2010, 2013 and 2015 data.

Table 3.7: Estimates of the effect of a lack of Spanish comprehension (the language barrier effect) on students' math test scores when controlling for student fixed effects

Second stages for math production functions[(a)]				
	(1) OLS	(2) 2SLS	(3) 2SLS	(4) 2SLS
Δ Do you not understand Spanish? ($Ho: \beta_3 \geqslant 0$)	-0.061^{**}	-0.346^{**}	-0.469^{*}	-0.705^{**}
Δ Do you intensively speak a non-Spanish language? ($Ho: \beta_4 \geqslant 0$)	0.032	0.099	0.121	-1.192
$\beta_3 + \beta_4$ ($Ho: \beta_3 + \beta_4 \geqslant 0$)	-0.029	-0.246^{**}	-0.348^{**}	-1.897^{*}
Obs.	3458	3393	3393	3393
Instrumented variables				
Do you not understand Spanish?		Yes	Yes	Yes
Sorting across schools[(b)]			Yes	Yes
Do you intensively speak a non-Spanish language?				Yes

Note: The dependent variable is the change of students' math test scores over time. Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. The variable Do not you understand Spanish? stands for non-SMT student's self-evaluation of not being proficient in Spanish. This variable takes a value of one only if non-SMT students reported not being proficient in Spanish and zero otherwise. The variable Do you intensively speak a non-Spanish language? stands for non-SMT student's mother tongue use. This variable takes a value of one only if non-SMT students reported intensive use of their mother tongue and zero otherwise. Significant levels: *** at 1%, ** at 5%, and * at 10%.
(a) All model specifications control for both all variables shown in Table 3.4.
(b) Models (3) and (4) treat both the sorting of students across schools with the bilingual program implemented, and the linguistic sorting of students across schools as endogenous.
Source: Author's calculation using the Ministry of Education's 2010, 2013 and 2015 data.

Table 3.8: Estimates of the effect of a lack of Spanish comprehension (the language barrier effect) on students' reading test scores when controlling for student fixed effects

Second stages for reading production functions[a]				
	(1) OLS	(2) 2SLS	(3) 2SLS	(4) 2SLS
Δ Do you not understand Spanish? ($Ho: \beta_3 = 0$)	0.038	−0.161	−0.149	−0.153
Δ Do you intensively speak a non-Spanish language? ($Ho: \beta_4 = 0$)	0.037	0.082	0.071	0.045
$\beta_3 + \beta_4$ ($Ho: \beta_3 + \beta_4 \geqslant 0$)	0.075	−0.078	−0.078	−0.108
Obs.	3458	3393	3393	3393
Instrumented variables				
Do you not understand Spanish?		Yes	Yes	Yes
Sorting across schools [b]			Yes	Yes
Do you intensively speak a non-Spanish language?				Yes

Note: The dependent variable is the change of students' reading test scores over time. Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. The variable Do you not understand Spanish? stands for non-SMT student's self-evaluation of not being proficient in Spanish. This variable takes a value of one only if non-SMT students reported not being proficient in Spanish and zero otherwise. The variable Do you intensively speak a non-Spanish language? stands for non-SMT student's mother tongue use. This variable takes a value of one only if non-SMT students reported intensive use of their mother tongue and zero otherwise. Significant levels: *** at 1%, ** at 5%, and * at 10%.
(a) All model specifications control for both all variables shown in Table 3.4.
(b) Models (3) and (4) treat both the sorting of students across schools with the bilingual program implemented, and the linguistic sorting of students across schools as endogenous.
Source: Author's calculation using the Ministry of Education's 2010, 2013 and 2015 data.

Table 3.9: Estimates of the effect of a lack of Spanish comprehension (the language barrier effect) on students' math test scores when controlling for both student fixed effects and students' innate ability[a]

Second stages for math production functions[b]				
	(1) OLS	(2) 2SLS	(3) 2SLS	(4) 2SLS
Δ Do you not understand Spanish? ($Ho: \pi_3 \geqslant 0$)	−0.065**	−0.278*	−0.443*	−0.659**
Δ Do you intensively speak a non-Spanish language? ($Ho: \pi_4 \geqslant 0$)	0.021	0.071	0.111	−1.113
$\pi_3 + \pi_4$ ($Ho: \pi_3 + \pi_4 \geqslant 0$)	−0.043	−0.207*	−0.332*	−1.771*
Obs.	3458	3393	3393	3393
Instrumented variables				
Do you not understand Spanish?		Yes	Yes	Yes
Sorting across schools[c]			Yes	Yes
Do you intensively speak a non-Spanish language?				Yes

Note: The dependent variable is the change of students' math test scores over time. Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. The variable Do you not understand Spanish? stands for non-SMT student's self-evaluation of not being proficient in Spanish. This variable takes a value of one only if non-SMT students reported not being proficient in Spanish and zero otherwise. The variable Do you intensively speak a non-Spanish language? stands for non-SMT student's mother tongue use. This variable takes a value of one only if non-SMT students reported intensive use of their mother tongue and zero otherwise. Significant levels: *** at 1%, ** at 5%, and * at 10%.
(a) Section 3.4.3 explains how to deal with students' abilitiy.
(b) All model specifications control for both all variables shown in Table 3.4.
(c) Models (3) and (4) treat both the sorting of students across schools with the bilingual program implemented, and the linguistic sorting of students across schools as endogenous.
Source: Author's calculation using the Ministry of Education's 2010, 2013 and 2015 data.

Table 3.10: Estimates of the effect of a lack of Spanish comprehension (the language barrier effect) on students' math test scores when controlling for students' innate ability[(a)]

Second stages for math production functions[(b)]						
	(1) OLS	(2) OLS	(3) 2SLS	(4) 2SLS	(5) 2SLS	(6) 2SLS
Non-SMT student (β_2)	−0.010	0.038	0.282***	0.329***	0.443**	0.559***
Do you not understand Spanish? ($Ho: \beta_3 \geqslant 0$)		−0.042	−0.110***	−0.523*	−1.523***	−1.084*
Do you intensively speak a non-Spanish language? ($Ho: \beta_4 \geqslant 0$)		−0.050*	−0.206***	−0.126*	0.118	−0.725*
$\beta_3 + \beta_4$ ($Ho: \beta_3 + \beta_4 \geqslant 0$)		−0.092**	−0.316***	−0.649**	−1.405***	−1.809***
Obs.	9032	9032	9024	9024	9024	9024
Instrumented variables						
Non-SMT student			Yes	Yes	Yes	Yes
Do you not understand Spanish?				Yes	Yes	Yes
Sorting across schools[(c)]					Yes	Yes
Do you intensively speak a non-Spanish language?						Yes

First stages for Do you not understand Spanish?[(d)]						
	(1)	(2)	(3)	(4)	(5)	(6)
Non-SMT father				0.018	0.021	0.012
Non-SMT mother				0.032	0.029	0.032
Non-SMT both parents				−0.063**	−0.067**	−0.049
Proportion SMT students at school s in $t-1$ $\left(\sigma^{SMT}_{s_i,t-1}\right)$				−0.050***	−0.038***	−0.080***
$\left(\sigma^{SMT}_{s_i,t-1}\right)^2$				−0.017***	−0.017***	−0.027***
Proportion SMT students surrounding school s in t $\left(\sigma^{SMT}_{-s,t}\right)$				−0.013	−0.026	−0.010
$\left(\sigma^{SMT}_{-s,t}\right)^2$				−0.017**	−0.025***	−0.017**
School proximity to main municipal city in t				0.389	0.383	0.488*
School proximity to main municipal city in t-1				−0.014	−0.010	0.000
$Pr\left(Ho: \sum_k \sigma_{z.k} = 0\right)$				0.00	0.00	0.00

First stages for Do you intensively speak a non-Spanish language?[(e)]						
	(1)	(2)	(3)	(4)	(5)	(6)
Non-SMT father						−0.032
Non-SMT mother						0.012
Non-SMT both parents						0.068
Proportion SMT students at school s in $t-1$ $\left(\sigma^{SMT}_{s_i,t-1}\right)$						−0.162***
$\left(\sigma^{SMT}_{s_i,t-1}\right)^2$						−0.038***
Proportion SMT students surrounding school s in t $\left(\sigma^{SMT}_{-s,t}\right)$						0.062**
$\left(\sigma^{SMT}_{-s,t}\right)^2$						0.030**
School proximity to main municipal city in t						0.399
School proximity to main municipal city in t-1						0.038**
$Pr\left(Ho: \sum_k \sigma_{z.k} = 0\right)$						0.00

Note: The dependent variable is students' math test scores. Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. Significant levels: *** at 1%, ** at 5%, and * at 10%.
(a) Section 3.4.3 explains how to deal with students' abilitiy.
(b) All model specifications control for all variables shown in Table 3.4, for municipal, rural area, school type, and year fixed effects. Guatemala is geographically divided by more than 300 municipalities.
(c) Models (3) and (4) treat both the sorting of students across schools with the bilingual program implemented, and the linguistic sorting of students across schools as endogenous.
(d) The variable Do not you understand Spanish? stands for non-SMT student's self-evaluation of not being proficient in Spanish. This variable takes a value of one only if non-SMT students reported not being proficient in Spanish and zero otherwise.
(e) The variable Do you intensively speak a non-Spanish language? stands for non-SMT student's mother tongue use. This variable takes a value of one only if non-SMT students reported intensive use of their mother tongue and zero otherwise.
Source: Author's calculation using the Ministry of Education's 2010, 2013 and 2015 data.

Table 3.11: Estimate of the effect of a lack of Spanish comprehension (the language barrier effect) on students' math test scores when controlling for students' unobserved heterogeneity

Do you not understand Spanish?[(a)]	−1.428***

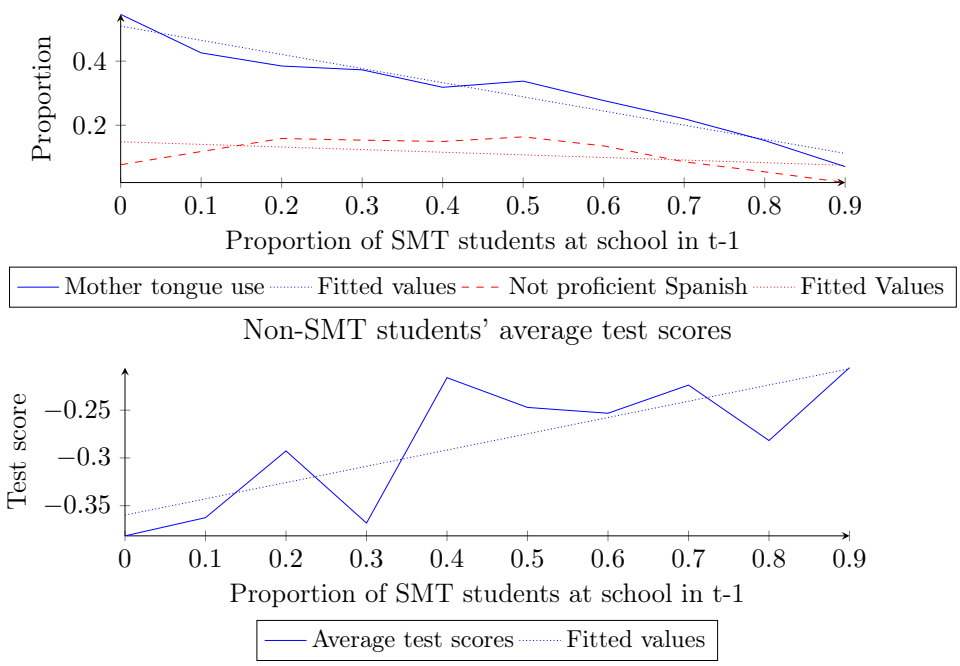
Note: The dependent variable is students' math test scores. Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. All model specifications control for all variables shown in Table 3.4, and for municipal, rural area, school type, and year fixed effects. Guatemala is geographically divided by more than 300 municipalities. This model does not include those variables treated as endogenous, but includes reading test scores as independent variables to control for students' innate ability.
(a) The variable Do you not understand Spanish? stands for non-SMT student's self-evaluation of not being proficient in Spanish. This variable takes a value of one only if non-SMT students reported not being proficient in Spanish and zero otherwise.
Significant levels: *** at 1%, ** at 5%, and * at 10%.
Source: Author's calculation using the Ministry of Education's 2010, 2013 and 2015 data.

Table 3.12: Estimate of the effect of a lack of Spanish comprehension (the language barrier effect) on students' math test scores when simulating Spanish as a second language program by rural vs. urban areas[(a)]

$E\left(y_{z_1,t} - y_{z_0,t} \mid Rural\ area\right)$	-0.705^{***}
$E\left(y_{z_1,t} - y_{z_0,t} \mid Urban\ area\right)$	0.001
$E\left(y_{z_1,t} - y_{z_0,t} \mid Rural\ area\right) + E\left(y_{z_1,t} - y_{z_0,t} \mid Urban\ area\right)$	-0.704^{***}

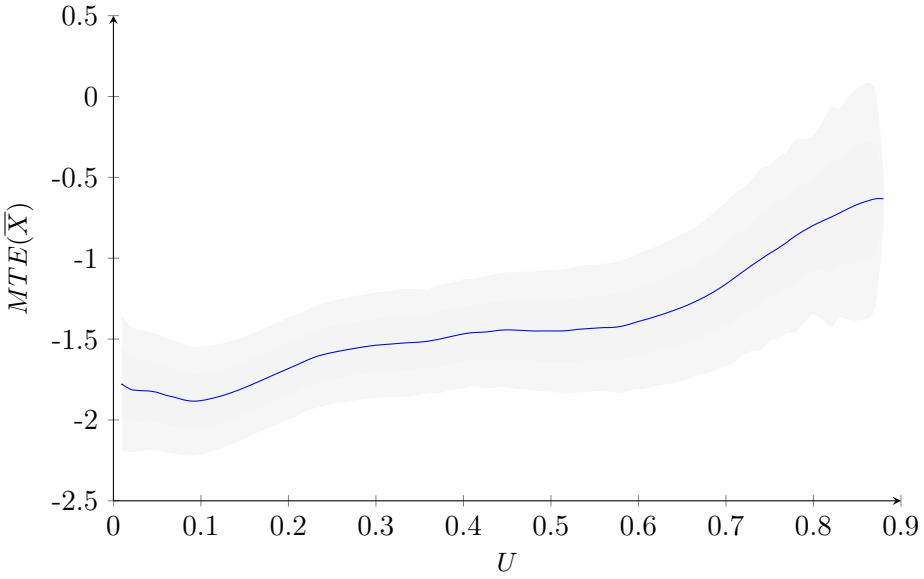
Note: The policy, z_1, artificially changes the students' linguistic sorting across schools to simulate an increase of non-SMT students' interaction with native Spanish speakers, which can result in non-SMT students' Spanish comprehension improvement. This policy can be understood as a second language program. Specifically, this policy simulation looks for evidence of what could have happened, in terms of math test scores, $y_{z_1,t}$, if SMT students would have attended schools in rural areas, where the majority of non-SMT students attend. Non-SMT students speak one of the Mayan, Xinka or Garífuna languages as a first language.
Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language.
Significant levels: *** at 1%, ** at 5%, and * at 10%.
Source: Author's calculation using the Ministry of Education's 2010, 2013 and 2015 data.

Figure 3.1: Non-SMT students' mother tongue use, self-evaluation of not being proficient in Spanish, and their test scores



Note: Spanish mother tongue (SMT) stands for students' report of having Spanish as a first language. Non-SMT students speak one of the Mayan, Xinka or Garífuna languages as a first language. This graph shows the average of math and reading test scores for non-SMT students.
Source: Author's calculation using the Ministry of education's data.

Figure 3.2: The marginal effect of a lack of Spanish comprehension (the language barrier effect) on math test scores (MTE)



Note: Conditional on X, the graph depicts the average language barrier effect on math test scores for a individual who is indifferente between improving or not his Spanish comprehension. The higher the probability that unobserved factors reduce students' lack of Spanish comprehension, U, the lower the language barrier effect on math test scores.
Source: Author's calculation using the Ministry of education's data.

Chapter 4

The root of Guatemalan students' language barrier: parental preferences for school attributes or spatial segregation of groups?

4.1 Introduction

In the development literature, educational achievement on school tests plays an important role in explaining growth differences across countries (Hanushek and Woessmann, 2012a). In fact, the perceived Latin American low growth puzzle is also resolved by educational achievement (Hanushek and Woessmann, 2012b). On the other hand, a number of papers find that social conflicts resulting from antagonism between ethnic groups (ethnic polarization) affects countries' development through investment.[1] Other empirical papers use country's ethnolinguistic composition or inequality as a proxy of ethnic polarization and find a negative and significant impact on either growth or educational attainment.[2] While the development literature finds evidence that all the above factors may affect countries' development, there is little empirical evidence to support how ethnic polarization and a country's ethnolinguistic composition or inequality may interact to affect educational achievement and, therefore, growth through other mechanisms rather than conventional mechanisms of investment or government's provision of public goods (Montalvo and Reynal-Querol, 2005a).

[1] For example, Montalvo and Reynal-Querol (2002) construct a variable that represents the distance across groups (ethnic polarization). The authors find evidence that countries with a highly polarized population are more likely to have social conflicts. These social conflicts affect countries' long run economic growth through investment or government consumption.

[2] See Burgess et al. (2015), Alesina et al. (2016), and Montalvo and Reynal-Querol (2005a) for instance.

To help interpret how ethnic polarization may interact with a society's ethnolinguistic composition, I estimate a model of demand for junior high schools in Guatemala in which parents consider schools as differentiated products. In particular, I allow for the degree of differentiation across schools to depend on characteristics of other parents who select the school in equilibrium. This school differentiation is in line with the international literature for school choice. For instance, Caetano and Maheshri (2017) find in their study for Los Angeles that parents select schools where other parents are of the same race. This relationship is stronger in the higher grades. Carneiro et al. (2016) find evidence that, in Pakistan, the lower the mother's educational attainment, the higher her preference for selecting a school where mothers have a similar level of education. Burgess et al. (2015) show that, in England, parents prefer schools where students have English as an additional language. Therefore, parents' school selection may play an important role in creating group segregation at school, which may lead students to have poor educational achievement at school.[3] In a polarized society, parents may have preferences for a school's race or language composition, which lead to linguistic segregation across schools.

The determinants of schools' linguistic composition are important to identify to observe whether parents' enrollment decisions show the effect of a polarized population: preferences for race or language. This appears to be a first-order issue for countries where more than one language is spoken. Guatemala has 25 ethnolinguistic groups, and about 60 percent of the total population has Spanish as a mother tongue (SMT). Spanish is the predominant language of instruction at secondary school.

Guatemala is an interesting country to address the gap of how ethnic polarization may interact with a society's ethnolinguistic composition. Guatemala has a polarized population, an ethnolinguistic composition which is close to the average world, and ethnolinguistic inequality (Desmet et al., 2018; Alesina et al., 2016; Montalvo and Reynal-Querol, 2005b). The 2005 report of the United Nations Development Program (UNDP) explains that, in Guatemala, there exists discrimination that mainly affects the indigenous population and that the existence of such practices may be due to either power, race, culture or social status. If there exists antagonism among ethnic groups in Guatemala, parents might prefer to sort their child into schools where parents or students have a similar race or language. This ethnolinguistic sorting in schools is hindering non-SMT students' Spanish comprehension (a language barrier) as shown in Chapter 3, which may have short and long-term consequences for the Guatemalan non-SMT population.

This linguistic sorting of students across schools is affecting non-SMT students' educational achievement. Table 4.1 shows standardized test scores for grade six students in 2010 by mother tongue.[4] The low performance of non-SMT students is persistent from the lower

[3]For instance, Hanushek et al. (2009) find that racial composition at school affects students' educational achievement.

[4]Test scores are standardized to have a mean of zero and standard deviation of one at each grade. Test scores represent students' school performance in the last year of elementary, junior high, and high school.

to the higher grades and does not show any significant improvement. For instance, some studies for Guatemala argue that the low educational achievement of non-SMT students is due to either schools' lack of a bilingual education program or books in non-Spanish languages (Marshall, 2009; Patrinos and Velez, 2009), and students' lack of interaction with people proficient in Spanish as Chapter 3 indicates.

Furthermore, the long-term consequences of non-SMT students' language barrier may be driving the non-SMT population's low educational attainment and, therefore, poverty. The 2002 Guatemalan census (see Table 4.2) shows the education levels of SMT and non-SMT parents with a 2-year-old child in 2002.[5] The numbers clearly reveal the lower education level of non-SMT parents, especially mothers.[6] In the context of poverty or inequality, the World Bank reported in 2009 indicates that Guatemala has one of the highest inequality rates in Latin America and some of the worst poverty, especially in rural and indigenous areas. This is also observed in the 2014 Guatemalan national survey. The survey shows that, on average, SMT workers earn about 37 percent more than non-SMT workers who still intensively speak their non-Spanish mother tongue. However, this gap is reduced to close to 6 percent for non-SMT workers who frequently speak Spanish.

The determinants of schools' linguistic composition are relevant to examine for Guatemala given the language barrier's existence and its possible long-term effects on non-SMT students. Specifically, this chapter looks for evidence of whether non-SMT parents have preferences for schools where other parents have a similar mother tongue, which can reflect social polarization associated with antagonistic feelings between groups. Ultimately, if antagonistic feelings between groups exist in Guatemala as the Guatemalan literature suggests, such a behavior must come from an underlying mechanism. While distinguishing such an underlying mechanism may be of interest per se, the main objective in this chapter is only to analyze parent's enrollment decisions. Therefore, with a polarization effect in this chapter I just mean the lack of interaction between non-SMT and SMT groups as measured by parental preferences to sort students by mother tongue.[7]

To identify the determinants of schools' linguistic composition, this chapter empirically examines parents' enrollment decisions for junior high schools in 2013 as in Carneiro et al. (2016) and Bayer and Timmins (2007). In Guatemala, a parent can enroll their child at any school. However, parents' enrollment decisions may depend not only on their own characteristics and on school attributes, but also on other parents' characteristics that select the school in equilibrium. Furthermore, non-SMT parents may prioritize schools in which

[5]This distribution may potentially represent the parents' educational distribution of junior high students at school in 2013.

[6]The Guatemalan civil war may have also played an important role in shaping the Guatemalan education attainment distribution, as discussed in Chamarbagwala and Morán (2011).

[7]In Section 4.5, I briefly discuss the possible underlying mechanism for antagonism among ethnic groups in Guatemala. The Guatemalan civil war affected the most the indigenous population (Chamarbagwala and Morán, 2011). I find that the higher the number of victims of the civil war in a specific region, the higher the parental preferences to sort students by mother tongue.

their child is likely to speak/learn Spanish, since Spanish is the predominant language of instruction at schools and communication at jobs. I include as a school attribute the child's probability of speaking a non-Spanish language at school. The estimation of a model of demand for schools allows me to control for all the above factors. Then, I document whether the root of the language barrier is specifically due to non-SMT parents' preferences for people with a similar mother tongue, or just the result of spatial segregation of groups. In this chapter, spatial segregation stands for household home-to-capital city proximity and school-to-nearest municipal city proximity.[8]

The data comes from the Ministry of Education (Mineduc) of Guatemala. This dataset contains household and student data in the last grade of junior high in 2013 such as parents' educational attainment and students' mother tongue. Additionally, Mineduc also collects detailed data about schools such as infrastructure and number of classes per week. Similar to the international literature, this dataset allows me to control for distance, peer effects and school fees.

Two features are important to highlight from the dataset in this chapter, since it is possible to identify which students attend which schools. First, the school-level data represents a household census of school choice which allows me to construct shares of enrolled students at school or to construct indices that represent the characteristics of the student body of each school, such as average test scores or shares of non-SMT students. Second, both school's shares of enrolled students and household-level data allow me to estimate parents' preferences for school attributes that depend on other parents' characteristics.

The estimation of a model of demand for differentiated schools for junior high students in Guatemala allows me to make three contributions to the literature. First, the development literature has shown that ethnic polarization affects economic outcomes such as growth via its effect on investment or government's provision of public goods. This chapter contributes to this literature by showing evidence of an alternative mechanism by which ethnic polarization and a country's ethnolinguistic composition interact to affect growth: ethnolinguistic segregation at school. In countries where more than one language is spoken, parents' preferences to segregate themselves by mother tongue can contribute to linguistic segregation at school. Then, not being proficient in the language of instruction at school may hinder human capital development which affects both growth and ethnolinguistic inequality.

Second, in the context of Guatemala, a school's ethnolinguistic composition is a prominent channel for students' language barrier (see Chapter 3). As a result of this language barrier, non-SMT students who are mostly taught in Spanish show low educational achievement at secondary school. This fourth chapter sheds light on the source of this language barrier. I contribute to the Guatemalan literature by showing that non-SMT parents value

[8]Guatemala is geographically divided by departments, and departments by municipalities. In total Guatemala has 22 departments and 337 municipalities. Municipalities contain both rural and urban areas.

schools in which their child is likely to speak/learn Spanish. However, non-SMT parents prefer to sort their child into schools where other parents have a similar mother tongue. This finding may be the root of the non-SMT students' language barrier. This latter preference dominates the former as we move away from the Guatemalan capital city, which leads to both spatial and linguistic segregation at school.

Third, a number of papers in the school choice literature find that key determinants of parental enrollment decisions are household home-to-school distance and school fees. Based on these determinants, these papers propose policies such as subsidies or cash transfer programs to increase educational attainment. However, targeting a cash transfer program to non-SMT parents may not eliminate non-SMT parents' preferences for their own mother tongue, keeping linguistic segregation at school. This chapter contributes to the school choice literature by recommending a policy to promote ethnolinguistic integration at rural school where the majority of non-SMT students attend. I calculate, in terms of school attributes, the compensation that SMT parents should receive if they enroll their child into their nearest rural school instead of their optimal urban school. I show that some school attributes that SMT parents value are school infrastructure and which classes are offered. However, the most important school attribute for SMT parents is school quality as measured by test scores.

The next section describes both the Guatemalan education system and the data. Section 4.3 presents the parents' school choice model and the estimation approach. How parents infer their child's expected mother tongue use at school is discussed as well. The main findings are presented in Section 4.4. In Section 4.5, I discuss a policy recommendation for school choice and last, I conclude.

4.2 The Guatemalan education system and the data

The Guatemalan education system can be divided into four school types: public, private, cooperative, and municipal. Public schools are completely managed by the Ministry of Education (Mineduc), and only public schools are free. Private schools are profit maximizing. Municipal schools are managed by local governments rather than by Mineduc. Finally, in cooperative schools, the schools' administration is formed by parents and the local government. Regardless of the school type, the curriculum and all syllabi are designed by Mineduc, and all schools are required to cover the Mineduc curriculum.

In 2013, Mineduc tested grade nine students in math and reading. Concurrently, students completed a survey which asked questions about school, family inputs and house infrastructure. In addition to the student-level data, school-level data are available for schools in 2013 as well. Principals provide information in the following areas: 1) school infrastructure, 2) school test frequency and 3) the students' preparation for the national tests. Both datasets

contain a school code variable that allows matching of students to schools.

The literature about students' sorting across schools has shown three main factors that influence parents' school choice: distance to school, peer effects, and fees.[9] To construct the variable home-to-school proximity, I employ the latitude and longitude coordinates of both students' home and junior high schools to calculate a measure of home-to-school distance in kms. However, I rescale this home-to-school distance.[10] A value of one means that students live near by the school, whereas a value of zero represents the highest distance observed in the dataset.

A particular characteristic of the Guatemalan education system is that parents can choose any school for their children. So, depending on parents' preferences for school characteristics such as a school's distance from home, parents select the best school given their own characteristics. Table 4.3 shows the average distance that students who live in rural areas travel to their schools conditional on family income and their home's proximity to the capital city. The table clearly reveals that the closer students live to the capital city, the lower both the distance from home to school and its dispersion. On average, those students who live near the capital city travel around 3.21 kms while those living far from the capital city travel 8.06 kms. A second feature displayed in this table is the role of family income on parents' school selection. Families with high income show greater dispersion, in terms of distance, than families with low income when selecting schools. Furthermore, among families with high income, this dispersion is greater for students living far away from the capital city than for those students living in the capital city. I control for family income in the estimation.

Then, one potential determinant of the language barrier may be spatial segregation of groups given that families with low socio-economic status cannot commute to places where Spanish is commonly spoken. Guatemala is geographically divided by departments and, then, by municipalities. Each municipality contains both rural and urban areas. Table 4.4 shows the share of non-SMT students who reported mother tongue use, for both junior high and high school, by municipalities' SMT student share. This table indicates that the higher the municipalities' SMT student share, the lower the proportion of non-SMT students who reported mother tongue use. For example, 32 percent of municipalities have less than 50 percent of SMT students at school, and in these municipalities more than 52 percent of non-SMT students reported using their non-Spanish mother tongue. Therefore, non-SMT students' linguistic sorting across municipalities may be an important factor influencing their Spanish proficiency. In the estimation I control for both school fixed effects and departmental

[9] A drawback in this dataset is that family income and school fees are not observed. To overcome this, I employ the 2014 Guatemalan national survey to construct these variables. See appendix for details.

[10] As a proxy of students' home location, I use schools' location where students attended grade six at public schools (last grade of elementary school). I assume grade six students live near by their schools. In Guatemala, public schools are different for elementary and secondary students. Then, these grade six students had to attend different schools at grade nine. The student-level data in this chapter includes only those students that I can observe in grade six at public schools and that I follow to grade nine. I rescale distances for computational purposes.

fixed effects.

If a language barrier exists as a result of municipalities' demographic characteristics, then parents may consider a school with the bilingual education program. In Guatemala, schools with the bilingual program have helped elementary non-SMT students. The aim of this program is to improve educational outcomes for non-SMT students such as increasing the likelihood of grade progression and reducing drop-out rates (Rubio, 2004). Some studies find that this program positively affects outcomes for non-SMT elementary students (Marshall, 2009; Rubio, 2004; Enge and Chesterfield, 1996); however, the program is mainly targeted towards students in the first three grades of elementary schools. Therefore, very few secondary schools offer a bilingual education program. For instance, less than 2 percent of junior high students attend schools with a bilingual education program. The bilingual education program is one possible school attribute parents might consider when choosing a school. In the estimation, I control for both bilingual education program at school and also for principal's non-Spanish language use.

The dataset contains a rich set of school attributes to control for in the estimation of the model. Controlling for many variables dramatically increase the computational burden. Therefore, I perform a data reduction by principal component analysis and employ the first component or index in the model estimation.[11] I group school attributes to construct six school quality indices: teachers, courses, classes, infrastructure, discipline, and distraction.[12] In short, the teacher quality index consists of the principal's perception of the teachers' ability to teach. The course index represents the type of courses that schools offer such as English or computation. The classes quality index incorporates variables such as number of classes per week and minutes per class. The infrastructure index includes variables like school's construction, ventilation and areas to practice sports. The second to last index is school discipline. This group has variables such as how many times students are tested at school, how often principals supervise teachers, and whether principals have improved their teachers' supervision. Finally, a school's index for distraction uses variables such as school temperature and noise.

Table 4.5 gives a description of both family and school variables. This table indicates that 20 percent of households have two non-SMT parents while just 5 percent have only one SMT parent. Similar to the Guatemalan educational distribution (see Table 4.2), in this chapter, parents also show a low educational attainment, especially for mothers. In this table, the family income variable is standardized so that a value of zero represents the average family income in the capital city. Households, on average, have a family income of 0.57 standard deviations lower than a representative family income in the capital city. The last family variable, home-to-capital city proximity, indicates that, on average, families do not reside in the capital city.

[11] I only include the first component due to the computational burden during the estimation. By adding the second component for all indices, the number of extra parameters to estimate is 64.

[12] See appendix for details.

Regarding school attributes, only 1.9 percent of schools have an established bilingual program. Private schools represent about 32.5 percent of all junior high schools, and 48.1 percent of all schools are located in rural areas. Interestingly, on average, school are closer to the capital city than students' home. In the table, school quality indices are also standardized so that a value of zero represents the average quality located in the capital city. All school indices are at least 0.28 standard deviations below the average school quality index in the capital city. The last group of variables indicates that parents, on average, pay Q103 ($13.30 USD) per month for non-public schools, and school test scores are 0.11 standard deviations below the average test score in the capital city.[13] The household's average educational attainment at school is seven years (one year more than elementary school).[14] The last three variables indicate that, on average, 24.6 percent of students still speak their mother tongue language rather than Spanish. On average, 38.4 percent of students have a non-Spanish mother tongue at school. Lastly, the percentage of non-SMT parents at school is 26.3, on average.

4.3 Model

In this section I first describe the model to understand parents' enrollment decision for their child. Then, I explain how the model is estimated, and the key features of data that makes possible the estimation of the model. Last, I discuss the assumptions I make to control for parents' priority for their child's non-Spanish language use.

4.3.1 Parents' school choice problem

Parents can choose any school for their child in the Guatemalan education system. So, depending on parents' preferences for school attributes, parents select the best school to enroll their child. However, these preferences may differ across parents due to parents' own characteristics.

In the model, a parent sorts their child into a school based on the attributes of the school. Specifically, a parent p faces a school choice set indexed by $s = 1, ..., S$. For each school, a parent observes both school attributes that are exogenous, $X_s = [x_{s,1} \cdots x_{s,J}]$, and also school attributes that are endogenous or that result from an equilibrium, $\sigma_s = [\sigma_{s,1} \cdots \sigma_{s,L}]$ such as the share of non-SMT students. In addition to these school attributes, X_s and σ_s, parent p also has preferences for both home-to-school proximity, $d_{p,s}$ and the Guatemalan department at which the school is located, δ_D. These departmental dummy variables may

[13] The abbreviations Q and USD stand for Quetzales and US dollars, respectively.
Relative to an exchange rate of Q7.75 per $1 USD, the schools fees would be about $13.30 USD.

[14] I define the household's educational attainment as the highest educational attainment between parents.

capture parents' preferences for other school attributes that are fixed at departmental level. Furthermore, a parent can derive utility from unobserved school attributes as well. I assume that a parent has unobserved preferences that are common across parents, ξ_s, and parent specific, $\xi_{p,s}$, for school s. I assume that ξ_s are invariant to the school decisions made by the parents in the model since σ_s controls for school attributes resulting from an equilibrium process.

A parent derives utility for a school by weighting its school attributes X_s and σ_s. These weights stand for the parent's preferences for school attributes. Let $\beta_{p,j}$ be the preference for the school's exogenous attribute j, and $\alpha_{p,l}$ be the preference for the school's endogenous attribute l, respectively, for parent p. I assume that the parent's preference for the school attribute j, $\beta_{p,j}$, consists of two parts: a common preference across parents $\overline{\beta}_j$, and a parent-specific preference which depends on parent's observed and unobserved characteristics, $Z_p = [z_{p,1} \cdots z_{p,K}]$ and ϵ_p, respectively. I make the same assumption for $\alpha_{p,l}$. See Table 4.5 for a description of both household and school variables.

Given preferences for school attributes, the utility that parent p derives from school s is given by:

$$U_{p,s} = \sum_{j=1}^{J} x_{s,j}\beta_{p,j} + \sum_{l=1}^{L} \sigma_{s,l}\alpha_{p,l} + \lambda d_{p,s} + \delta_D + \xi_s + \xi_{p,s} \tag{4.1}$$

$$\beta_{p,j} = \overline{\beta}_j + \sum_{k=1}^{K} z_{p,k}\beta_{k,j} + \beta_j^u \epsilon_p \tag{4.2}$$

$$\alpha_{p,l} = \overline{\alpha}_l + \sum_{k=1}^{K} z_{p,k}\alpha_{k,l} + \alpha_l^u \epsilon_p. \tag{4.3}$$

The parameters $\beta_p = [\beta_{p,1} \cdots \beta_{p,J}]$, $\alpha_p = [\alpha_{p,1} \cdots \alpha_{p,L}]$ and λ capture the parent's preferences for exogenous and endogenous school attributes, and home-to-school proximity, respectively. Parameters β_j^u and α_l^u stand for preferences from a parent's unobserved characteristics.

Reorganizing the parent's total utility for school s, $U_{p,s}$, into three types of utilities: (1) school mean utility, δ_s, (2) utility from parents' observed characteristics, $\delta_{p,s}$, and (3) utility from parents' unobserved characteristics, $\psi_{p,s}$, then the system can be rewritten as:

$$U_{p,s} = \delta_s + \delta_{p,s} + \psi_{p,s} + \xi_{p,s} \tag{4.4}$$

$$\delta_s = X_s\overline{\beta} + \sigma_s\overline{\alpha} + \delta_D + \xi_s \tag{4.5}$$

$$\delta_{p,s} = \sum_{j=1}^{J}\sum_{k=1}^{K} x_{s,j} z_{p,k} \beta_{k,j} + \sum_{l=1}^{L}\sum_{k=1}^{K} \sigma_{s,l} z_{p,k} \alpha_{k,l} + d_{p,s}\lambda \tag{4.6}$$

$$\psi_{p,s} = \sum_{j=1}^{J} x_{s,j}\epsilon_p\beta_j^u + \sum_{l=1}^{L} \sigma_{s,l}\epsilon_p\alpha_l^u. \tag{4.7}$$

Equation 4.4 indicates the total utility that parent p would derive for school s. Then, the probability that parent p chooses school s is $Pr\,(S = s) = Prob\,(\xi_{p,r} < U_{p,s} - U_{p,r} + \xi_{p,s}\ \forall r \neq s)$.

To represent parent's utility function as a conditional logit model, the parent's unobserved utilities for schools, $\xi_{p,s}$, do not have to be correlated over school alternatives. By treating parent's mean utilities, δ_s, as school fixed effects in equation 4.4, parent's unobserved utilities, $\xi_{p,s}$, are not correlated with any unobserved attributes of schools, ξ_s.[15] Therefore, I assume that $\xi_{p,s}$ follows an independent and identically distributed extreme value type one distribution to represent the parent's utility function as a conditional logit model:

$$Pr(S = s \mid X_s, \sigma_s, Z_p, \epsilon_s) = \frac{e^{\delta_s + \delta_{p,s} + \psi_{p,s}}}{\sum_{g=1}^{S} e^{\delta_g + \delta_{p,g} + \psi_{p,g}}}. \tag{4.8}$$

Equation 4.8 gives the probability that parent p selects the school s for their child i given both observed ($\delta_{p,s}$) and unobserved ($\psi_{p,s}$) characteristics of the parent (equations 4.6 and 4.7), and school mean utility, δ_s.

4.3.2 Estimation of the random coefficient model

Estimation of the model consists of two steps given both the representation of the parent's utility function as a conditional multinomial logit model, and also that school-specific constants in the multinomial logit model control for any school fixed effect on parents' utility

[15] The distributional assumption of ϵ_p does not have any effect on the representation of parent's utility function as a conditional logit model since $\xi_{p,s}$ is not correlated over school alternatives for parent p.

(Carneiro et al., 2016; Bayer and Timmins, 2007).[16]

In the first step, the estimation procedure is similar to standard discrete choice estimation, in which parameters in equation 4.8 are chosen to best fit the observed choices made by parents in the data. The difference with standard discrete choice estimation is that at each iteration of the optimization algorithm, the parent's unobserved characteristic, ϵ_p, is integrated out from equation 4.8. I assume that ϵ_p follows a standard normal distribution.[17] In this step, I estimate parameters that determine utilities from observed ($\delta_{p,s}$) and unobserved ($\psi_{p,s}$) characteristics of the parent (equations 4.6 and 4.7), and also school-specific constants in the multinomial logit model.[18] The constants stand for the parents' mean utility for schools, δ_s.

An important feature of the dataset in this chapter is that this represents a household census of school choice. Therefore, it is possible to equate both observed and simulated school shares of enrolled students to recover school-specific constants or parents' mean utilities, δ_s, by using the contraction mapping procedure proposed in Berry (1994). By doing so, not only the computation burden is reduced, given the 2495 schools I consider in this chapter, but also school-specific constants are accurately estimated since the student-level dataset includes about 11 students per school.[19]

The estimates of parents' preferences for school attributes are sensitive to unobserved school attributes. This means that any school attribute that parents consider such as those resulting from an equilibrium, σ_s, may be correlated with unobserved school attributes, ξ_s. Therefore, in the second step, I confront the fact that school attributes resulting from an equilibrium process, σ_s, are endogenous. This stage uses the estimated vector of school-specific constants, δ_s, school attributes and instruments to estimate all parameters in equation 4.5 that represent parent's mean preferences for school attributes, $\overline{\beta}$, $\overline{\alpha}$ and δ_D, via two stage least squares (2SLS).[20] The estimation of parents' mean preferences for school attributes through 2SLS allows me to estimate unbiased preferences due to unobserved or measurement errors in school attributes, even without making any distributional assumption on

[16]An assumption in the estimation approach is that the probability that an equilibrium is selected is not affected by any individual's particular tastes.

[17]To calculate the mixed logit probability $\overline{Pr}(S = s \mid X_s, \sigma_s, Z_p, \beta, \alpha, \lambda, \gamma)$, I use a 7-point Gaussian quadrature rule. The mixed logit probability is given by: $\overline{Pr}(S = s \mid X_s, \sigma_s, Z_p, \beta, \alpha, \lambda, \gamma) = \int Pr(S = s \mid X_s, \sigma_s, Z_p, \epsilon_p, \beta, \alpha, \lambda, \gamma) f(\epsilon_p)\, d\epsilon_p$, For instance, the parent's preference for exogenous school attributes is: $\beta_{p,j} \sim N\left(\overline{\beta}_j + \sum_{k=1}^{K} z_{p,k}\beta_{k,j}, \beta_j^u\right)$.

[18]The parameters of the model are globally identified under the regularity conditions that are generally required for standard discrete choice. This is because the mixed logit probability, $\overline{Pr}(S = s \mid X_s, \sigma_s, Z_p, \beta, \alpha, \lambda, \gamma)$, is a weighted average of the logit formula evaluated at different values of ϵ_p, with the weights given by density $f(\epsilon_p)$.

[19]In the estimation, I do not calculate derivatives for school-fixed effects.

[20]Parameters in equation 4.5 (mean preferences) are identified under conditions that are generally required for linear regression models. However, note that in equation 4.8, the vector of parameters δ_s is estimated rather than observed in the data, which depend on the distribution of $\xi_{p,s}$.

ξ_s.

I construct two types of instruments. The first set of instruments follows the industrial organization literature and Caetano and Maheshri (2017). The industrial organization literature indicates that the demand for school s may be affected by any quality change of surrounding schools. Therefore, attributes of surrounding schools may be potential instruments for endogenous attributes, σ_s, of school s. In this fashion, I employ the average quality of surrounding schools, $\bar{\sigma}_{-s}$, as an instrument for school s. These instruments are weighted by distance. Additionally, Caetano and Maheshri (2017) take advantage of students' linguistic grade composition for a particular school across time to identify parents' preferences for school attributes. The underlying economic framework is that parents may form their expectations for a school's ethnic sorting from previous periods when making an enrollment decision. This can create a parent's positive-feedback mechanism that may keep or change a school's ethnic composition through time.

In the context of Guatemala, parents may expect a 65 percent of SMT students at school since this percentage represents a uniform sorting of these students across schools. Therefore, I standardized schools' shares of SMT students so that a value of zero represents a uniform sorting of SMT students across schools. Positive values indicate a higher share of SMT students relative to the uniform sorting. Following the industrial organization literature, I construct the instrument $\bar{\sigma}_{-s,smt}$ by averaging the standardized share of SMT students in surrounding schools. Similar to Caetano and Maheshri (2017), I employ $\bar{\sigma}_{-s,smt,2013}$ and $\bar{\sigma}_{-s,smt,2010}$ and their interaction as instruments for the school's share of non-SMT students, $\sigma_{s,nsmt}$. For instance, if $\bar{\sigma}_{-s,smt,2010}$ is positively correlated with the share of non-SMT students, this correlation may imply that households are forming their expectations for schools' ethnic composition from previous periods to sort their child.

Similar to Carneiro et al. (2016) and Bayer and Timmins (2007), the second set of instruments represents the predicted shares of non-SMT parents at school. The assumption is that by using only exogenous school attributes, the model can compress in one index or instrument the attributes of surrounding schools.[21] To do so, and given estimates for parents' heterogeneous preferences (parameters in equation 4.8), first I simulate parents' school choice model with the parameters of the endogenous variables, σ_s, parents' unobserved characteristics, ϵ_p, and school-specific constants, δ_s, equal to zero. The estimated school-specific constants, δ_s, include the effects of unobserved school attributes. Then, I calculate the expected share of non-SMT parents at each school, and I employ this new variable as an instrument to get consistent estimates for the vector of parameters $\overline{\beta}$. Secondly, with the consistent estimates for $\overline{\beta}$ from the previous step, I calculate a mean utility, $\widehat{\delta}_s$, that contains only the effect of exogenous school attributes. Then, I simulate the model with $\widehat{\delta}_s$

[21] Bayer and Timmins (2007) explain that many functional forms can be use to construct instruments. By using the parents' probability of selecting a school, variation in the instrument is determined in part by nonlinearities implied by the assumption about the error distribution, $\xi_{p,s}$. However, variation in the proposed instrument also relies on the interaction of individual characteristics with school attributes.

to construct new instruments and employ 2SLS again.

4.3.3 Parents' priority for their child's non-Spanish language use

In Guatemala, an important school attribute for parents may be whether their child speaks/learns Spanish or not at school, since Spanish is the predominant language of instruction at schools and communication at jobs. This is a school attribute that parents have to infer when considering schools given that friendships are stronger intra-race.[22] For instance, non-SMT parents may prefer schools where the number of non-SMT students is low. Then, parents' expectation for their child's non-Spanish language use at a potential school may be based on school attributes, such as the share of non-SMT speakers. Parents may use this expectation and, concurrently, with other school attributes, when deciding which school to enroll their child.

To control for a parent's expectation for their child's non-Spanish language use in equation 4.4, I first estimate a logit model for students' non-Spanish language use as a function of parents' characteristics and school attributes, and conditional on schools where these students attended. Then, for each school, I employ the logit model to calculate the student's probability of speaking a non-Spanish language, and I use this probability as the parent's expectation.

Specifically, if I observe whether the student i reports speaking their non-Spanish language, $m_i = 1$, I assume that the student's utility is larger or equal to 0. Otherwise, $m_i = 0$ and student's utility is lower that zero. I also assume an additive utility framework, where the characteristics X_s, σ_s and Z_p may affect a student's utility to speak their mother tongue. Let $\sigma_{s,m}$ be the share of non-SMT students who reported non-Spanish language use at school s, $\sigma_{s,-m}$ be all other endogenous school attributes in this chapter, and $X = [X_s, Z_p, \sigma_{s,-m}]$, then utility function of student i, $U_{i,m}$, is given by:

$$U_{i,m} = X\gamma_0 + \sigma_{s,m}\gamma_1 + \sigma_{s,m}X\gamma_2 + \xi_{i,m}. \tag{4.9}$$

The vector of parameters γ_0 and γ_2, and the parameter γ_1 stand for students' preferences for speaking their non-Spanish mother tongue. The vector of parameters γ_2 in equation 4.9 allows me to observe if there exists complementarity between the school's share of non-SMT speakers and other school and household's variables.

I assume that $\xi_{i,m}$ follows an extreme value type one distribution to represent students'

[22] For instance, see Fruehwirth (2013), Hanushek et al. (2009), and Marmaros and Sacerdote (2006).

utility function as a logit model.[23] Therefore, I estimate this logit model for students' non-Spanish language use as a function of parents' characteristics and school attributes. Then, conditional on estimates and parents' characteristics, the parent infers their child's non-Spanish language use at each school by observing school attributes and calculating the child's probability of speaking a non-Spanish language at school:

$$E\left(m_i \mid X_s, \sigma_s, Z_p\right) \quad = Pr\left(m_i = 1 \mid X_s, \sigma_s, Z_p\right). \tag{4.10}$$

I include this parent's expectation for their child's non-Spanish language use, $E\left(m_i \mid X_s, \sigma_s, Z_p\right)$, in the parent's utility function as an additional school attribute when estimating the random coefficient model. Equation 4.6 includes this extra term $\theta E\left(m_i \mid X_s, \sigma_s, Z_p\right)$, where the parameter θ represents the parent's priority for their child's expected non-Spanish language.

4.4 Results

A language barrier can appear if non-SMT students are not surrounded by people proficient in Spanish. Therefore, to find out the underlying mechanism (parents' preferences) that influences students' linguistic sorting across schools, I first discuss students' preferences for speaking their non-Spanish mother tongue (parameters in equation 4.9). Then, I analyze parents' mean preferences for school attributes (parameters in equation 4.5). Lastly, I discuss parents' heterogeneous preferences for school attributes, $\delta_{p,s}$ and $\psi_{p,s}$.

4.4.1 Students' preferences for speaking a non-Spanish language

In this section, I interpret students' preferences for speaking a non-Spanish language (parameters in equation 4.9). The dependent variable in the logistic regression takes a value of one if non-SMT students reported frequent use of their mother tongue and zero otherwise. The independent variables are grouped by either family or school variables. Among family variables, I consider the following variables to influence students' non-Spanish language use: parents' educational attainment, family income, child's gender, whether the student has one or two non-SMT parents, and the students' home proximity to the capital city.

On the other hand, school attributes may heavily affect student's decisions as well. I con-

[23] After controlling for school attributes and family characteristics, I assume that the student's unobserved portion of utility, $\xi_{i,m}$, in equation 4.9 are not correlated between alternatives. This means that any parental influence on a student's language decision is entirely controlled for a parent's characteristics. Therefore, variables $\xi_{i,m}$, $\xi_{p,s}$ and ϵ_p are independent one from each other.

sider as main determinants whether schools have a bilingual program implemented, whether principals speak a language other than Spanish, and the school's share of non-SMT students who reported non-Spanish language use (non-SMT speakers), $\sigma_{s,m}$. In addition to these variables, I employ school quality indices and the average test scores at each school to control for the degree of quality of the school. For example, if high quality schools, as measured by test scores, frequently evaluate students, this may force non-SMT students to rely on their mother tongue to learn a subject if they are not yet proficient in Spanish. Finally, I control for whether the school is located in a rural area and for the school's proximity to the nearest main municipal city.

In short, the results indicate that variables that heavily affect non-SMT students' utility to speak their non-Spanish language, as measured by the coefficients' magnitude, are school test scores, the school's share of non-SMT speakers, $\sigma_{s,m}$, and their interaction. These coefficients are statistically significant even when controlling for family variables. The higher both the school test scores and the share of non-SMT speakers at school, the higher the probability for non-SMT students' mother tongue use. Furthermore, this probability is even higher if both parents have a non-Spanish language as a mother tongue, while the opposite happens with only one non-SMT parent.

Table 4.6 displays students' preferences for speaking a non-Spanish language. First, I discuss the results in column one, which only control for school attributes. Then, relative to the findings in column one, I highlight only the main changes from the second to the last column (when gradually introducing family variables).

In the first column, I only control for school attributes. The results in this column show that school test scores, the school's share of non-SMT speakers, $\sigma_{s,m}$, and their interaction, have a significant effect on students' non-Spanish language use with coefficients of -0.643, 5.842 and 1.061 respectively. These coefficients imply that the higher the school quality non-SMT students attend, the lower Spanish must be spoken at school so that these non-SMT students have a positive utility. Dividing these coefficients by 0.643 will provide students' preferences in terms of test scores' standard deviations. For instance, relative to a school with a school quality of one standard deviation as measured by test scores, non-SMT students would have ten times higher utility when attending a school where Spanish is not spoken than when Spanish is mainly spoken.[24] Other coefficients that statistically affect the students' decisions are school's proximity to the nearest main municipal city, and whether the school is private or not. Interestingly, the rural area dummy variable is not significant once the model incorporates the variable school proximity to the nearest main municipal city.

From the second column to the last one, departmental dummies are included and family variables are gradually added. Based on these columns, the main findings are that, first, the effects of school variables are robust to model specifications. Second, if non-SMT students have non-SMT parents at home, these students would speak their non-Spanish language

[24] $\left(\frac{5.842+1.061}{0.643}\right) = 10.74$

more frequent than non-SMT students with only one SMT parent at home (see fifth column). Third, if non-SMT students attend a high quality school as measured by test scores, non-SMT parents may help them by speaking their mother tongue which deter non-SMT students' Spanish learning process. Lastly, students' home proximity to a main municipal city reduces the probability of speaking a non-Spanish language. This last finding provides support for the distribution of non-SMT students' mother tongue use in Table 4.4. Those families who live in remote areas or far from main municipal cities may not have enough interaction with people proficient in Spanish.

The effects of school test scores, the school's share of non-SMT speakers, $\sigma_{s,m}$, and their interaction on non-SMT students' mother tongue use are relevant in the context of the Guatemalan education system. Mineduc establishes the curriculum for all subjects, and schools are required to cover all of this curriculum. Furthermore, Mineduc requires that students must be promoted to the next grade if their marks are above 60 percent in all classes they take each academic year. How students are tested and the frequency of subject tests, however, are determined completely by principals. Therefore, top ranking schools may demand more effort from students by designing more challenging tests than their counterparts. Then, if non-SMT students are not yet proficient in Spanish, and given the probability of scoring below 60 percent in some subject, then it is likely that non-SMT students devote more time either to studying by themselves or to asking for help using their mother tongue. By doing so, non-SMT students devote less time to learning Spanish than other subjects. As mentioned, non-SMT parents may help students by speaking a language other than Spanish, but they might face the cost of affecting the Spanish learning process. This is a possible interpretation of such coefficients' robustness and magnitude.

In terms of non-SMT students' language barrier, the results also provide some insight. Spatial segregation also influences students' Spanish learning process. First, the closer non-SMT students' home to the capital city, the more likely non-SMT students are going to speak Spanish. This variable, home-to-capital city proximity, is significant and with a coefficient of -0.433 (see sixth column). For instance, a non-SMT student who resides in the capital city would have a utility reduction equivalent to $\frac{0.433}{0.643} = 0.673$ standard deviations in their test score. Furthermore, the results also indicate that non-SMT students talk in Spanish with non-SMT classmates only if non-SMT students attend schools near by a main municipal city.[25]

4.4.2 Parents' mean preferences for school attributes

The international literature about school choice indicates that parents may sort their child into schools where other parents have a similar race, income or educational attainment. In the context of Guatemala, parents may have preferences for the school attributes that come from

[25]The interaction between $\sigma_{s,m}$ and school-municipal city proximity, in column 6, show this result.

an equilibrium process, σ_s. I consider the following school attributes as endogenous: (1) share of non-SMT speakers, $\sigma_{s,m}$, (2) share of non-SMT students, $\sigma_{s,nsmt}$, (3) share of non-SMT parents, $\sigma_{s,nsmt\ households}$, (4) fees, $\sigma_{s,fees}$, (5) average test score, $\sigma_{s,score}$, (6) average family income, $\sigma_{s,family\ income}$, and (7) household's average years of education, $\sigma_{s,education}$. The school's share of non-SMT speakers stands for the share of students who reported frequent use of a non-Spanish language, while the school's share of non-SMT students represents students who do not have Spanish as a mother tongue (regardless of their mother tongue language use).

Table 4.7 displays the parents' mean preferences for school attributes. The dependent variable is the school-specific constants, δ_s, while the independent variables are all school attributes. The upper part of the table shows all school attributes, while the lower part contains the first stage estimation of the 2SLS for the school fees, $\sigma_{s,fees}$, and school's share of non-SMT parents, $\sigma_{s,nsmt\ household}$. Only the first column of this table shows the results for ordinary least squares.

In terms of ethnic segregation at school, the main message of the table is that, on average, parents do not prefer schools where parents have a non-Spanish mother tongue. This finding is significant only in the last column, when including the two types of instruments. This finding may be the root of non-SMT students' language barrier, but this becomes clear in the section 4.4.3, when analyzing the parents' heterogeneous preferences for school attributes.

Other school attributes can shape ethnic segregation at school as well. Recall that Table 4.2 shows that non-SMT parents have low educational attainment, which may lead non-SMT parents to live in poverty. With this in mind, results show that parents look for private schools, and also schools with a bilingual education program. Furthermore, the closer a school's location to the main municipal city, the stronger parents' preferences are for these schools. Therefore, if non-SMT parents cannot pay for school fees or commute to municipal cities, this may lead to ethnic segregation. On the other hand, parents dislike, on average, to pay school fees. This last finding is significant, in the last two model specifications and is in line with the international literature of school choice.

The first step of the 2SLS for the schools' non-SMT parent shares, $\sigma_{s,nsmt\ household}$, also sheds light on how parents sort their child into schools. The coefficients for the instruments $\bar{\sigma}_{-s,smt,2013}$ and $\bar{\sigma}_{-s,smt,2010}$ are negative and significant, while the interaction is positive and significant. For example, if schools that surround school s have high shares of SMT students, the results indicate that school s will be selected by non-SMT parents. This finding shows that parents may be forming their expectations for school ethnic composition from previous periods when making an enrollment decision (Caetano and Maheshri, 2017).

To conclude this section and in terms of the root of students' language barrier, results indicate that parents dislike schools where the other parents have a non-Spanish mother tongue. Furthermore, as interpreted by the first stage in the 2SLS, parents form their

expectations for students' linguistic sorting across schools from previous periods when making an enrollment decision.

4.4.3 Parents' heterogeneous preferences for school attributes

In this section, I discuss a parent's heterogeneous preferences for school attributes (parameters in equation 4.8). I assume that the main household's characteristics influencing parent's preferences for school attributes are their mother tongue, years of education, family income, child's gender, and home-to-capital city proximity. I interact these household variables with exogenous and endogenous school attributes X_s and σ_s, respectively, but I focus the discussion on variables that can affect the students' linguistic sorting across schools. These school attributes are (1) the share of non-SMT speakers, $\sigma_{s,m}$, (2) the share of non-SMT students, $\sigma_{s,nsmt}$, (3) the share of non-SMT parents, $\sigma_{s,nsmt\ household}$, and (4) student's non-Spanish language use, $E\left(m_i \mid X_s, \sigma_s, Z_p\right)$.[26] These first three school attributes are in line with the international literature where parents select schools based on language or race. The last variable follows the peer effect literature in which parents' decision of enrollment may depend on their child's probability to speak/learn Spanish given that peer effects are stronger intra-race. By focusing the discussion on these four variables, I help to identify the root of non-SMT students' language barrier, since linguistic segregation across schools is partially to blame for non-SMT students' lack of Spanish comprehension (see Chapter 3).

In short, four main findings emerge in this section. First, non-SMT parents sort their child into schools where parents have a similar mother tongue (polarization effect) and also where students frequently speak their non-Spanish language. These preferences shape students' linguistic sorting across schools. Second, this polarization effect is stronger the further away parents reside from the capital city. Third, non-SMT parents do not have preferences, statistically speaking, for schools where students have a non-Spanish mother tongue. Last, non-SMT parents prioritize their child's Spanish learning process, and this preference is slightly higher for households with only one SMT parent at home.

Table 4.8 shows parents' preferences for school attributes that may shape schools' ethnolinguistic composition. Columns in this table differ from each other by controls for spatial segregation of groups or for parents' priority for their child's Spanish learning process, $E\left(m_i \mid X_s, \sigma_s, Z_p\right)$.[27] Only for the first column, the benchmark specification, I discuss the effects of school's shares of non-SMT speakers, students and parents on parents' utility. Then, depending on the column, I highlight how spatial segregation, parents' priority for their child's Spanish learning process and parents' unobserved characteristics affect parents'

[26]The school's share of non-SMT speakers stands for the share of students who reported frequent use of a non-Spanish language, while the school's share of non-SMT students represents students who do not have Spanish as a mother tongue (regardless of their mother tongue language use).

[27]I refer to spatial segregation as both school-to-nearest municipal city proximity and home-to-capital city proximity.

preferences relative to the benchmark specification.

The first column does not include the effects of both spatial segregation and parents' priority for their child's Spanish learning process, $E\left(m_i \mid X_s, \sigma_s, Z_p\right)$. This is the benchmark specification. This specification allows me to document whether or not Guatemala shows the effect of a polarized population as argued in Montalvo and Reynal-Querol (2005b). If so, non-SMT parents may sort their child into schools where parents or students have a similar mother tongue.

Results in column one indicate that households with just one non-SMT parent does not prefer schools where students are non-SMT speakers, $\sigma_{s,m}$. Furthermore, depending on households' own characteristics such as parents' years of education, parents may have even lower preferences for this school attribute, $\sigma_{s,m}$. For the second school attribute in this table, school's share of non-SMT students ($\sigma_{s,nsmt}$), households with only one non-SMT parent dislike this school attribute.

The third school attribute in the table, school's share of non-SMT parents, may be the root of students' language barrier. On average, parents' dislike for schools where other parents have a similar mother tongue is 3.331, but statistically insignificant, in monetary terms is equivalent to Q179.96, \$23.22 USD, or about 1.75 months of tuition at school.[28] However, non-SMT parents have a strong preference for their own mother tongue. The coefficients or preferences are significant and with a magnitude of 7.010 and 8.631 for one and two non-SMT parents at home, respectively, or in monetary terms of Q378.71 (\$48.87 USD) and Q466.29 (\$60.17 USD), respectively. Then, the non-SMT parents' total preference for a school where all parents are non-SMT speakers is 8.631-3.331=5.3 or equivalent to Q286.33 or \$36.95 USD, respectively. Therefore, non-SMT parents will sort their child into schools where other parents have a similar mother tongue. The finding that non-SMT parents show preferences for the school's share of non-SMT parents, $\sigma_{s,nsmt\ household}$, provide evidence of the effects of polarization in Guatemala.

In the second column in Table 4.8, I observe whether students' linguistic sorting across schools is the result of spatial segregation of groups. Desmet et al. (2018) argue that Guatemala is a country where groups have a low degree of interaction, and that such behavior may be the result of spatial segregation of groups.

The main change relative to the benchmark specification is that the parental mean preference for school's shares of non-SMT speakers becomes insignificant. Parent's mean preferences for the school's share of non-SMT parents remains as the only significant preference. This finding provides two noteworthy points. First, the probable root of students' language barrier is not spatial segregation of groups, but the result of parents' preferences for being with other parents who speak the same language. The parents' mean and heterogeneous

[28] $\left(\frac{3.331}{1.851}\right)\left(\frac{100}{7.75}\right) = 23.22$ with an exchange rate of Q7.75/USD. Recall from Section 4.2 that the average tuition at school is Q103 per month.

preferences for school's share of non-SMT parents are both significant, but with different signs. This sign difference may be due to the effect of a polarized population. Second, the low degree of interaction among groups is not completely the result of spatial segregation as hypothesized by Desmet et al. (2018). The results indicate that both home-to-capital city proximity and school-to-nearest municipal city proximity only reduces this polarization effect on parents' utility from 5.3 to 4.104, or in monetary terms from Q286.33 to Q221.72 (from $36.95 USD to $28.61 USD).

In the third column, I control for students' expected non-Spanish language use at school, $E(m_i \mid X_s, \sigma_s, Z_p)$. One important change happens relative to the benchmark specification. Non-SMT households also sort their child into schools where students are non-SMT speakers, $\sigma_{s,m}$. However, these non-SMT household preferences decrease for parents with high educational attainment. Furthermore, the sorting of non-SMT parents into schools where parents have a similar mother tongue is robust to students' expected use of a non-Spanish language at school, $E(m_i \mid X_s, \sigma_s, Z_p)$.

The lower part of column three shows the parent's preferences for their child's non-Spanish language use. On average, parents do care about non-Spanish language use. This preference is positive and significant with a coefficient of 3.158 or in monetary terms of Q170.61 or $22.01 USD. On the other hand, non-SMT parents want their child to speak Spanish. For example, if non-SMT parents consider a school where their child may speak a non-Spanish language with probability one, these parents would face a significant utility reduction of -1.984 for this school or equivalent to Q107.19 or $13.83 USD.

In the fourth column, I control for both effects: spatial segregation and students' expected use of a non-Spanish language at school. The results still show the effect of a polarized Guatemalan population. Furthermore, the results indicate that non-SMT parents prefer schools where students still speak a non-Spanish language, but non-SMT parents want their child to speak Spanish. A possible interpretation for this finding is school thresholds for grade progression which students face in the Guatemalan education system as mentioned in section 4.4.1. Non-SMT parents sort their child into schools where students speak a non-Spanish language to compensate for their child' lack of Spanish proficiency and, therefore, to increase their child's probability of grade progression. However, non-SMT students may not improve their Spanish comprehension.

Lastly, the fifth column shows the results when integrating out parents' unobserved characteristics, ϵ_p, in the parents' utility function. Parents' unobserved characteristics may also control for measurement error in variables (at household level) given the data limitations in this chapter.

Two important findings emerge. First, even after controlling for parents' unobserved characteristics, ϵ_p, in the parents' utility function, the results still indicate that Guatemala shows the effect of a polarized population. Parents' average preference for schools where

parents have a similar mother tongue is significant and negative, -4.751, or in monetary terms is about Q256.67 or $33.12 USD. However, if the household has one or two non-SMT parents, these parents would have a total preference of 2.364 and of 4.021, respectively, when the school is selected only by non-SMT parents. In monetary terms this preference is about Q127.71 ($16.48 USD) and Q217.23 ($28.03 USD) respectively. In the section 4.5, I provide an interpretation for this finding, as well as how this polarization effect may be shaping school linguistic composition.

Second, parents' heterogeneous preferences for school's shares of non-SMT parents depend on how close non-SMT parents live to the capital city (home-to-capital city proximity). For instance, if non-SMT parents reside in the capital city, their utility for schools where all parents are non-SMT speakers is lower than zero which means that parents may prefer other schools. In other words, the further way non-SMT parents reside from the capital city, the higher non-SMT parents' preferences to sort their child into a school highly populated by non-SMT parents, which lead to spatial and linguistic segregation at school.

The fact that non-SMT parents prefer to sort their child into schools where parents or students have a similar mother tongue is affecting non-SMT students' interaction with people proficient in Spanish. Recall that Chapter 3 shows that schools' linguistic composition is partially to blame for non-SMT students' lack of Spanish comprehension, which affects their performance at school. Therefore, spatial segregation of groups is not a key factor to understand non-SMT students' lack of Spanish proficiency.

Another variable that can shape the ethnolinguistic segregation at school is home-to-school proximity, which is also significant in this chapter. Results in Table 4.9 suggest that the distance from students' home to school reinforces non-SMT households' spatial segregation, which influences parents' school choice. Nevertheless, spatial segregation is not the main factor influencing students' Spanish learning process as implied by the fact that non-SMT parents sort their child into schools where parents have a similar mother tongue, even after controlling for home-to-school proximity.

Regarding the parent's preferences for exogenous school attributes, X_s, I just provide the main findings here (see Table 4.10 for estimates). Parents prefer schools that are located close to main municipal cities, that are private, and that have a good infrastructure. Furthermore, if parents reside close to the capital city, they have strong preferences for schools located close to main municipal cities. A similar pattern is found in terms of the school quality as measure by the index of classes. The closer parents reside to the capital city, the higher parents' preferences for schools that offer more classes per week, for instance.

To conclude this section, the main finding is that non-SMT households prefer to sort their child into schools where parents have a similar mother tongue. This non-SMT parents' preference may be the root of the non-SMT students' language barrier. Non-SMT parents' self-selection by mother tongue is robust to the inclusion of variables that the literature on

school choice find to be determinant factors. This self-selection is robust to the addition of school fees, home-to-school proximity, and spatial segregation.

4.5 Non-SMT parents' demand for schools

The main finding of this chapter is that non-SMT parents' preference for their own mother tongue is creating ethnolinguistic segregation at school. As shown in Chapter 3, ethnolinguistic segregation at school puts non-SMT students in a disadvantaged position in the Guatemalan education system.

Therefore, in this section, I provide some policy recommendations by first identifying the Guatemalan departments that display strong polarization effects. To do so, I calculate the average departmental elasticities of non-SMT parents' demand for schools highly populated by non-SMT parents. Second, I shed light on how to reduce ethnolinguistic segregation at rural schools by recommending how to improve the quality of rural schools. By doing so, SMT parents who prefer urban schools may consider their closest rural school as an option. Lastly, non-SMT households' preferences for their own mother tongue must come from an underlying mechanism. The identification of this mechanism may be the key factor to eliminating non-SMT students' language barrier.

Table 4.11 displays the average departmental elasticities of households with non-SMT parents. The first three columns show the elasticities by school type (municipal-cooperative, public or private) while the last column shows the average departmental elasticity. Departments with strong polarization effects are located in the north-west region of Guatemala: Sololá, Alta Verapaz, Quiché, Totonicapán, Huehuetenango and Chimaltenango. These departments show elasticities higher than elasticities without parental preferences to sort students by mother tongue as shown next. The table also indicates that non-SMT parents who reside in these departments are more likely than SMT parents to choose a school where parents are also non-SMT speakers if the school is not private.

Figure 4.1 shows the effect of a polarized population. This figure depicts the last column of Table 4.11 together with the elasticities of households with a non-SMT parent. The magnitude of these elasticities clearly differs between non-SMT households and among departments. Departments located in the north-west region of Guatemala show a strong effect of ethnic polarization. For example, in the department of Sololá, on average, non-SMT parents' probability of selecting a school increases by 0.6 percent given a 1 percent increase in the proportion of non-SMT parents at school. However, the same elasticity for households with only one SMT parent is about 0.41. On average, the non-SMT households' elasticity difference in the north-west region is about 0.2, while in other departments it is nearly zero. In a non-polarized country these elasticities should be close to zero.

Figure 2 shows an alternative scenario where non-SMT parents do not have preferences for schools highly populated by non-SMT parents (no polarization effect). Both parents' mean and specific preferences for schools where parents have a similar mother tongue are set to zero.[29] This figure shows that the polarization effect (shadow area in the figure) is an important factor influencing non-SMT household's demand elasticities. In the case of Sololá, the elasticity of a non-SMT parent who has the polarization effect is 0.6, otherwise the elasticity is 0.2. The difference in elasticities implies that, on average, the polarization effect accounts for 66 percent of the non-SMT parent's demand elasticity for schools in north-west departments.

To reduce non-SMT students' language barrier in the north-west region of Guatemala, policy makers can implement Spanish as a second language program, alter schools' linguistic composition or both.[30] Different than a second language program, changing school attributes in public rural schools not only may lead non-SMT students to improve their educational achievement in schools, but also SMT parents can consider enrolling their children in those rural schools.

To change school linguistic composition, first, I show evidence about SMT parents' preferences and their willingness to pay for school attributes that principals or Mineduc can alter. This evidence is for parents who prefer to enroll their child in urban schools. Second, I provide measures to improve the quality of rural schools, so that SMT parents who prefer urban schools may consider their closest rural school as an option.

The average of SMT parents' marginal utility and their willingness to pay for school attributes is displayed in Table 4.12.[31] The first four columns show the average of SMT parents' marginal utilities by distance from home to the capital city. The fifth column shows the school attribute increment to interpret the SMT parents' willingness to pay for school attributes. This willingness to pay is also split by distance from home to the capital city. The table's marginal utility section indicates that school's average test scores, infrastructure, and courses are the school attributes that SMT parents value the most, in order of preference respectively. Furthermore, the closer the school is to the capital city, the higher the SMT

[29]Recall for section 4.3 that a parent's preference for the school attribute l is given by $\alpha_{p,l} = \overline{\alpha}_l + \sum_{k=1}^{K} z_{p,k}\alpha_{k,l} + \alpha_l^u \epsilon_p$. Then, I set equal to zero $\overline{\alpha}_l$ and $\alpha_{k,l}$, where l stands for the share of non-SMT parents at school, and k stands for whether or not the household has two non-SMT parents at home.

[30]For instance, in Chapter 3 the observed linguistic sorting in Guatemalan secondary schools is altered. This artificial setting can be understood as a Spanish second language program. Chapter 3 shows that non-SMT students perform better on national tests. However, SMT students still outperform non-SMT students.

[31]To calculate parents' willingness to pay for school attributes, I first calculate the total derivative of the parent's utility function, $\partial U_{p,s} = \partial X_{s,j}\beta_{p,j} + \sum_{j=1}^{J} X_{s,j}\beta_{j,k}\partial Z_{p,k}$. Then, the willingness to pay is given by $\frac{\partial Z_{p,k}}{\partial X_{s,j}} = \frac{\beta_{p,j}}{\sum_{j=1}^{J} X_{s,j}\beta_{j,k}}$; where $Z_{p,k}$ stands for family income and $X_{s,j}$ any school attribute.

parents' valuation for school attributes. The first row of the table, test score, shows that for an increment of 0.1 standard deviation in test scores, SMT parents are willing to pay Q208.10 ($26.85 USD) per month if the school is practically located in the capital city. Similarly, SMT parents who live further away than 200 kms from the capital city are willing to pay Q148.07 ($19.11 USD) per month. In monetary terms, the second highest attribute SMT parents are willing to pay for is school infrastructure.

Having identified what variables SMT parents who enroll their children in urban areas value the most, Table 4.13 shows the SMT parents' monetary and school quality compensation for considering the closest rural school from their child's urban school enrollment. Rows in this table display the departments with the strongest polarization effects. The first column shows the difference in SMT parents' utilities for considering their child's urban school enrollment relative to the closest rural school. For example, the highest difference in terms of parents' utilities is Sololá with a value of 6.674. This column indicates that, on average, schools in Sololá, Quiché and Chimaltenango are of lower quality relative to urban schools. The second column shows the monetary compensation per month that an SMT parent should receive to get the same level of utility relative to their child's school of enrollment. On average, an SMT parent in Sololá should receive Q477.66 ($61.63 USD) per month.

Although, the implementation of a cash transfer program for attending rural schools may change the linguistic distribution across schools, the non-SMT students' school performance improvement would be only through both an improvement in Spanish comprehension and student peer effects. However, the same result may be achieved by improving rural school quality such as infrastructure or courses offered. By changing rural public school quality, SMT parents may start considering this school as an attractive option. The last two columns of the table provide some guidance on how much to improve rural school attributes relative to a school located in the capital city with an average quality, so that SMT parents receive, on average, the same utility relative to their child's urban school enrollment. For the case of Sololá this represents an infrastructure quality improvement of around 4.06 times greater than an average school in the capital city and a course offering list that is 30.99 times greater than an average school in the capital city.

The non-SMT households' preferences for their own mother tongue must come from an underlying mechanism. The identification of this mechanism may be the key factor to eliminate non-SMT students' language barrier. Chamarbagwala and Morán (2011) discuss that after Guatemala gained its independence from Spain in 1821, an authoritarian state was created that excluded the indigenous population, was racist in its precepts and practices, and served to protect the economic interests of the privileged minority. Due to the chronic status quo of inequality and social exclusion, the Guatemalan civil war started. The Guatemalan civil war lasted for 36 years, and the peace agreement was signed in 1996. Indigenous people, especially the Mayans, suffered the consequences of the civil war. The authors define the departments of Quiché, Alta Verapaz, Baja Verapaz, Petén and Huehuetenango as those

that were most affected by the civil war.

The discussion in Chamarbagwala and Morán (2011) suggests that non-SMT parents' preferences for their own mother tongue may have originated in or been reinforced by the civil war. Furthermore, The United Nations Development Program's (UNDP) report in 2005 explains that in Guatemala there exists discrimination that mainly affects the indigenous population. Since culture is comprised of a group of people's beliefs and values which are passed down through generations, prejudices that existed before the civil war continue to exist in the Guatemalan culture. In addition, the genocide perpetuated against non-SMT indigenous people during the civil war may have severely reinforced non-SMT parents' preferences for their children attending school with other non-SMT children. As a result of such behavior, the ethnic discrimination or segregation can limit human development (UNDP, 2005).

To help understand whether this past behavior in the civil war influences non-SMT parents demand for schools where parents have a similar mother tongue, I regress the difference between non-SMT households' elasticities (shadow area in Figure 1) on the number of victims of the Guatemalan civil war per 1000 population in departments.[32] I refer to this difference as the polarization effect. Additionally, I control for the population's illiteracy rates in 1973, and for a fixed effect for the departments where the civil war affected the most. These are the variables Chamarbagwala and Morán (2011) employ in their work. Furthermore, since the non-SMT parents' preferences may be the result of departmental inequality, I employ both a departmental Gini index calculated in 2011 by the Instituto Nacional de Estadística de Guatemala (Guatemala's National Statistic Institute) and a departmental poverty index calculated by UNDP.[33]

Table 4.14 shows the estimates for the effects of civil war victims on non-SMT parents' polarization effects. First, notice that only 22 observations are included which correspond to the 22 Guatemalan departments. In some columns the Guatemalan department is not included because the Gini index was not calculated for this department. Across model specifications, the coefficient on the number of victims in the civil war is positive and significant, even when controlling for the fixed effect for the departments where the civil war was most intense. To help interpret the effect of the number of victims on the polarization effect, I compare the predicted polarization effect between the high and low intensity areas of the civil war, $0.097\left(E\left(ln\left(victims\right) \mid intensive\ area\right) - E\left(ln\left(victims\right) \mid not\ intensive\ area\right)\right)$. I divide this predicted difference by the observed polarization effect between these two areas, $E\left(polarization\ effect \mid intensive\ area\right) - E\left(polarization\ effect \mid no\ intensive\ area\right)$. This ratio indicates that about 30.786 percent of the polarization effect between these two areas is due to differences in the number of victims of the Guatemalan civil war.

[32] Approximately, the youngest parents in this research lived through at least 12 years of the civil war, while the oldest parents lived through all of it. As part of the civil war peace agreement, only aggregate data of victims can be released.

[33] According to this institute, the Gini index is calculated using data from rural areas of Guatemala.

If policy makers can change school attributes in rural schools or eliminate non-SMT parents' polarization effect, then the subsequent change in linguistic sorting of students in these schools would affect the non-SMT students' Spanish proficiency, school performance, and in the long term the non-SMT population's education and poverty distribution.

4.6 Conclusion

The existence of a learning barrier to a country's predominant language may affect human capital development and, as a result, influence poverty in the long run. The first sign of this language barrier may be reflected in students' school performance. In Guatemala, the focus of this study, non-SMT students' performance at school is poor relative to SMT students. Although factors such as parents' educational attainment and family income can affect the accumulation of students' human capital, non-SMT students' poor performance may also be resulting from a language barrier.

To find the root of the non-SMT students' language barrier, it is necessary to understand what parents consider important when making enrollment decisions. This is relevant, because parents' decisions about which schools to send their child may play an important role in the determination of friendships. Friendships, which play an important role in learning a new language, are stronger intra-race as the peer effect literature shows (Fruehwirth, 2013; Hanushek et al., 2009; Mayer and Puller, 2008; Marmaros and Sacerdote, 2006; Hoxby, 2000).

A rich dataset of school and households' characteristics allows me to analyze Guatemalan parents' school choices. The estimation of a random coefficient model of demand for junior high schools allows me to control for unobserved school and households' characteristics, and also to observe whether parents value the fact that other parents with similar characteristics select the same school. By doing so, I find that the possible root of language barrier for non-SMT students' lack of Spanish proficiency is that non-SMT parents prefer schools where other parents have a non-Spanish language as a mother tongue; this creates segregation. This finding is robust even after controlling for school attributes that the international literature shows as important factors in school choice and for the non-SMT population's spatial distribution.

The demand estimates highlight that non-SMT parents' segregation by mother tongue at school is high in five Guatemalan departments. This non-SMT parents' segregation by mother tongue seems to be driven by differences in the number of victims between the high and low intensity areas of the civil war. For instance, in the department with the highest segregation, non-SMT parents' probability of selecting a school will increase by 0.6 percent if the proportion of non-SMT parents at this school increases by 1 percent. However, if such preferences did not exist, the probability of selecting this school would increase only by 0.2.

The analysis indicates that measures could be taken to reduce student linguistic segregation and thus lower the intergenerational language barrier for the non-SMT population in Guatemala. The parents' preferences for school attributes suggest that by improving rural school infrastructure and courses offered in rural schools, SMT parents who prefer urban schools would start considering their closest rural school instead. Specifically, after a quality improvement in rural schools, only those SMT parents who benefit from the quality improvement would make an enrollment decision in rural schools. Therefore, by integrating SMT and non-SMT students in schools, the language barrier's consequences on the non-SMT population's educational attainment can be mitigated.

Bibliography

Alberto Alesina, Stelios Michalopoulos, and Elias Papaioannou. Ethnic Inequality. *Journal of Political Economy*, 124(2):428–488, March 2016. ISSN 0022-3808. doi: 10.1086/685300. URL https://doi.org/10.1086/685300.

Patrick Bayer and Christopher Timmins. Estimating Equilibrium Models Of Sorting Across Locations. *Economic Journal*, 117(518):353–374, 2007. URL https://ideas.repec.org/a/ecj/econjl/v117y2007i518p353-374.html.

Steven T. Berry. Estimating Discrete-Choice Models of Product Differentiation. *The RAND Journal of Economics*, 25(2):242–262, 1994. ISSN 07416261. doi: 10.2307/2555829. URL http://www.jstor.org/stable/2555829.

Simon Burgess, Ellen Greaves, Anna Vignoles, and Deborah Wilson. What Parents Want: School Preferences and School Choice. *Economic Journal*, 125(587):1262–1289, 2015. URL https://EconPapers.repec.org/RePEc:wly:econjl:v:125:y:2015:i:587:p:1262-1289.

Gregorio Caetano and Vikram Maheshri. School segregation and the identification of tipping behavior. *Journal of Public Economics*, 148:115–135, April 2017. ISSN 0047-2727. doi: 10.1016/j.jpubeco.2017.02.009. URL http://www.sciencedirect.com/science/article/pii/S0047272717300221.

Pedro Carneiro, Jishnu Das, and Hugo Reis. The value of private schools: evidence from Pakistan. Technical report, Centre for Microdata Methods and Practice, Institute for Fiscal Studies, 2016. URL https://ideas.repec.org/p/ifs/cemmap/22-16.html.

Rubiana Chamarbagwala and Hilcías E. Morán. The human capital consequences of civil war: Evidence from Guatemala. *Journal of Development Economics*, 94(1):41–61, January 2011. ISSN 0304-3878. doi: 10.1016/j.jdeveco.2010.01.005. URL http://www.sciencedirect.com/science/article/pii/S0304387810000076.

Klaus Desmet, Joseph Gomes, and Ignacio Ortuño-Ortín. The Geography of Linguistic Diversity and the Provision of Public Goods. Technical Report w24694, National Bureau of Economic Research, Cambridge, MA, June 2018. URL http://www.nber.org/papers/w24694.pdf.

Kjell I. Enge and Ray Chesterfield. Bilingual education and student performance in Guatemala. *World Bank's Education Sector Review: Priorities and Strategies for Education*, 16(3):291–302, July 1996. ISSN 0738-0593. doi: 10.1016/0738-0593(95)00038-0. URL http://www.sciencedirect.com/science/article/pii/0738059395000380.

Jane Cooley Fruehwirth. Identifying peer achievement spillovers: Implications for desegregation and the achievement gap. *Quantitative Economics*, 4(1):85–124, 2013. ISSN 1759-7331. doi: 10.3982/QE93. URL http://dx.doi.org/10.3982/QE93.

Eric A. Hanushek and Ludger Woessmann. Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth*, 17(4):267–321, December 2012a. ISSN 1573-7020. doi: 10.1007/s10887-012-9081-x. URL https://doi.org/10.1007/s10887-012-9081-x.

Eric A. Hanushek and Ludger Woessmann. Schooling, educational achievement, and the Latin American growth puzzle. *Journal of Development Economics*, 99(2):497–512, November 2012b. ISSN 0304-3878. doi: 10.1016/j.jdeveco.2012.06.004. URL http://www.sciencedirect.com/science/article/pii/S0304387812000491.

Eric A. Hanushek, John F. Kain, and Steven G. Rivkin. New Evidence about Brown v. Board of Education: The Complex Effects of School Racial Composition on Achievement. *Journal of Labor Economics*, 27(3):349–383, 2009. URL https://ideas.repec.org/a/ucp/jlabec/v27y2009i3p349-383.html.

Caroline Hoxby. Peer Effects in the Classroom: Learning from Gender and Race Variation. *National Bureau of Economic Research Working Paper Series*, No. 7867, 2000. doi: 10.3386/w7867. URL http://www.nber.org/papers/w7867. featured in NBER digest on 2001-04-01.

David Marmaros and Bruce Sacerdote. How Do Friendships Form?*. *The Quarterly Journal of Economics*, 121(1):79–119, February 2006. ISSN 0033-5533. doi: 10.1093/qje/121.1.79. URL http://dx.doi.org/10.1093/qje/121.1.79.

Jeffery H. Marshall. School quality and learning gains in rural Guatemala. *Economics of Education Review*, 28(2):207–216, April 2009. ISSN 0272-7757. doi: 10.1016/j.econedurev.2007.10.009. URL http://www.sciencedirect.com/science/article/pii/S0272775708000745.

Adalbert Mayer and Steven L. Puller. The old boy (and girl) network: Social network formation on university campuses. *Journal of Public Economics*, 92(1):329–347, February 2008. ISSN 0047-2727. doi: 10.1016/j.jpubeco.2007.09.001. URL http://www.sciencedirect.com/science/article/pii/S0047272707001181.

Jose G. Montalvo and Marta Reynal-Querol. Ethnic diversity and economic development. *Journal of Development Economics*, 76(2):293–323, April 2005a. ISSN 0304-3878. doi: 10.1016/j.jdeveco.2004.01.002. URL http://www.sciencedirect.com/science/article/pii/S0304387804001129.

José G. Montalvo and Marta Reynal-Querol. Ethnic Polarization, Potential Conflict, and Civil Wars. *American Economic Review*, 95(3):796–816, 2005b. doi: 10.1257/0002828054201468. URL http://www.aeaweb.org/articles?id=10.1257/0002828054201468.

José Garcia Montalvo and Marta Reynal-Querol. Why ethnic fractionalization? Polarization, ethnic conflict and growth. Technical report, Department of Economics and Business,

Universitat Pompeu Fabra, 2002. URL https://ideas.repec.org/p/upf/upfgen/660.html.

Harry Anthony Patrinos and Eduardo Velez. Costs and benefits of bilingual education in Guatemala: A partial analysis. *International Journal of Educational Development*, 29(6): 594–598, November 2009. ISSN 0738-0593. doi: 10.1016/j.ijedudev.2009.02.001. URL http://www.sciencedirect.com/science/article/pii/S0738059309000182.

F.E. Rubio. Educación Bilingüe en Guatemala: Situación y desafios. 2004.

UNDP, editor. *Diversidad étnico-cultural: la ciudadanía en un Estado plural*. Number 7 in Informe nacional de desarrollo humano. Programa de las Naciones Unidas para el Desarrollo, Guatemala, 1. ed edition, 2005. ISBN 978-99939-69-77-8. URL http://desarrollohumano.org.gt/biblioteca/informes-nacionales/. OCLC: 254910041.

4.7 Tables and figures

Table 4.1: Scores on standardized tests by Spanish vs. non-Spanish mother tongue students

	Math			Reading		
	Mother tongue			Mother tongue		
	Non-Spanish (a)	Spanish (b)	Gap (a-b)	Non-Spanish (a)	Spanish (b)	Gap (a-b)
Last grade in						
Elementary school	−0.19	0.12	−0.31***	−0.40	−0.24	−0.64***
Junior high school	−0.25	0.14	−0.39***	−0.40	−0.22	−0.62***
High school	−0.32	0.11	−0.43***	−0.49	−0.17	−0.66***
Average			−0.38			−0.64

Note: Spanish mother tongue (SMT) students' self-report of having Spanish as a first language. Non-SMT students speak one of the Mayan laguages, or Xinka or Garífuna languages as a first language. Test scores have a mean of zero and standard deviation of one at each grade. Significant levels: *** at 1%, ** at 5%, and * at 10%.
Source: Author's calculation using the Ministry of Education's 2013 data.

Table 4.2: Percentage of parents with children at school in 2002 by education level and mother tongue

	Fathers		Mothers	
	SMT speaker	Non-SMT speaker	SMT speaker	Non-SMT speaker
Parental educational attainment				
Lower than elementary	21.1	43.0	30.6	69.2
Elementary (grades 1-3)	25.3	28.8	24.0	17.6
Elementary (grades 4-6)	28.5	21.4	24.1	10.4
Junior high	9.8	3.5	8.0	1.4
High school	10.2	2.7	10.0	1.2
More than high school	4.9	0.6	3.3	0.2
	100%	100%	100%	100%

Note: SMT speaker stands for people who report Spanish as a first language. Non-SMT students speak one of the Mayan languages, or Xinka or Garífuna languages as a first language. The 2002 Guatemalan census shows the education levels of SMT and non-SMT parents with a 2-year-old child in 2002. This distribution may potentially represent the parents' educational distribution of junior high students at school in 2013.
Source: Guatemala's 2002 national census.

Table 4.3: Distance in Kms. from home to school for students who live in rural areas

	Family income's quintile distribution				
	First	Second	Third	Fourth	Fifth
Home to capital city's proximity distribution					
<30 Kms	3.05	2.09	2.95	3.50	4.46
	(8.40)	(2.95)	(8.39)	(5.51)	(11.41)
30-85 Kms	3.13	5.38	5.27	6.17	5.28
	(5.83)	(13.20)	(14.30)	(15.64)	(13.00)
85-125 Kms	4.14	3.94	4.80	6.20	6.52
	(10.64)	(8.05)	(14.91)	(19.06)	(20.38)
>125 Kms	6.98	7.20	5.82	9.56	10.74
	(12.00)	(18.98)	(19.53)	(32.43)	(32.54)

Note: Standard deviation in parenthesis. The first quintile represents the lower family income.
Source: Author's calculation using the Ministry of Education's 2013 data.

Table 4.4: Distribution of SMT students across municipalities and non-SMT students' mother tongue use[(a)]

		Non-SMT students who reported mother tongue use at	
Municipal distribution	Percentage of SMT students	Junior high school	High school
18%	0-25%	81%	75%
14%	25-50%	54%	52%
20%	50-75%	30%	31%
48%	75-100%	17%	13%

(a) Way of reading the table: in 18% of municipalities (1) between 0 and 25% of SMT students reside, and (2) 81% of non-SMT students speaks their mother tongue at junior high school.
Note: Spanish mother tongue (SMT) stands for students' self-report of having Spanish as a first language. Non-SMT students speak one of the Mayan laguages, or Xinka or Garífuna languages as a first language. Guatemala is geographically divided by more than 300 municipalities.
Source: Author's calculation using the Ministry of Education's 2013 data.

Table 4.5: Statistics for household and school's variables

	(1) Mean	(2) Std.	(3) Min.	(4) Max
Household variables				
Non-SMT Parent	0.052		0	1
Non-SMT Parents	0.201		0	1
Student' gender (Male)	0.521		0	1
Mother's educational attainment				
Elementary	0.486		0	1
Junior high	0.107		0	1
High school	0.094		0	1
Undergraduate or more	0.047		0	1
Father's educational attainment				
Elementary	0.490		0	1
Junior high	0.134		0	1
High school	0.112		0	1
Undergraduate or more	0.064		0	1
Family Income	−0.572	0.760	−1.943	4.621
Home-to-Capital city proximity	0.046	0.096	0.003	1
Total households	27948			
School attributes				
Exogenous				
Private	0.325		0	1
Bilingual program	0.019		0	1
Principal's non-SMT speaker	0.218		0	1
Rural area	0.481		0	1
School-to-nearest municipal city proximity	0.241	0.369	0.008	1
Quality indices				
Teachers	−0.537	1.865	−5.923	3.315
Courses	−0.993	1.730	−3.983	0.652
Classes	−0.844	2.047	−4.022	4.523
Discipline	−0.281	1.892	−10.245	2.578
Infrastructure	−0.968	1.582	−8.851	8.685
Distractor	0.411	1.582	−1.283	8.359
Endogenous				
Fees in Q ($\sigma_{s,fees}$)	103.203	1.956	30.560	358.810
Test scores ($\sigma_{s,scores}$)	−0.086	0.738	−3.946	3.503
Years of education ($\sigma_{s,Education}$)	7.296	2.415	0	18.141
Family income ($\sigma_{s,Family\ Income}$)	−0.650	0.531	−1.693	1.938
Share of non-SMT speakers ($\sigma_{s,m}$)	0.246	0.192	0	1
Share of non-SMT students ($\sigma_{s,nsmt}$)	0.384	0.284	0	1
Share of non-SMT parents ($\sigma_{s,nsmt\ household}$)	0.263	0.317	0	1
Total schools	2495			

Note: $\sigma_{s,m}$ represents the school's share of non-Spanish mother tongue students who reported mother tongue use. $\sigma_{s,nsmt}$ represents the school's share of non-Spanish mother tongue students. $\sigma_{s,nsmt\ household}$ represents the school's share of non-Spanish mother tongue parents. $\sigma_{s,fees}$ stands for school's fees. While $\sigma_{s,score}$, $\sigma_{s,family\ income}$ and $\sigma_{s,education}$ stand for the school's average of students' test scores, parents' family income and parents' years of education respectively.
Source: Author's calculation using the Ministry of Education's 2013 data.

Table 4.6: Results of the logistic regression for students' non-Spanish language use

	(1)	(2)	(3)	(4)	(5)	(6)
School attributes						
Bilingual program	0.146	0.205	0.185	0.187	0.182	0.146
Private	−0.188**	−0.216***	−0.223***	−0.222***	−0.215***	−0.212***
Principal's non-SMT speaker	−0.029	0.027	0.019	0.018	0.015	0.016
Test score	−0.643***	−0.649***	−0.646***	−0.645***	−0.635***	−0.643***
Rural area	−0.086	−0.082	−0.075	−0.074	−0.080	−0.098
School-to-nearest municipal city proximity	0.150*	0.165**	0.176**	0.175**	0.181**	0.214**
Index Teachers	0.005	0.003	0.001	0.001	0.001	0.001
Index Courses	0.057***	0.057***	0.057***	0.057***	0.058***	0.059***
Index Classes	0.017	0.016	0.015	0.015	0.015	0.016
Index Discipline	−0.027	−0.028	−0.028	−0.028	−0.028	−0.029*
Index Infrastructure	0.033	0.020	0.018	0.018	0.020	0.017
Index Distractor	−0.003	−0.009	−0.009	−0.009	−0.010	−0.010
$\sigma_{s,m}$	5.842***	5.918***	5.781***	5.877***	5.687***	6.071***
Household characteristics						
Non-SMT Parent			0.249	0.249	0.243	0.243
Non-SMT Parents			0.219**	0.219**	0.197**	0.197**
Family Income				0.002	0.009	0.009
Years of education					−0.013	−0.013
Child's gender (Male=1)						0.138**
Home-to-Capital city proximity						−0.433*
Interaction of $\sigma_{s,m}$ with School attributes						
Bilingual program	−0.483	−0.545	−0.555	−0.558	−0.544	−0.417
Private	0.339	0.373	0.411	0.410	0.397	0.397
Principal's non-SMT speaker	−0.322	−0.339	−0.541**	−0.540**	−0.533**	−0.521**
Test score	1.061***	1.063***	1.090***	1.089***	1.062***	1.099***
Rural area	0.306	0.311	0.280	0.259	0.275	0.343
School-to-nearest municipal city proximity	−0.515**	−0.589**	−0.640***	−0.630**	−0.648***	−0.681**
Index Teachers	−0.032	−0.038	−0.028	−0.027	−0.026	−0.025
Index Courses	−0.121**	−0.125**	−0.132**	−0.131**	−0.134**	−0.136**
Index Classes	−0.021	−0.018	−0.015	−0.016	−0.016	−0.018
Index Discipline	0.033	0.033	0.028	0.026	0.026	0.028
Index Infrastructure	−0.008	0.007	0.012	0.014	0.010	0.017
Index Distractor	−0.012	0.006	0.009	0.009	0.010	0.011
Household characteristics						
Non-SMT Parent			−1.059**	−1.061**	−1.047**	−1.038**
Non-SMT Parents			0.389	0.384	0.436*	0.456*
Family Income				−0.022	−0.043	−0.041
Years of education					0.036	0.033
Child's gender (Male=1)						−0.710***
Home-to-Capital city proximity						0.112

Note: Spanish mother tongue (SMT) students' self-report of having Spanish as a first language. Non-SMT students speak one of the Mayan laguages, or Xinka or Garífuna languages as a first language. The dependent variable in the logistic regression takes a value of one if a non-SMT student reported that he frequently speaks his mother tongue and zero otherwise. $\sigma_{s,m}$ represents the school's share of non-Spanish mother tongue students who reported mother tongue use. All models include departmental dummy variables except the first one. Robust standard errors are clustered at municipal level.
Significant levels: *** at 1%, ** at 5%, and * at 10%.
Source: Author's calculation using the Ministry of Education's 2013 data.

Table 4.7: Parents' mean preferences for school attributes

	(1) OLS	(2) 2SLS	(3) 2SLS	(4) 2SLS
School attributes for model 5				
Guatemala department	3.542***	9.423***	7.896***	7.594***
Bilingual program	1.662***	1.197	1.286***	1.254***
Private	0.290	7.794	4.277**	4.288**
Rural area	0.627***	−0.468	−0.175	−0.161
School-to-nearest municipal city proximity	0.731***	1.563***	1.462***	1.429***
Principal's non-SMT speaker	0.159*	−0.584	−0.290	−0.241
Index Teachers	0.202***	−0.123	0.028	0.065
Index Courses	−0.554***	−0.297	−0.256	−0.309*
Index Classes	0.290***	0.474*	0.348**	0.347**
Index Discipline	−0.162***	−0.097	−0.122	−0.148
Index Infrastructure	0.353***	0.966*	0.906**	0.868***
Index Distractor	−0.065	0.041	−0.016	−0.029
$\sigma_{s,fees}$	0.106	−3.748	−1.810*	−1.851*
$\sigma_{s,score}$	−0.065**	3.518	1.396	1.062
$\sigma_{s,m}$	−2.564***	1.191	−0.493	0.090
$\sigma_{s,nsmt}$	−0.156	1.865	−0.343	−1.477
$\sigma_{s,nsmt\ household}$	−5.429***	−6.800	−5.466	−4.751**
$\sigma_{s,education}$	−3.247***	−5.285	−5.819	−4.799*
$\sigma_{s,family\ income}$	1.710**	−10.231	−5.211	−5.641
First stage for				
$\sigma_{s,fees}$				
$\sigma_{-s,smt,2013}$		0.010	0.020	0.020
$\sigma_{-s,smt,2010}$		0.008	0.011	0.012
$\sigma_{-s,smt,2013} * \sigma_{-s,smt,2010}$			0.037	0.037
$Test\ score_{-s,2013}$		−0.035*	−0.021	−0.029
$Test\ score_{-s,2010}$		0.026	0.041	0.039
$Test\ score_{-s,2013} * Test\ score_{-s,2010}$			0.065	0.014
$Fees_{-s,2013}$ (in hundreds)		0.811***	−1.749***	−1.793***
$Fees^2_{-s,2013}$ (in hundreds)			9.557***	9.643***
$Family\ education_{-s,2013}$		−0.206**	−0.053	−0.046
$Family\ income_{-s,2013}$		0.084	−0.051	−0.073
$\sigma^{nsmt\ household}_{-s,2013}$		0.023	0.073	0.074
$\widehat{\sigma^{nsmt\ household}_{-s,2013}}$				0.056
Statistics				
R^2		0.939	0.940	0.940
$F_{\beta_Z=0}$		4.776	9.425	10.663
$\sigma_{s,nsmt\ household}$				
$\sigma_{-s,smt,2013}$		−1.176***	−1.189***	−1.189***
$\sigma_{-s,smt,2010}$		−0.320***	−0.241***	−0.226**
$\sigma_{-s,smt,2013} * \sigma_{-s,smt,2010}$			1.125**	1.120**
$Test\ score_{-s,2013}$		0.388*	0.563***	0.447*
$Test\ score_{-s,2010}$		−0.202	−0.218	−0.243
$Test\ score_{-s,2013} * Test\ score_{-s,2010}$			2.002*	1.260
$Fees_{-s,2013}$ (in hundreds)		2.611**	5.306	4.651
$Fees^2_{-s,2013}$ (in hundreds)			−10.551	−9.287
$Family\ education_{-s,2013}$		1.138	1.089	1.190
$Family\ income_{-s,2013}$		−0.905	−1.256	−1.587*
$\sigma^{nsmt\ household}_{-s,2013}$		0.666	0.842	0.859
$\widehat{\sigma^{nsmt\ household}_{-s,2013}}$				0.822
Statistics				
R^2		0.727	0.730	0.730
$F_{\beta_Z=0}$		30.841	26.773	24.685

Note: The dependent variable is the estimates of parents' mean utilities, δ_s, (see eq. 4.5). $sigma_{s,m}$ represents the school's share of non-Spanish mother tongue students who reported mother tongue use. $sigma_{s,nsmt}$ represents the school's share of non-Spanish mother tongue students. $sigma_{s,nsmthousehold}$ represents the school's share of non-Spanish mother tongue parents. $sigma_{s,fees}$ stands for school's fees. $sigma_{s,scores}$ and $sigma_{s,familyincome}$ and $sigma_{s,education}$ stand for the school's average of students' test scores, parents' family income and parents' years of education respectively. $widehatsigma^{nsmthousehold}_{-s,2013}$ represents the simulated school's share of non-Spanish mother tongue parents in period t. All models include departmental dummy variables, and interactions between school and family variables. Robust standard errors are clustered at municipal level.
Significant levels: *** at 1%, ** at 5%, and * at 10%.
Source: Author's calculation using the Ministry of Education's 2013 data.

Table 4.8: Parents' preferences for endogenous school attributes (A)

School attributes * Household characterisitcs	(1)	(2)	(3)	(4)	(5)
$\sigma_{s,m}$					
Mean	7.870*	4.217	4.713	0.100	0.090
Interaction with					
Non-SMT Parent	−1.376**	−1.352**	4.091**	4.470***	4.468***
Non-SMT Parents	−0.834	−0.802	4.707**	5.259***	5.257***
Years of Education	−0.126***	−0.127***	−0.126***	−0.129***	−0.130***
Child's gender (Male=1)	−0.561**	−0.602**	−0.563**	−0.587**	−0.591***
Family income (in thousands)	−0.015	−0.009	−0.017	−0.009	−0.010
Home-to-capital city proximity		−9.873		−9.286	−9.285
$\epsilon_p^{(a)}$					0.046
$\sigma_{s,nsmt}$					
Mean	−5.390	−1.911	−7.063	−1.537	−1.477
Interaction with					
Non-SMT Parent	−1.213**	−1.217**	−1.439**	−1.465**	−1.467***
Non-SMT Parents	−0.006	−0.111	−0.086	−0.219	−0.225
Years of Education	−0.006	−0.003	−0.020	−0.017	−0.018*
Child's gender (Male=1)	0.260	0.319	0.367	0.440	0.432***
Family income (in thousands)	0.059	0.064	0.068*	0.073*	0.072***
Home-to-capital city proximity		−3.446		−3.708	−3.708
$\epsilon_p^{(a)}$					−0.013
$\sigma_{s,nsmt\ household}$					
Mean	−3.331	−4.700**	−1.969	−4.735**	−4.751**
Interaction with					
Non-SMT Parent	7.010***	7.071***	7.034***	7.118***	7.115***
Non-SMT Parents	8.631***	8.804***	8.582***	8.778***	8.772***
Years of Education	0.237***	0.254***	0.243***	0.261***	0.260***
Child's gender (Male=1)	0.491	0.570*	0.420	0.494*	0.486***
Family income (in thousands)	−0.015	−0.015	−0.016	−0.016	−0.017
Home-to-capital city proximity		−9.356		−9.836	−9.836*
$\epsilon_p^{(a)}$					0.005
$E\left(m_i \mid X, \sigma_{s,m}\right)$					
Mean			3.158*	3.659**	3.650***
Interaction with					
Non-SMT parent			−5.658***	−5.944***	−5.945***
Non-SMT parents			−5.142***	−5.617***	−5.620***

Note: $sigma_{s,m}$ represents the school's share of non-Spanish mother tongue students who reported mother tongue use. $sigma_{s,nsmt}$ represents the school's share of non-Spanish mother tongue students. $sigma_{s,nsmthousehold}$ represents the school's share of non-Spanish mother tongue parents. $sigma_{s,fees}$ stands for school's fees. $sigma_{s,scores}$ and $sigma_{s,familyincome}$ and $sigma_{s,education}$ stand for the school's average of students' test scores, parents' family income and parents' years of education respectively. All models include departmental dummy variables, and interactions between school and family variables. (a) Estimates multiplied by 10. Robust standard errors are clustered at municipal level.
Significant levels: *** at 1%, ** at 5%, and * at 10%.
Source: Author's calculation using the Ministry of Education's 2013 data.

Table 4.9: Parents' preferences for endogenous school attributes (B)

School attributes * Household characterisitcs	(1)	(2)	(3)	(4)	(5)
$\sigma_{s,fees}$					
Mean	−0.588	−1.834*	−0.688	−1.856*	−1.851*
Interaction with					
Non-SMT Parent	0.099	−0.028	0.130	−0.007	−0.007
Non-SMT Parents	0.144	−0.054	0.174	−0.034	−0.034
Years of Education	−0.052	−0.067	−0.051	−0.066	−0.066***
Child's gender (Male=1)	−0.067	−0.150	−0.069	−0.155	−0.155***
Family income (in thousands)	−0.062**	−0.060**	−0.062**	−0.061**	−0.061***
Home-to-capital city proximity		10.353**		10.401**	10.401
$\epsilon_p^{(a)}$					−0.004
$\sigma_{s,scores}$					
Mean	0.330	0.707	−0.163	1.060	1.062
Interaction with					
Non-SMT Parent	0.280***	0.288***	−0.225**	−0.241**	−0.241***
Non-SMT Parents	0.429***	0.439***	−0.067	−0.100	−0.100
Years of Education	0.020***	0.020***	0.019***	0.020***	0.020***
Child's gender (Male=1)	0.059	0.062	0.063	0.068	0.068***
Family income (in thousands)	0.003	0.002	0.003	0.003	0.003
Home-to-capital city proximity		0.570		0.510	0.510
$\epsilon_p^{(a)}$					−0.003
$\sigma_{s,education}$					
Mean	−4.923**	−4.878*	−3.136	−4.799*	−4.799*
Interaction with					
Non-SMT Parent	0.646*	0.801*	0.612	0.787*	0.788***
Non-SMT Parents	0.449	0.669	0.384	0.622	0.628***
Years of Education	0.732***	0.755***	0.732***	0.755***	0.756***
Child's gender (Male=1)	−0.393	−0.271	−0.388	−0.266	−0.271***
Family income (in thousands)	−0.361***	−0.334***	−0.362***	−0.336***	−0.336***
Home-to-capital city proximity		−13.707***		−13.930***	−13.925***
$\epsilon_p^{(a)}$					0.042
$\sigma_{s,family\ income}$					
Mean	−6.233	−5.005	−6.077	−5.622	−5.641
Interaction with					
Non-SMT Parent	−0.910*	−0.934*	−0.695	−0.705	−0.705***
Non-SMT Parents	−0.314	−0.282	−0.085	−0.035	−0.031
Years of Education	−0.522***	−0.529***	−0.521***	−0.527***	−0.527***
Child's gender (Male=1)	−0.551	−0.594	−0.565	−0.606	−0.608***
Family income (in thousands)	0.438***	0.429***	0.439***	0.430***	0.430***
Home-to-capital city proximity		3.211		3.159	3.162
$\epsilon_p^{(a)}$					−0.010
Other coefs.					
Home-School proximity (<=75 kms)			7.912***	12.522***	12.522***
Home-School proximity (>75 kms)			−6.565***	−6.356***	−6.356***
Home-School proximity (<=75 kms) – Home-Cap.City proximity				−126.256*	−126.256***
Home-School proximity (>75 kms) – Home-Cap.City proximity				−4.005	−4.005
Other controls.					
$\sigma_{s,m}$	Yes	Yes	Yes	Yes	Yes
$\sigma_{s,nsmt}$	Yes	Yes	Yes	Yes	Yes
$\sigma_{s,nsmt\ household}$	Yes	Yes	Yes	Yes	Yes
$E(m_i \mid X, \sigma_{s,m})$	Yes	Yes	Yes	Yes	Yes

Note: $sigma_{s,m}$ represents the school's share of non-Spanish mother tongue students who reported mother tongue use. $sigma_{s,nsmt}$ represents the school's share of non-Spanish mother tongue students. $sigma_{s,nsmthousehold}$ represents the school's share of non-Spanish mother tongue parents. $sigma_{s,fees}$ stands for school's fees. $sigma_{s,scores}$ and $sigma_{s,familyincome}$ and $sigma_{s,education}$ stand for the school's average of students' test scores, parents' family income and parents' years of education respectively. All models include departmental dummy variables, and interactions between school and family variables. (a) Estimates multiplied by 10. Robust standard errors are clustered at municipal level.
Significant levels: *** at 1%, ** at 5%, and * at 10%. Source: Author's calculation using the Ministry of Education's 2013 data.

Table 4.10: Parents' preferences for school exogenous attributes

	Mean	Non-SMT Parent	Non-SMT Parents	Years of Education	Child's gender (Male=1)	Family income (in thousands)	Home-to-capital city proximity	$\epsilon_p^{(a)}$
Model 5								
Bilingual program	1.254***	−0.778***	−1.115***	−0.044***	0.305*	−0.050***	−28.120	0.003
Private	4.288**	−0.666***	−0.681***	0.006	−0.103*	0.107***	−16.295	−0.551
Rural area	−0.161	−0.281*	−0.122	0.030**	0.149***	−0.295***	−1.726	−0.856
School-to-nearest municipal city proximity	1.429***	−0.356***	−0.320***	−0.099***	−0.413***	−0.027***	10.580**	−0.084
Non-SMT Principal	−0.241	−0.381***	−0.392	0.004	−0.013	−0.013*	−0.527	−0.657
Index Teachers	0.065	−0.309***	−0.214***	−0.017***	−0.014	−0.013***	1.648	−0.006
Index Courses	−0.309*	0.289***	0.542***	0.034***	0.169***	0.027***	4.664**	−0.049
Index Classes	0.347**	−0.201***	−0.109**	−0.015***	−0.134***	−0.014***	0.548	0.019
Index Discipline	−0.148	0.120***	0.194*	0.008	0.110***	−0.001	−3.403	−0.005
Index Infrastructure	0.868***	0.385***	0.070	−0.008*	0.089***	0.005	2.411***	−0.032
Index Distractor	−0.029	0.082***	0.186***	−0.004	−0.014	−0.001	2.460***	0.016

Note: This model includes departmental dummy variables, and interactions between school and family variables. Robust standard errors are clustered at municipal level. (*) Estimates multiplied by 100. Significant levels: *** at 1%, ** at 5%, and * at 10%.
Source: Author's calculation using the Ministry of Education's 2013 data.

Table 4.11: Non-SMT parents' departmental demand elasticity for schools where parents have a similar mother tongue by school type

	(1)	(2)	(3)	(4)
Department	Muni-Coope$^{(a)}$	Public	Private	Mean
Sololá	0.645***	0.644***	0.581***	0.625***
Alta Verapaz	0.586***	0.584***	0.454***	0.550***
Quiché	0.565***	0.561***	0.475***	0.544***
Totonicapán	0.590***	0.575***	0.370***	0.527***
Huehuetenango	0.512***	0.441***	0.326***	0.416***
Chimaltenango	0.349***	0.353***	0.264***	0.314***
Baja Verapaz	0.262***	0.290***	0.302***	0.286***
Quetzaltenango	0.272***	0.294***	0.257***	0.277***
Izabal	0.273***	0.282***	0.158***	0.225***
Sacatepéquez	0.160***	0.187***	0.126***	0.153***
Petén	0.222***	0.151***	0.093***	0.149***
Suchitepéquez	0.123***	0.128***	0.139***	0.130***
San Marcos	0.160***	0.128***	0.038***	0.127***
Guatemala	0.077***	0.108***	0.083***	0.090***
Retalhuleu	0.085***	0.084***	0.075***	0.083***
Escuintla	0.032***	0.047***	0.089***	0.059***
Jalapa	0.011***	0.016***	0.035***	0.021***
Chiquimula	0.007***	0.020***	0.007***	0.015***
El Progreso	0.009***	0.001***	0.021***	0.008***
Zacapa	0.001***	0.008***	0.011***	0.008***
Santa Rosa	0.006***	0.004***	0.018***	0.008***
Jutiapa	0.003***	0.004***	0.015***	0.007***

(a) Muni-Coope stands for municipal or cooperative school types. Significant levels: *** at 1%, ** at 5%, and * at 10%.
Source: Author's calculation using the Ministry of Education's 2013 data.

Table 4.12: SMT parents' maginal utilities and their willingness to pay for school attributes

	Marginal utility for school attribute by school-capital city distance (Kms.)				Increment in school attribute	Willigness to pay in Q (US) for school attribute by school-capital city distance (Kms.)			
	>200	200-100	100-50	50-0		>200	200-100	100-50	50-0
School attribute									
$Testscore_{s,t}$	12.815***	12.898***	12.824***	13.752***	0.1 std	148.07 (19.11)	515.53 (66.52)	315.10 (40.66)	208.10 (26.85)
Index Teachers	−1.506***	−1.581***	−1.329***	0.679***	1%	-3.47 (-0.45)	-12.61 (-1.63)	-6.52 (-0.84)	2.05 (0.26)
Index Courses	2.209***	2.643***	2.573***	9.674***	1%	2.82 (0.36)	11.65 (1.50)	6.98 (0.90)	16.15 (2.08)
Index Classes	0.800***	0.700***	0.897***	1.375***	1%	1.09 (0.14)	3.29 (0.43)	2.60 (0.34)	2.45 (0.32)
Index Discipline	−0.439***	−0.566***	−0.793***	−5.469***	1%	-1.74 (-0.22)	-7.76 (-1.00)	-6.69 (-0.86)	-28.42 (-3.67)
Index Infrastructure	8.723***	8.812***	8.942***	12.278***	1%	39.80 (5.14)	139.08 (17.95)	86.77 (11.20)	73.36 (9.47)
Index Distractor	−0.684***	−0.610***	−0.445***	2.937***	1%	-2.73 (-0.35)	-8.42 (-1.09)	-3.78 (-0.49)	15.36 (1.98)

Note: to calculate parents' willingness to pay for school attributes, I first calculate the total derivative of the parent's utility function. Then, the willingness to pay is given by the amount of money parents are willing to pay for a change in one school attribute.
Significant levels: *** at 1%, ** at 5%, and * at 10%.
Source: Author's calculation using the Ministry of Education's 2013 data.

Table 4.13: SMT parents' compensation in both monetary and school quality terms for attending their nearest rural school

Department	Difference in parents' utilities (+)	Monthly payment in Q (USD)	Times to increment rural school quality in terms of Infrastructure	Courses
Sololá	6.674***	477.66 (61.63)	4.06	30.99
Alta Verapaz	0.637	45.60 (5.88)	0.38	3.69
Quiché	2.715***	194.34 (25.08)	1.58	21.77
Totonicapán	0.433	30.97 (4.00)	0.25	5.58
Huehuetenango	0.651	46.59 (6.01)	0.37	18.49
Chimaltenango	1.591***	113.86 (14.69)	0.88	68.26

Note: Spanish mother tongue (SMT) stands for students' self-report of having or that their parents have Spanish as a first language . Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. The third column represents the SMT parents' average monthly payment to select the nearest rural school conditional on a school' share of non-SMT parents higher or equal to 0.35. The fourth column indicate the increase on the rural school infrastructure relative to a school with an average quality. The last column has the same interpretation as the third one, but with the classroom index.
(+) The first column shows the difference in SMT parents' utilities for considering their child's urban school enrollment relative to the closest rural school.
Source: Author's calculation using the Ministry of Education's 2013 data.

Table 4.14: Effects of the Guatemalan civil war on the difference of non-SMT households' departmental demand elasticities for schools where parents have a similar mother tongue

	(1)	(2)	(3)	(4)
Ln(Victims per 1000 population)	0.136**	0.118**	0.119**	0.097*
Ln(Illiteracy in 1973)			1.028	5.532*
Ln(Gini coefficient in 2013)		−5.434***	−5.117***	−4.523***
Ln(Poverty index UNDP)				−3.722*
Dummy variable		0.985	0.748	0.736
Constant	−3.459***	14.486***	15.333***	17.676***
R^2	0.188	0.549	0.557	0.617
Observations	22	21	21	21

Note: Spanish mother tongue (SMT) stands for students' self-report of having or that their parents have Spanish as a first language . Non-SMT students speak either one of the Mayan languages, or Xinka or Garífuna languages as a first language. The dependent variable is the difference of non-SMT households' departamental demand elasticities (shadow area in Figure 1). The variables victims per 1000 population, Illiteracy in 1973, and the dummy variable's specification are the ones employed by Chamarbargala and Moran (2012). The dummy variable takes a value of one for the departments that experienced the worst period of the civil war as in Chamarbargala and Moran (2012). The Gini index in 2011 is calculated by Instituto Nacional de Estadistica de Guatemala while the proverty index is calculated by United Nations Development Program. The Gini index is not calculated for the Guatemalan department. Significant levels: *** at 1%, ** at 5%, and * at 10%. Source: Author's calculation using the Ministry of Education's 2013 data.

Figure 4.1: Departmental demand elasticity for schools where parents have a non-Spanish mother tongue by number of non-SMT parents at home

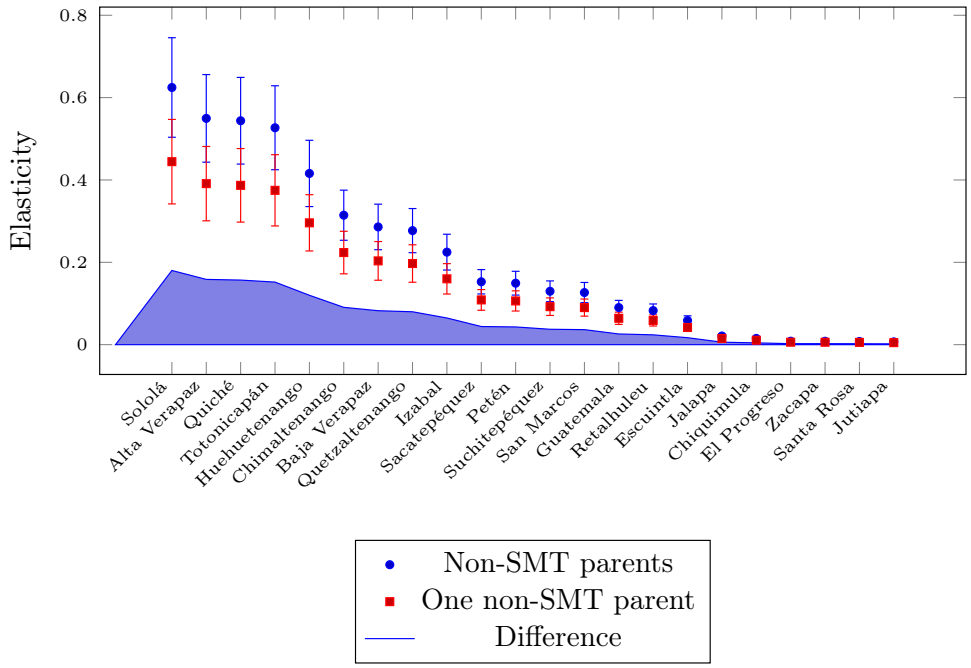
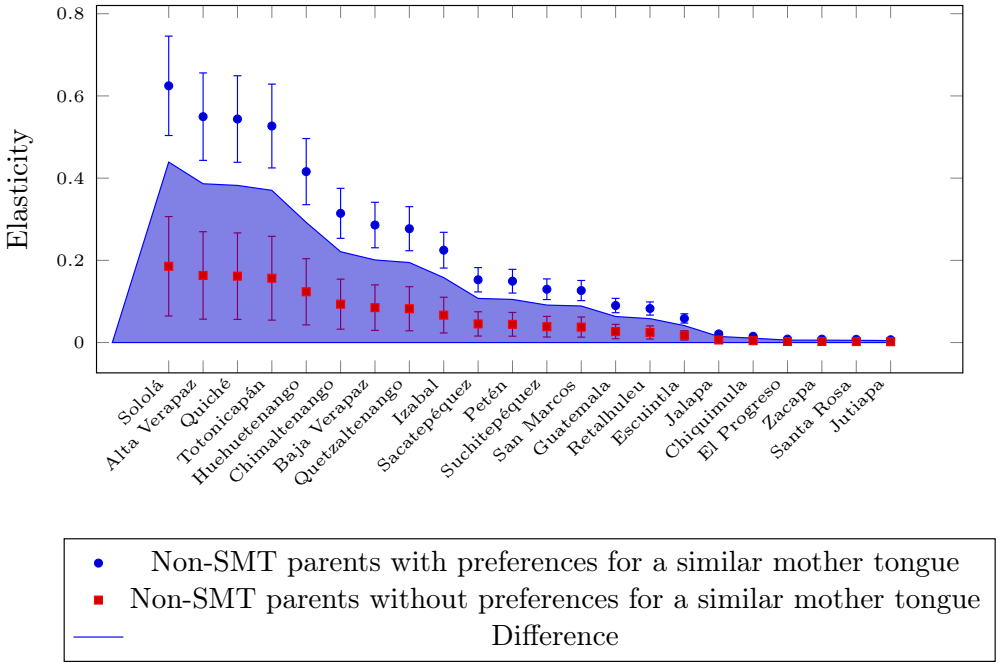


Figure 4.2: Non-SMT parents' departmental demand elasticity for schools where parents have a similar mother tongue by mother tongue preference



Appendix A

Third chapter's appendix

A.1 Appendix

Model for school progression

The model for students' selection of school progression is the following:

$$\begin{matrix} student & i & observed \\ in & period & t \end{matrix} =$$

$$\begin{cases} & \beta_1 \left(y^r_{i,t-1} - y^r_{s,t-1}\right) + \beta_2 \left(y^m_{i,t-1} - y^m_{s,t-1}\right) \\ 1 \quad if & +\beta_4 \left(y^r_{i,t-1} - y^r_{s,t-1}\right)\left(y^m_{i,t-1} - y^m_{s,t-1}\right) \quad \geqslant 0 \\ & +\beta_4 ParentsMaxGrade_{t-1} + \epsilon_{i,t} \\ 0 \quad Otherwise & \end{cases}$$

The inclusion of the parents' educational attainment is to control for those who had scores higher than the schools' thresholds, but not observed after one period.

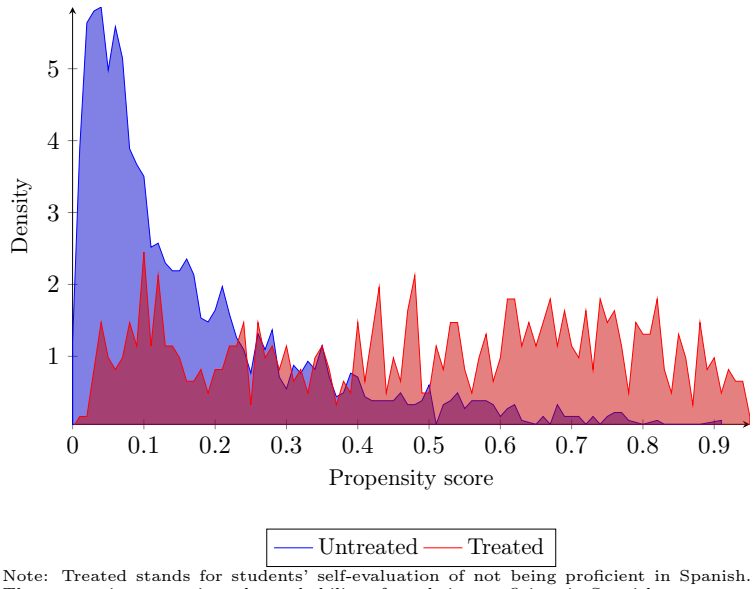
Table A.1: Logit model for students' grade progression

	Coef.
Scores	
$\left(y^r_{i,t-1} - y^r_{s,t-1}\right)$	0.106***
$\left(y^m_{i,t-1} - y^m_{s,t-1}\right)$	0.280***
$\left(y^r_{i,t-1} - y^r_{s,t-1}\right)\left(y^m_{i,t-1} - y^m_{s,t-1}\right)$	0.072***
Parents' educational attainment in $t-1$	
Mother	
Elementary	0.149***
Junior high	0.558***
High school	0.231***
Undergrad	0.448***
College	1.286***
Father	
Elementary	0.244***
Junior high	0.590***
High school	0.042
Undergrad	0.253***
College	0.899**
Constant	-0.321***

Note: The dependent variable in the logit model takes a value of one if students attend school at time t, and zero otherwise. The variable, y_i, stands for students' test scores. These students are those who can be followed over grades. The superscripts m or r stand for reading and math respectively. The variable, y_s, stands the average test score at school s. These average test scores are calculated using all junior high and high school students. Significant levels: *** at 1%, ** at 5%, and * at 10%. Source: Author's calculation using the Ministry of Education's 2010, 2013 and 2015 data.

Frequency of propensity scores by treatment status

Figure A.1: Frequency of propensity scores by treatment status



Note: Treated stands for students' self-evaluation of not being proficient in Spanish. The propensity score gives the probability of not being proficient in Spanish.
Source: Author's calculation using the Ministry of education's data.

Appendix B

Fourth chapter's appendix

B.1 Appendix

Principal component analysis

The dataset used in this document was collected from the Ministry of Education of Guatemala in 2013 by surveying schools' principals. The dataset contains a rich set of school variables to control for in the baseline specification. This includes data about infrastructure and principals' educational attainment for instance. Controlling for many variables will dramatically increase the computational burden; therefore, I employ a data reduction by principal component analysis (PCA) and employ the first component or index in the model estimation. I group school variables to construct six quality indices: teachers, courses, classroom, infrastructure, discipline, and distraction. See next table for a detailed description of the variables included in each index and for the proportion of variance that explains each index.

Table B.1: School variables for school quality indices

	Teacher	Discipline	Infrastructure	Classes	Distraction	Courses
Variables						
	Teacher's reading performance	Reading test frequency	Ilumination	Reading classes per week	Interior noise	Laboratory computation
	Teacher's math training	Math test frequency	Ventilation	Math classes per week	Exterior noise	English
	Train teachers due to previous results	Share test score with students	Electricity at school	Time per class	Classroom is hot	Library
	Share test scores with teachers	Share test score with parents	Water service (pipeline)	Use support material from Mineduc	Classroom is cold	
		Principal's interaction with teachers	Basketball field	Use previous national test	Pollution	
		Principal's interaction with students	Volleyball field	Improve reading class due to previous test scores		
		Principal's interaction with parents	Football field	Improve math class due to previous test scores		
		Principal's frequency of supervision	Swimming pool			
			Roof infrastructure			
			Floor infrastructure			
			Wall infrastructure			
			Number of toilets			
			Number of handwashers			
			Number of urinals			
			Typing room			
First component						
Eigenvalue	1.77637	2.70614	5.07096	1.96463	2.15729	2.01027
Proportion of variance	0.9688	0.5697	0.4609	0.5867	1.1717	0.8731

Note: All categorical variables are ordered in such a way that a higher value of the index represents a better school quality.
Source: Author's calculation using the Ministry of Education's 2013 data.

Family income and school fees

An important drawback in this dataset is that, while parents' education attainment is observed, family income is not. To overcome this, I employ the 2014 Guatemalan national survey to construct a family income variable as a function of education attainment. Specifically, I regress $log(wage)$ on the mother's educational attainment, departmental dummies, and a rural area dummy. With the estimates of this regression, I employ the same variables, using the dataset in this document, to construct the mother's expected wage. I replicate this procedure for the father's expected wage. Finally, the expected family income is the sum of the mother and father's expected wages. I employ the same procedure to obtain an amount paid by parents at each private school. Then the private school fees are just the mean of the parents' payments at each school.[1]

[1]This mean of household payments includes all households at each school, and not only those households that the estimation employs.

Table B.2: Wage equations for Fathers and Mothers

	Father	Mother
Years of education		
Six	0.284***	0.264***
Nine	0.600***	0.490***
Twelve	0.730***	0.744***
Eighteen	1.080***	0.768***
Twenty	1.428***	2.193***

Note: The dependent variable is the logarithmic transformation of wages reported in the Guatemala's 2014 national survey. Both equations include departmental and rural area dummy variables. Significant levels: *** at 1%, ** at 5%, and * at 10%.
Source: Author's calculation using the Guatemala's 2014 national survey.

Table B.3: Parents' school fees

	Coef.
Variables	
Private school	9.423***
Municipal school	8.386***
Cooperative school	8.303***
Log(family income)	0.460***

Note: The dependent variable is the parents payment at school reported in the Guatemala's 2014 national survey. The equation includes departmental and rural area dummy variables. Significant levels: *** at 1%, ** at 5%, and * at 10%.
Source: Author's calculation using the Guatemala's 2014 national survey.

Curriculum Vitae

Name:	Fidel Pérez Macal
Post-Secondary Education and Degrees:	University of Western Ontario London, ON, Canada 2014-2020 (Expected) Ph.D. in Economics
	University of Western Ontario London, ON, Canada 2013-2014 M.A. in Economics
	Universidad Católica de Chile Santiago de Chile, Chile 2009-2010 M.A. in Applied Macroeconomics
	Universidad de San Carlos de Guatemala Guatemala, Guatemala 1998-2005 Mechanical-Industrial Engineer
Related Work Experience:	Central Bank of Guatemala Economic Analyst 2005-2013