Essays on Education Policy and Student Achievement in Colombia

by Diana Patricia Hincapié

B.A. in Economics, September 2005, Universidad de Los Andes
Master in Economics, March 2007, Universidad de Los Andes

A Dissertation submitted to

The Faculty of
The Columbian College of Arts and Sciences
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

May 18th, 2014

Dissertation directed by

Stephanie R. Cellini
Associate Professor of Public Policy and Public Administration, and Economics

The Columbian College of Arts and Sciences of The George Washington University certifies that Diana Patricia Hincapié has passed the Final Examination for the degree of Doctor of Philosophy as of March 6th, 2014. This is the final and approved form of the dissertation.

Essays on Education Policy and Student Achievement in Colombia

Diana Patricia Hincapié

Dissertation Research Committee:

Stephanie R. Cellini, Associate Professor of Public Policy and Public Administration, and Economics, Dissertation Director

Dylan Conger, Associate Professor of Public Policy and Public Administration, Committee Member

Burt Barnow, Amsterdam Professor of Public Service, Committee Member

Dedication

Dedicado a mi familia.

A mi mamá y mi papá, por enseñarme el valor de la educación, y por inculcarme el amor por el estudio y las ganas de aprender algo nuevo cada día.

A mi hermana y hermano, por escucharme, aguantarme y apoyarme siempre.

A mi tía Myriam, por contagiarme el amor a la investigación, los datos, y el estudio.

A Leo, por el apoyo y ánimo permanentes, y por creer siempre en mí.

En memoria de mis tías Myriam, Dolly y Flor, quienes se nos fueron para siempre mientras escribía esta Disertación.

Acknowledgements

I am specially indebted to my advisor and mentor, Stephanie Cellini, for being the best advisor that could possibly exist. For going beyond her duties as an advisor every single time we met, talked, or emailed, in order to help me and support me. For providing great advice and feedback, enthusiastically discussing every issue I brought up in our meetings and conversations, and encouraging me to think critically. For always being available, even throughout her months of leave, when she would go to her office just to meet with me, or will call between her children's naps to discuss anything I needed. I once read that she hoped to become as good as an advisor as her own advisor had been to her, and I am sure she has exceeded those expectations in becoming such an amazing advisor to me. But above all, I thank her for always believing in me, and for being a great mentor and friend. Thanks, Stephanie.

I am also grateful to my committee members, Dylan Conger, Burt Barnow, and Hal Wolman, for their constant support and encouragement, their insightful discussions and suggestions, and for always being available for me and giving me so much of their time. I thank Dylan, for convincing me to attend GWU in the first place; Hal, for believing in me since the first day we met; and Professor Barnow, for editing my papers and improving my English, and for helping me understand the links between economic theory and public policy, and the benefits of being an economist specializing in public policy. I also wish to thank Felipe Barrera, for encouraging me to study education issues, and for his valuable advice and feedback.

I will always be grateful to (and for) my family. My mom is the most amazing educator that I know, and the greatest inspiration of my life and work. Because of her, many years ago I realized the importance of having great schools, devoted educators, and a good education. My dad taught me the value of studying and working hard, and through his example, taught me the importance of helping others every time I could, which led me to pursue a career and a professional path that will allow me helping other people. I thank my sister, for being the greatest company, confident, supporter, and friend that I could have, and my brother, for his support and for always providing great advice on any possible issue. I also wish to thank: my aunt Myriam, for passing on to me the love for research, studying, data and statistics; Susy, for her constant encouragement; and my nephews Tomas and Lucas, for bringing so much joy to our lives.

Finally, I would like to thank Leo, for his infinite patience, encouragement, and his constant support through this PhD process. But above all, for believing in me even more than I believe in myself. I could not have done it without him by my side all these years. Thanks.

Abstract of the Dissertation

Essays on Education Policy and Student Achievement in Colombia

The main objective of this dissertation is to analyze the impact that two notable school reforms have had on student achievement in Colombia. The dissertation consists of three essays. The first essay lays out the conceptual framework for the dissertation. It describes the education production function that underlies most analyses in the economics of education, and reviews the main evidence on the impact of school resource policies on student outcomes.

The second essay analyzes the impact of longer school days on student achievement in Colombia, where primary and secondary students attend schools that have either a complete (7-hour) or a half-day (4-hour) schedule. Using test score data from 5^{th} and 9^{th} graders in 2002, 2005, and 2009, along with school administrative data, this study identifies the effect of longer school days by implementing a school fixed effects model. The main model compares variation in average test scores across cohorts for schools that switched from a complete schedule to a half schedule and vice versa. I find that among schools that switch schedules between 2002 and 2009, the cohorts exposed to complete schedules have test scores that are about one tenth of a standard deviation higher than cohorts that attended half schedules. The impact of a complete schedule is larger for math test scores than for language test scores, and it is larger for 9^{th} grade test scores than for 5^{th} grade test scores. Effects are largest among the poorest schools in the sample, and

those in rural areas. The results suggest that lengthening the school day may be an effective policy for increasing student achievement, particularly for the lowest-income students in Colombia and other developing countries.

The third essay analyzes the impact of the "Escuela Nueva" (EN) model (New School) on student achievement, using test score data from SABER 2002 and 2005, a national standardized test administered to 5^{th} and 9^{th} graders in Colombia. EN is an educational model originally designed to improve the effectiveness of rural schools. It is characterized by multigrade classrooms (i.e., one instructor teaches students in various grades in the same classroom), a child-centered curriculum, flexible systems of grading and promotion, intensive teacher training, and parental involvement. To mitigate the concerns about systematic selection of schools into EN that might bias the estimations of the EN impact, this study implements a school fixed effects model that controls for time-invariant characteristics within the school. Results show that among schools that switched models between 2002 and 2005, the cohorts of 5^{th} grade students exposed to EN have on average 0.135 of a standard deviation higher language test scores than cohorts exposed to other models, while there is no statistically significant impact on switching to EN for 9^{th} graders. The impact of EN is largest among rural schools and the poorest schools in the sample.

Table of Contents

List of Tables

Essay 1: Education production function and the effectiveness of education policies

1.1 Introduction

According to Human Capital Theory (Becker, 1964), education can generate higher incomes, both individually and at the society level, mainly through an increase in labor productivity. On one hand, education increases people's knowledge, abilities, and skills, which allows them to produce more output per unit of time, and this higher productivity translates into a higher growth for a given country. On the other hand, at the individual level, the increase in the productivity of workers raises their incomes. For this reason, improving access to education and increasing the education levels of the population is a central development strategy in most countries. It has also become one of the most important goals in developing countries, as reflected in the Millennium Development Goals (MDGs), a set of agreements made by most of the countries and the world's leading development institutions. For example, the second MDG is to achieve universal primary education by 2015.[1] Partly motivated by this goal, during recent decades many developing countries have expanded access to primary education, and some of them have already achieved or are close to achieving universal coverage.

However, expanding coverage *per se* might not necessarily translate into more or better skills for students. In fact, despite their dramatic increase in access to primary and secondary education, most developing countries are a long way from achieving

[1] Millennium Development Goal No. 2: "Ensure that, by 2015, children everywhere, boys and girls alike, will be able to complete a full course of primary schooling". http://www.un.org/millenniumgoals/education.shtml.

educational levels that make them competitive in the globalized world. The OECD Programme for International Student Assessment (PISA), which has made possible comparing student performance in Language, Mathematics, and Science among 65 countries, has shown that students in developing countries are far from the achievement levels of students in develop countries. For example, in 2012, the performance of the eight Latin American countries that participated in the PISA assessment (Argentina, Brazil, Chile, Colombia, Costa Rica, Mexico, Peru, and Uruguay) was in the lowest third of the distribution among all participating countries, and the average student in seven of these countries only achieved the lowest level of performance (Level 1 out of 6 possible levels), which means that he or she does not have even the most basic skills in the respective subject (Bos, Ganimian, & Vegas, 2013).

Additionally, the quality of education in developing countries is highly unequal, and there are huge achievement gaps within the countries. For example, there are large differences in the quality of private and public schools (Gaviria & Barrientos, 2001a), between schools in rural and urban areas, and between students of different socioeconomic backgrounds (World Bank, 2008).

In this context, the focus of education policy research and public policy has shifted from increasing coverage to improving the quality of education, with policy makers and researchers arguing that it is at least as important as the quantity of education in effectively increasing people's skills and abilities, and consequently, in raising individual income and in generating economic growth.[2] In developing countries, the low

[2] See for example, Hanushek, E. and Wossermann, L. (2007). Education quality and economic growth. World Bank. Washington, DC.

achievement levels of students compared to those in developed countries is seen as an obstacle to economic growth and becoming competitive in the global economy. At the same time, the gaps in the educational achievement between private and public schools, between urban and poor areas, and between high-income and low-income families, are seen as a big obstacle to achieving equality of opportunity for all and reducing poverty.

Despite the importance that has been given to the quality of education, there is still no agreement about what exactly we mean by "quality." Is a higher quality of education one that allows students to achieve higher test scores? Is it one that better prepares students for future jobs? Is it one that forms better citizens? Part of these disagreements come from the differences in what parents, policy makers, and educators expect from the education system and what they think is really important for children to learn at school. Usually, people expect children who go to school to be able to at least learn some basic concepts in core areas that allow them to go through higher grades or levels of education, or to work, but this is difficult to determine and to measure. There is even less agreement about which education policies are effective in improving the quality of education.

The main objective of this Essay is to provide an overview of the status of the quality of education in Colombia, and to analyze the effectiveness of some of the most common education policies by looking at the international and domestic evidence that has analyzed their impact. The rest of the paper is organized as follows: Section 1.2 describes the status of the quality of education in Colombia. Section 1.3 provides a description of the education production function, a conceptual framework frequently used in developed and developing countries to analyze education policy. Section 1.4 discusses the evidence

on the effectiveness of education policies, and Section 1.5 concludes and provides some policy recommendations for improving the quality of education in Colombia.

1.2 Education in Colombia

In the last two decades, Colombia has significantly increased primary and secondary education enrollment rates. In 1991, 71% of children of primary school age were attending a primary school grade, while 88% were attending a primary school grade in 2011 (World Bank EdStats, 2014). However, in terms of quality, performance on international and national tests shows that student achievement is still very low compared to industrialized countries and even to other similar developing countries, and there are huge achievement gaps within the country. Part of this could be explained by the increase in the number of children attending school, if the incoming students are less qualified and pull down the average scores (Glewwe, Hanushek, Humpage, & Ravina, 2011). For example, if mostly the poorer students entered the system in recent years, it is possible that the decline in average student achievement is driven by the characteristics of these new students, who are less likely to do well in school. However, part of the poor achievement of students might be caused by the government's focus on increasing access to education at the expense of improving its quality. In particular, both the central and local governments have focused on increasing school resources in order to increase school capacity and enrollment (e.g., building new schools).[3]

[3] For example, one of the central education policies in Bogota during the second half of the previous decade was to construct public "mega-schools", with the capacity to attend about 1000-1500 students each.

In the previous decade (2000-2010), the Colombian Ministry of Education made some attempts to reform the educational system with the aim of improving the quality of education. One of them was the New Teacher's Statute of 2002. The goal of this statute was to "look for, attract, and keep the best professionals at the service of the public education."[4] It established new policies for teachers, including new merit-based recruiting and promotion policies (e.g. changes in the teachers' wage structure, commonly known as the "Teachers' Ladder"), and the evaluation of teachers' abilities and competences in order for them to move up in the Teachers' Ladder or to remain in the system. The objective of these policies was improving teachers' quality, and consequently, the quality of the education that students received in public schools. Policymakers hoped that as a result of the policy, student achievement in public schools would improve as well. The government also moved forward in the process of decentralizing education, which changed the way in which the central government assigned resources to local governments (departments and municipalities), hoping that these changes would generate incentives for them to improve the efficiency of education expenditures and the quality of education.

Despite all the government efforts to improve the quality of education in Colombia, student achievement in the country is still very low compared to other countries. In 2009, the country ranked 52 in reading, 54 in science, and 58 in math, among 65 developed and emerging economies that participated in the PISA assessments (OECD, 2010). In 2012, test scores in the three areas evaluated went down, and the country ranked even worse than in 2009: it ranked 62 in math, 58 in science and 55 in

[4] Estatuto de Profesionalización Docente, Decreto 1278 de 2002.

reading, among the 65 countries and economies evaluated in that round (OECD, 2013). Most of the tested countries were developed countries and therefore not totally comparable to Colombia, but the country's educational achievement lagged even with respect to relatively similar Latin American economies like Mexico, Chile, and Uruguay. These results are worrisome because students that are not learning and that do not have the necessary skills cannot help make the country competitive and cannot contribute to economic growth.

In addition to the poor performance of Colombian students on the PISA tests and other international assessments, there is another important problem: as in many other developing countries, the educational system is highly unequal. Colombia has the second highest level of inequality in Latin America (after Brazil) and one of the highest levels of inequality in the world, with a Gini coefficient of income[5] of 0.54 for 2012 (DNP, 2013). This huge inequality is also reflected in the distribution of education both in terms of access and of quality. Access to primary and secondary education has significantly increased in the last decade; however, there are significant differences in achievement between rural and urban areas, students with different socioeconomic backgrounds, public and private schools, and schools in different departments or municipalities. This not only prevents students from developing their full potential and hurts economic growth, but also traps a significant part of the population in poverty and perpetuates income inequality.

[5] The Gini coefficient is a measure that represents the income distribution of a nation's citizens. The Gini coefficient goes from zero to one, where zero represents perfect equality and one represents perfect inequality.

Regardless of the fact that student achievement in Colombia seems to be lagging behind other countries, and of the huge achievement gaps that exist within the country, there is relatively little evidence about the factors associated with student achievement, and even less about which type of interventions *cause* an increase in student achievement. This could be, in part, due to the lack of large datasets or representative surveys about education in the country. For example, only a few years ago the *Colombian Institute for Educational Evaluation (*ICFES) released to the public data from the first three rounds of the SABER 5^{th} and 9^{th}. SABER 5^{th} and 9^{th} are national standardized tests designed to evaluate 5^{th} grade and 9^{th} grade students in language, math, and science all over the country.

Colombia has also participated in some international assessments, like the TIMSS (Trends in International Mathematics and Science Study), PIRLS (Progress in International Reading Literacy study, CIVED, ICCS (Civic Education Study), PERCE, SERCE (Regional Comparative and Explicative studies), and, most notably, PISA. The PISA is an international study jointly developed by the Organization for Economic Cooperation and Development (OECD) countries to evaluate education systems worldwide. It consists of testing the mathematics, science, and reading skills and knowledge of 15-year-old students in schools in participating countries or economies. Arguably, PISA is the most comprehensive of all, both in terms of coverage (it tested students in 65 OECD and developing countries in 2012, 8 of them in Latin America), and in terms of information collected (student, household, school, institutional characteristics, and test scores in math, science, and language).

1.3 Education production function

The economics of education and education policy literature frequently use an education production function approach to analyze the relationship between school inputs, educational policies, and educational outcomes. An education production function is an economic formulation of an input-output model, in which measured input variables, such as teacher quality, are mapped to a certain measured outcome, usually student achievement test scores or test-score gains (Pelayo & Brewer, 2010). In this approach, student achievement is directly related to educational inputs, which can be either directly controlled by policymakers (e.g. characteristics of schools, teachers, curriculum), or not controlled by policymakers (e.g. families, friends, and the innate endowments and abilities of the students) (Hanushek, 2007). Therefore, an educational outcome is usually represented as a function of the educational inputs, such as in the following simplified example:

Educational Outcome = f (Individual, Household, and School characteristics)

The empirical estimation of an education production function is not simple: there are many complexities and issues that need to be considered and analyzed. First of all, despite the importance that education is given in most countries nowadays, there is no agreement about what should be the goal of education, and therefore, about which educational outcome should be used. What should students get when they attend school? What should they be able to achieve as they go through each grade, or stage? What should they be able to do once they leave the school system? As mentioned above, most people would say that students should be able to learn some basic concepts in important

areas, which allow them to move into higher grades or levels of education, or to work. Therefore, some of the most frequently used "measures" of education quality are test scores, graduation rates, drop-out rates, the percentage of students pursuing further education (like college), or educational attainment, all which are correlated with attending higher levels of education, finding a job, or earning a higher income.[6]

Achievement test scores are probably the most common educational outcome studied. In theory, achievement tests should be able to measure the knowledge and skills acquired by students in the educational process. Additionally, test scores are frequently used because of their common availability, their correlation with continuation in schooling, or because they are valued in and of themselves (Hanushek, 1979). However, there are three main problems with using student achievement as the main measures of educational output. First, achievement is a multidimensional concept, so it is important to determine what exactly each test is measuring (Koretz, 2008). Questions in achievement tests are generally selected by criteria internal to the tests, but they are not necessarily directly related to, and do not necessarily appropriately evaluate the material, knowledge, or skills valued by society (Hanushek, 1979). Second, is it difficult to differentiate in these tests what is measuring achievement (the skills and knowledge acquired from education) and what measures ability (innate skills) (Barnow, 1979). Third, achievement tests have measurement error and that biases the coefficients in the education production function.

Another important aspect that has been discussed about the use of student achievement as an educational outcome, is whether student achievement should be

[6] See for example: Hanushek, 1986, 1989; Goldhaber and Brewer, 1997; World Bank, 2008.

measured at a point in time or as a change in a period of time. What is it that we should value: What students know at one moment in time, or how much they learn from one period of time to another? The education production function can be used to address this question. It can be analyzed using a *status* approach, which means that the outcome of interest in the production function is the *level* of achievement in a specific point of time, for example, achievement in a particular test in one year. This is the most common approach because frequently only input measures for one period of time are available (Hanushek, 2007). However, the differences in achievement at one point of time may reflect differences in many other characteristics. To help mitigate this problem, the education production function can also be analyzed using a value-added approach, which means that the outcome of interest is the student's achievement gains, for example, the growth or change in the a students' test-scores between the beginning and end of an academic year, or from one year to another. Because the value-added approach measures prior or initial achievement levels, it alleviates the problem of omitting prior inputs of schools and families that are related to student achievement (Hanushek, 1979).[7]

The education production function can also be affected by data, specification or estimation problems, and consequently, to bias in the estimated effects of educational inputs (Hanushek, 1986). On one hand, a frequent problem in the estimation of education production functions is related to specification errors, such as those caused by omitted

[7] There have been a lot of recent debates about the value and usefulness of value-added models. For example, Rothstein (2010) developed falsifications tests for three typical value-added models, based on the idea that teachers cannot affect students' past achievement. He found that using value-added models, fifth grade teachers seem to have large effects on fourth grade test score gains, suggesting teacher effects that are not possible. He also found that the common measures of individual teachers' value-added fade out very quickly and are best weakly related to long-run effects. On the other hand, Goldhaber and Chaplin (2012) defend value added models, arguing that under some conditions, Rothstein's falsification test rejects value added models even when students are randomly assigned (conditional on the covariates in the model), and even when there is no bias in estimated teacher effects.

variables that can be correlated with student achievement. For example, how a student performs on a test maybe related to unobservable characteristics such as ability, or motivation. On the other hand, there are three main issues that can affect the selection and measurement of the educational inputs. First, the educational process is cumulative, that is, past inputs and innate abilities affect current educational outcomes. However, because of the unavailability of data, usually only current input measures are used, and this can lead to measurement and specification errors (Hanushek, 1986). Second, measures of innate ability, such as IQ tests, are rarely available. Arguably, when available, IQ tests can be a good measure of intelligence. However, IQ tests attempt to measure a particular type of intelligence, failing to measure intelligence in a broad sense. For example, they might not capture other aspects of intelligence, like creativity. They also fail to capture other things associated with innate abilities, like motivation. Third, in estimating the effect that an input has on an educational output, like student achievement, it is very difficult to separate the effects of the different inputs that contribute to student learning and performance, making it difficult to establish a causal relationship between resources and performance (Hanushek, 2007). For example, if policy makers provide more resources to those they considered to be most in need, higher resources could simply signal students that usually have lower achievement.

Governments in many developed and developing countries seem to be spending a lot of money and effort increasing school inputs. However, it is not clear that this will necessarily improve the quality of education. Hanushek (1989, 1997) has reviewed extensively the literature on the education production function, and he concludes that the evidence overwhelmingly supports that variations in school expenditures are not

systematically related to variations in student performance. In fact, he suggests that education policies should move away from traditional "input directed policies" to ones providing performance incentives (Hanushek, 1989). This does not necessarily mean that investing in school resources never matters. For example, after reviewing the literature, Card and Krueger (1998) conclude that there is some evidence that school resources affect earning and educational attainment. Although they recognize that there is a lot of uncertainty in the literature.

As mentioned before, for many years the Colombia government spent most of its investments in education increasing infrastructure and school inputs. This was a necessary policy to increase enrollment rates, and reach full education coverage, at least in primary grades. However, as the country's priorities move from increasing the quantity of education toward improving the quality of education, it is important to analyze which policies have been effective in increasing student achievement and improving the quality of education, and which ones have not. In some areas, such as public-private partnerships and programs that directly aim to change household characteristics, such as Conditional Cash Transfer programs, Colombia has already implemented various successful policies. This evidence, as well as the international evidence, might suggest some policies that could help improve the quality of education in Colombia, although more evidence is needed in some areas.

1.4 The impact of education policies on the quality of education

The following sections review the education literature that has analyzed the impact of school resources and other type of education policies on the quality of education, typically measured by student achievement in standardized tests, or other student outcomes such as dropout and graduation rates. Most of the literature has focused on analyzing the importance of school characteristics or resources, particularly those that can be directly affected by public policies, such as teacher characteristics and class size. Sections 1.4.1 and 1.4.2 describe the most relevant evidence with respect to the impact of teachers and class size, respectively. Section 1.4.3 looks briefly at the evidence regarding the length of the school day. Section 1.4.4 reviews the relationship between institutional characteristics and the quality of education. Section 1.4.5 discusses pedagogical models. And Section 1.4.6 looks at the impact of individual and household characteristics on student outcomes.

1.4.1 Teacher policies

The literature on the impact of teachers is probably the most extensive. Summarizing this literature, Hanushek and Woessmann (2007) concluded that there is increasing evidence that the quality of teachers is very important for student achievement. For example, Rivkin, Hanushek and Kain (2000) and Snaders and Horn (1994) found that achievement gains by students over an academic year are strongly impacted by the teachers whose classes they attend. The former study found that 7% of the total variance in test-score gains could be attributed to the differences in teachers (cited in Loeb, 2001).

However, *how* teachers affect students is still considered a "black box" as there is no conclusive evidence on which specific learning practices contribute to student achievement, or how specific and measurable teacher characteristics like experience, education, and certification affect student outcomes. Some studies have found positive effects of some of these characteristics. For example, according to Loeb's (2001) review: Ferguson (1991) found that in Texas, teacher performance on a state-wide certification exam was positively related to student outcomes; Ehrenberg and Brewer (1994) found that the selectivity of the college a teacher attends positively influenced test-score growth; Monk (1994) found a positive effect of teachers' subject area preparation on student learning; and Hakel, Keoning, and Elliot (2008) found that teachers certified by the National Board of Professional Standards (NBPS) are more effective teachers than those that are not certified. In contrast, Hanushek and Rivkin (2007) found little evidence of teacher credentials' effects on student achievement. Reviewing the literature on the topic, Hanushek (1986) concluded that teachers' education and experience are not consistently found to be associated with higher student achievement. In sum, even though the literature recognizes the importance of teachers, in general the studies have not found consistent results on what attributes of teachers signify "quality" and which ones translate into better educational outcomes for their students.

Studies about the impact of teachers on student performance in Colombia have focused on observable characteristics of teachers, such as their educational level, and the evidence is not conclusive either. Gaviria and Barrientos (2001) and Bonilla and Galvis (2011) found a statistically significant association between the average education of

teachers and a school's performance in the ICFES tests,[8] although in the former study, this effect was limited to private schools and not significant for public schools. Also using the ICFES tests, Barrera-Osorio (2003) found positive and significant effects of teachers' educational level on student achievement. Other studies also reach the conclusion that there is a positive association between "teacher quality" and student achievement in international assessments: World Bank (2006) found that schools with certified teachers have on average higher test scores in PISA 2006, while Garcia et al. (2014) found that teacher quality, as measured by student perception, is correlated with higher test scores in PISA 2009. In contrast, Woessmann and Fuchs (2005) did not find a statistically significant relationship between teachers having a professional degree or pedagogical training and student performance in the Progress in International Reading Literacy Study (PIRLS).

As mentioned above, in the previous decade, the Colombian Ministry of Education implemented a new Teacher's Statute (2002). The goal of this Statute was to "look for, attract, and keep the best professionals at the service of the public education".[9] A first evaluation of this policy found that the new Teacher's Statute had a positive impact on SABER test scores. In particular, the results showed that there was a statistically significant effect of the policy on 9th grade test scores, some positive effects on 5th grade test scores, and no effects on 11th grade test scores (Ome, 2013).

[8] ICFES is the acronym of the Colombian Institute for Educational Evaluation, but given that the agency is in charge of administering the high school exit tests, "ICFES" is also the name given to this test. In recent years, the name of the test has changed to "SABER 11", to be consistent with the rest of national tests administered by the agency: SABER 3, 5, 9 (tests 3rd, 5th, and 9th graders respectively), SABER 11 (the previous "ICFES", which tests students in 11th grade, the last high school year), and SABER PRO (tests university students in different programs).

[9] Estatuto de Profesionalización Docente, Decreto 1278 de 2002.

1.4.2 Class size

Regarding school resources and class size, the evidence generally suggests that class size can increase test scores, at least in the short run. Tennessee's STAR project, a large-scale randomized trial in Tennessee to test the impact of reducing class size, has been perhaps the major experiment in class size reduction, and the evidence from this program is frequently cited as providing support for class reduction policies. Krueger (1999) and Krueger and Schanzenbach (2002) found that a reduction of eight students in a classroom would increase average student achievement by 0.2 standard deviations, but only in the first grade that students attend a smaller class (kindergarten or first grade). Exploiting Israel's Maimonides rule, Angrist and Lavy (1999) found a negative effect of larger classes on fifth grade student test scores (between 0.089 and 0.202 standard deviations) and a smaller impact on fourth grade reading test scores (between 0.061 and 0.075 standard deviations), compared to students in the same grade assigned to smaller classes. More recently, Dynarski, Hyman, and Schanzenbach (2011) analyzed the impact of STAR on college entry, college choice, and degree completion. They found that assignment to small classes in the STAR project increased the probability of attending college by 2.7 percentage points, increased the likelihood of earning a college degree by 1.6 percentage points, and shifted students toward high earning fields, such as Science, Technology, Engineering, and Mathematics (STEM), business, and economics. They also confirmed the previous finding that the impact of STAR on test scores faded out by middle school, but they found that test scores at the time of the experiment are an excellent predictor of long-term improvements in postsecondary outcomes. Another frequently cited study analyzed the case of Israel, where the "Maimonides rule" imposes

a maximum class size of 40 in public schools. This allowed the authors to construct instrumental variables estimates of the effects of class size, and find that reducing class size increased test scores for fourth and fifth graders, but not for third graders (Angrist & Lavy, 1999).

However, the evidence on class size, and particularly that of the STAR program, has been controversial. For example, Hanushek (1999) argued that the STAR studies have significant design and implementation issues that make the magnitude of the treatment effects uncertain. For him, the "positive" effects of a reduction in class size are limited to very large reductions, and he finds no support for the impacts of smaller reductions in class size (i.e. resulting in class sizes greater than 13-17) or for the reductions in later grades. Looking at variations in class size caused by variations in the population, Hoxby (2000) also concluded that class size does not have a statistically significant effect on student achievement, even for reductions of class size of about 10 percent.

In developing countries, the evidence on the impact of class size has generally found fairly modest impact on student test scores. For example, a program in Kenya that hired an additional local teacher to teach a first grade class, which effectively reduced class size by about half (from 82 to 44), found that a 10-student reduction in class size was associated with an increase of 0.042 to 0.064 standard deviation in test scores. Given that there was an observed class reduction of about 40 students, the program could have increased test scores by 0.17 to 0.26 standard deviations (Duflo, Dupas, & Kremer, 2012).

In Colombia, the evidence on the impact on class size is scarce. Analyzing schools in Bogota, Gaviria and Barrientos (2001) found that class size had a statistically significant effect on student achievement in the ICFES tests, although the impact of increasing the number of teachers seem to be counter-productive beyond some point: an increase of one teacher per 100 students increased test scores by 5 percentage points, although an increase in the ratio beyond 0.1 was associated with a lower average test score for the school. Additionally, the effect seems to be more important for public schools: a 0.03 increase in the teacher student ratio was associated with 10 percentage points higher test scores in public schools, and only 5 percentage points in private schools. In contrast, Woessmann and Fuchs (2005) find no statistically significant impact between smaller classes and performance in the PIRLS tests for Colombian students.

1.4.3 Length of the school day

The international evidence suggests that instructional time is an important input in the education process: the length of the school year and the length of the school day seem to have a positive impact on student achievement and other student outcomes. This is the central topic of Essay 2 of this Dissertation ("Do longer school days improve student achievement? Evidence from Colombia"). Essay 2 covers the international evidence on this topic in detail, so in this section, I will only briefly refer to the evidence on the impact on longer school days in Colombia.

In Colombia, primary and secondary students attend schools that have either a complete (7-hour), or a morning or afternoon (4-hour) schedule. Most students attending public schools attend schools with the shorter schedules. There have three recent studies

estimating the impact of longer school days on different outcomes. Bonilla-Mejia (2011) found that schools with a complete schedule have average ICFES test scores that are 2.5 percentage points higher than those attending half-day schedules. Garcia and Fernandez (2011) found that a complete schedule reduced the probability of early dropout and grade repetition. In Essay 2 of this dissertation, I find that among schools that switched schedules, cohorts that attended complete schedules have test scores that are about one tenth of a standard deviation higher than cohorts that attended half day schedules. After analyzing the international and national evidence supporting the benefits of extending the length of the school day, Barrera-Osorio, Maldonado, and Rodriguez (2012) recommended extending the length of the school day as an important policy to improve the quality of education in Colombia.

1.4.4 Institutional characteristics

Some authors have highlighted the importance of institutional characteristics, like the level of school or institutional autonomy, accountability, the type of school administration, and the methods of assessment (Vegas & Petrow, 2007). Fuchs and Woessmann (2004) have found that these institutional factors explain about one-fourth of the variation in test scores between countries. Analyzing the impact of institutional characteristics, Hanushek and Woessmann (2007) concluded that some of the variables that are correlated with student achievement are: teachers' influence in the instructional content, school autonomy to hire staff and decide on wages, limited influence of unions in the study programs, centralized control of budget, centralized assessments, evaluation of student achievement through exams, homework and meetings, competition with

private institutions, and stimulating parents to get involved in the learning processes. Summarizing the literature, the World Bank (2008) concluded that school and institutional autonomy are factors that significantly affect the decision making process and that it is important that this autonomy is matched with a centralized evaluation of student achievement and institutional capacity. Research has also shown the importance of competition between institutions, achievement accountability, and efficient governance.

A large debate in the education literature refers to who should provide education: the public or the private sector. It is known that the social returns to education are much larger than the private returns, and therefore, the government needs to intervene (Friedman, 1962). Before, this meant that the government had to finance and provide education. However, in some countries, governments have allowed the private sector to be involved in the provision of education services in different ways. In the United States, for example, charter schools receive public funding but operate independently, having great autonomy over curriculum, instruction, and the operation of the school. The evidence on charter schools is mixed. A study by CREDO (2009) found that only 17% of charters schools had better results than the local public schools, while 37% of charter schools had worse results than the students in traditional public schools. Zimmer et al. (2009) also find that charter schools in most of the states in their study have test scores that are on average neither better nor worse than the traditional public schools. The U.S. has also experimented with vouchers, which are government-financed entitlement that students and their parents can use to attend any school, including private schools.

Colombia has had several successful experiences with the private provision of education. In 1991, the *Programa de Ampliación de Cobertura en Educación Secundaria* (PACES) was implemented to provided vouchers for 125,000 poor students that covered about half the cost of attending a private secondary school. Since vouchers were allocated by lottery, this enabled researchers to evaluate the program using randomization methods. Angrist et al. (2002) found that lottery winners were about 10 percentage points more likely to finish 8th grade than lottery losers, and their test scores were on average 0.2 of a standard deviation higher. The program also had long-term impacts: students that received the vouchers had a larger probability of graduating from high school, performed better in math and language tests (Angrist et al., 2006), and had a 10% larger probability of attending higher education, a 25% larger probability of staying in higher education, and higher wages (Saavedra et al., 2012).

The country has also experimented with contracting out public schools to private institutions, which sign an agreement with the government to manage and operate public schools in exchange for public resources (Patrinos, Barrera-Osorio, & Guaqueta, 2009). The city of Bogotá implemented the Concession Schools program in 1991. This program allowed a group of private schools and universities to provide education for low-income students under a contract with the public education system. The evidence suggests that this model has been widely successful: dropout rates are lower and ICFES test scores are higher in concession schools than in similar public schools (Barrera-Osorio, 2007). Bonilla-Angel (2011) also found that the concession schools in Bogotá had an impact of between 0.2 and 0.6 standard deviations in language and math ICFES test scores, respectively.

1.4.5 Pedagogy and education models

In contrast to the extensive literature on the impact of school resources on the quality of education, the evidence on pedagogical methods and education models is very scarce. This is an area where a lot more research is needed in order to understand what works. In Essay 3 of this Dissertation (The effectiveness of multigrade classrooms: Evidence from Colombia's New School model), I analyze the impact of *Escuela Nueva* (New School), and education model that has been often cited as best practice in terms of rural schools reform (McEwan, 2008). *Escuela Nueva* is characterized by multigrade classrooms (i.e., one instructor teaches students in various grades in the same classroom), a child-centered curriculum, flexible systems of grading and promotion, intensive teacher training, and parental involvement. According to the World Bank (2008), the model "has improved student achievement in rural areas by enabling students to progress to a flexible curriculum, by engaging them with active pedagogy supported by teachers training, and by adapting to local needs through democratic decision-making and community engagement." In Essay 3, I find that among schools that switched models between 2002 and 2005, the cohorts of 5^{th} grade students exposed to Escuela Nueva have on average 0.135 of a standard deviation higher language test scores than cohorts exposed to other models, while there is no statistically significant impact on switching to EN for 9^{th} graders.

1.4.6 Individual characteristics and household policies

Other educational inputs that have been analyzed in the literature are individual and household characteristics. Individual characteristics refer to the things that the student himself or herself brings to the educational process. In this category, the literature has analyzed how inherent characteristics such as age, gender, and the health status of the student, are associated with student learning and achievement. For example, Datar (2006) found that delaying kindergarten entrance age for some months improved student achievement. And the World Bank (2008), summarizing the literature on gender, concluded that "research indicates that girls tend to do better in language tests, while boys tend to do better in math and science" (p.37). Students also have some innate endowments, a natural capacity to learn and different levels of cognitive development, which affect their learning in school (Hanushek, 2007). However, as mentioned before, measures of innate ability are frequently not available.

Students also learn to speak during their childhood and develop different levels of early literacy, and this determines their ability to learn and express what they learn later in life (World Bank, 2008). Therefore, a strand of the literature has analyzed the impact of early childhood development (ECD) on educational outcomes later in life. For example, for the United States, Galinsky (2006) found that children who participated in high quality ECD programs had a higher cognitive development and achievement. Schutz (2009) compared various countries and found that the association of pre-primary attendance with test scores is larger for countries with higher per-pupil spending in pre-primary education, larger shares of children attending privately managed pre-primary institutions, higher relative pay, and higher levels of training of pre-primary teachers.

On the other hand, household characteristics usually refer to socio-demographic characteristics, such as parental education, income, and family size (Hanushek, 2007). Recent literature has also given a lot of attention to the issue of educational resources at home, like books and computers. After the Coleman Report (1966) reported the importance of family background on student achievement in the United States, a lot of studies have followed supporting this as the most important determinant (World Bank, 2008). For example, Patrinos and Psacharopoulos (1995) found that the family socioeconomic variables that are related to poverty are strongly (negatively) associated with student achievement. The World Bank (2005) also found that there are strong associations between parent involvement in the education process and student achievement in math test scores.

In recent decades, governments in many developing countries have implemented programs that aim to affect in some way the characteristics of the student or the household. The most common of these has been Conditional Cash Transfer programs (CCT). CCTs provide financial support to poor families, conditional on children attending school regularly and obtaining basic medical care. There have many studies analyzing the impact of CCTs in different countries. Most of them have found strong evidence of the impacts that these programs have on increasing enrollment rates, and reducing dropout rates. For example, Skoufias (2005) found that Progresa (now known as Oportunidades), Mexico's pioneer CCT program, increased enrollment rates by 7.2 to 9.3 percentage points for girls, and 2.5 and 5.8 percentage points for boys in secondary school, and Behrman, Parker and Todd (2005) found that the program significantly reduced dropout rates by 5.9 to 11.6 percentage points for youth over 10. On the other

hand, the evidence of the effects of CCTs on student achievement are limited and inconclusive (Garcia & Hill, 2009).

Familias en Acción, Colombia's CCT program, has been widely studied. Evaluating the impacts of the program, Attanasio et al. (2005; 2006) found that the program increased school enrollment by 5.5 percentage points for urban youth between 14 and 17 years old, and 7 percentage points for those in rural areas. For younger children the impact on enrollment ranged from 1 to 3 percentage points. Garcia and Hill (2009) looked particularly at the impact on grade retention and student test scores. They found that for children 7 to 12 years old who live in rural areas, the program has an impact of 3.8 points in math test scores (40% of a standard deviation) and 3.2 increase in language test scores (47 percent of a standard deviation), although the impact was not significant for those in urban areas. They also found a small, but negative effect, on school achievement of adolescents who lived in rural areas. Baez and Camacho (2011) analyzed the impact of *Familias en Acción* on human capital formation in the long term, and found that children that participated in the program were 4 to 8 percentage points more likely to finish high school, particularly girls and those living in rural areas. However, they do not find a statistically significant impact of the program on the ICFES test scores.

1.5 Conclusions and policy recommendations

The review of the existing research suggests that provision of education is generally an inefficient process. If one assumes that schools wish to maximize student achievement, the evidence suggests that in fact, schools are economically inefficient, because they pay for attributes that are not systematically related to achievement

(Hanushek, 2007). As the evidence discussed in this essay suggests, this happens in Colombia (as well as in many other developing countries), where for many years the government has been significantly increasing education expenditures to increase the quantity of school resources, although these have not been found to have an impact on student achievement, or more generally, have not improved the quality of education. This implies that the government has been investing its scarce resources in a system that is basically failing to offer the quality of education that children and adolescents need in order to succeed later in life. Therefore, a deep change is needed in the education system if the country wishes to remain in the path of economic growth, to become competitive in the global economy, and to reduce poverty and provide equal opportunities to all its citizens.

The literature and the evidence presented in this Essay, as well as in Essays 2 and 3 of this Dissertation, provide a discussion of the different policies and programs that might be useful to improve the quality of education in Colombia. In particular, I think that there are four types of policies that are very important for Colombia:

First, the international and national literature has found strong evidence on the importance of having good teachers. However, the way in which the Colombian teacher's labor market was organized for many years did not provide the incentives to attract and retain good teachers. For example, high school students with the worst achievement levels are the ones pursuing education degrees or a teaching career; teachers are promoted automatically based on their education and experience (and not on their performance); and "bad" teachers can remain in the system without any type of accountability. Consequently, it is important to promote an education system where the

best high school students are attracted to professional programs in education and to become teachers, to improve the quality of professional education programs and pedagogical training, and to create incentives for teachers that are aligned with those of the education system. This means that teachers should be accountable for their performance: they should not be promoted only on the basis of their education and experience, and they should be permanently evaluated so that only the good teachers remain in the system (Barrera-Osorio, Maldonado, & Rodriguez, 2012; Garcia et al., 2014). The new Teacher's Statute of 2002 made some progress in all these aspects, but most of these changes still need to be implemented.

Second, it is important to extend the length of the school day in all public schools, so that all students can attend schools for a complete schedule. The evidence, including the results from Essay 2 of this Dissertation, has shown that students attending complete schedules do better than students attending half schedules. Additionally, the international evidence has found that extending the length of the school day could also have other type of positive non-academic impacts, such as reducing crime, drug consumption, or teen pregnancies. Extending the length of the school day might require a large investments, but the possibility of improving the quality of education and improving other outcomes such as the ones mentioned above, makes this a policy that it worthwhile implementing. In order to have a positive impact, it is also important that the increase in the length of the school day goes together with improving the quality of teachers.

Third, the successful experiences that Colombia has had with public-private partnerships, particularly with the PACES voucher program, and Bogota's concession schools, suggest that other cities could experiment with this type of partnership with the

private sector. However, as the evidence on charter schools in the U.S. has shown, there might be a lot of heterogeneity in the quality of the services provided in privately-managed schools. Therefore, it is important that any expansion of the role of the private sector in the provision of education is closely supervised by the government, so that bad schools are quickly identified and closed, or their administration replaced.

Finally, given the positive impacts that *Familias en Acción*, Colombia's CCT program, has had on enrollment rates, graduation rates, and other education outcomes, it is important to continue with this program. However, it is important to adjust the program so that it also incentivizes student achievement. For example, some additional conditions could be included to promote student achievement, and the program could be complemented with other programs that promote food security and health outcomes (Camacho, 2013).

Despite the existing evidence, more research needs to done in all of the areas discussed in this essay. In particular, more studies are needed that identify the factors that are associated with student achievement, and more impact evaluations that rigorously analyze the impact that education policies have on student achievement in Colombia. One area in particular where a lot more research is needed is on the effectiveness of pedagogical practices, and education models. That is, it is very important to better understand the processes that take place inside each school, and to identify the type of pedagogical practices and education models that are more effective in improving the quality of education in Colombia. Essay 3 in this Dissertation contributes to this area, but much more research is needed.

Finally, in order to generate the strong evidence needed, more and better education data is needed. This is necessary to improve the quantity and the quality of the research, but most importantly, to provide evidence that can better inform and improve the quality of education policies so that they can effectively improve the quality of education in Colombia.

1.6 References

Angrist, J., & Lavy, V. (1999). "Using Maimonides' rule to estimate the effect of class size on scholastic achievement." Quarterly Journal of Economics, 114 (2), 533-575.

Angrist, J., Bettinger, E., Bloom, E., King, E., & Kremer, M., (2002). Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment. *American Economic Review, 92*(5), 1535-58.

Angrist, J., Bettinger, E., & Kremer, M. (2006). Long-term educational consequences of secondary school vouchers: Evidence from administrative records in Colombia. *American Economic Review, 96*, 847-862.

Baez, J., & Camacho, A. (2012). Assessing the Long-term Effects of Conditional Cash Transfers
on Human Capital. Evidence from Colombia. The World Bank Policy Research Working Paper 5681.

Barnow, B. (1979). Theoretical issues in the estimation of production functions in manpower programs. In Evaluating Manpower Training Programs, Research in Labor Economics, Supplement 1, 295-338.

Barrera-Osorio, F. (2006). The impact of private provision of public education: Empirical evidence from Bogota's Concession Schools. The World Bank Policy Research Working Paper 4121.

Barrera-Osorio, F. (2003). Decentralization and Education: An Empirical Evaluation. Ph.D Thesis, University of Maryland College Park.Retrieved from ProQuest Dissertations and Theses.

Barrera-Osorio, F., Bertrand, M., Linden, L., & Perez-Calle, F. (2011). Improving the Design of Conditional Transfer Programs: Evidence from a Randomized Education Experiment in Colombia. *American Economic Journal: Applied Economics, 3*(2*),*167-195.

Becker, G. (1964). Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education, 3rd edition. The National Bureau of Economic Research. University of Chicago Press. Chicago and London.

Behrman, J. R., Parker, S. W., & Todd, P. (2005). Long-term impacts of the Oportunidades conditional cash transfer program on rural youth in Mexico (Discussion Paper 122). Goettingen: Ibero-American Institute for Economic Research.

Bonilla-Angel, J.D. (2011). Contracting out public schools and academic performance. Evidence from Colombia (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses.

Bonilla-Mejia, L. (2011). Doble jornada escolar y calidad de la educación in Colombia. Documentos de trabajo sobre Economía Regional, No. 143.

Bonilla, L., & Galvis, L.A. (2011). Profesionalización docente y calidad de la educación en Colombia. Documentos de trabajo sobre economía regional. Banco de la Republica No. 154.

Bos, M.S., Ganimian, A., & Vegas, E. (2013). América Latina en PISA 2012: ¿cómo les fue a la región?

Center for Research on Education Outcomes (CREDO) (2009). Multiple Choice: Charter School performance in 16 States.

Camacho, A., (2012). Familias en Acción: un programa con alcances adicionales a la formación de capital humano. Universidad de Los Andes. Notas de Política No. 12.

Card, D., & Krueger, A. (1998). School resources and student outcomes. The ANNALS of the American Academy of Political and Social Science, 559, 39.

Coleman, J. (1966). Equality of educational opportunity. Washington: US. Government printing office.

Datar, A. (2006). Does delaying kindergarten entrance give children a head start? *Economics of Education Review*, *25*, 43-62.

Departamento Administrativo Nacional de de Estadística (DANE) (2013). Pobreza en Colombia.

Duflo, E, Dupas, P., & Kremer, M. (2012). School governance, teacher incentives, and pupil-teacher ratios: Experimental evidence from Kenyan primary schools. National Bureau of Economic Research WorkingPaper No. 14475. Cambridge, MA.

Dynarski, S., Hyman, J., & Schanzenbach, D. W. (2011). Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion. NBER working paper No. 17533.

Friedman, M. (1962). Capitalism and Freedom. Chicago: University of Chicago Press.

Fuchs, T., & Woessmann, L. (2004). What accounts for the international differences in student performance? A re-examination using PISA data. Working Paper No. 1235. Center for Economic Studies and Ifo Institute for Economic Research. Munich, Germany.

Galinsky, E. (2006). "The economic benefits of high-quality early childhood programs: what makes the difference?" Families and Work Institute.

García, S., Fernández, C., & Weiss, C. (2012). Does lengthening the school day reduce the likelihood of early school dropout and grade repetition: Evidence from Colombia. Paper presented at the Population Association of America 2012 Annual Meeting, San Francisco, CA, May 3-5, 2012.

García, S., Maldonado, D., Perry, G., Rodríguez, C., & Saavedra, J.E. (2014). Tras la excelencia docente. Como mejorar la calidad de la educación para todos los colombianos. Fundación Compartir.

Gaviria, A., & Barrientos, J. (2001a) Características del plantel y calidad de la educación en Bogotá. Coyuntura Social, 25.

Gaviria, A., &Barrientos, J. (2001b) Determinantes de la calidad de la educación en Colombia. Archivos de Economía, No. 159.

Glewwe, P., Hanushek, E., Humpage, S., & Ravina, R. (2011). School resources and educational outcomes in developing countries: A review of the literature from 1990 to 2010. National Bureau of Economic Research Working Paper 17554, Cambridge, MA.

Hakel, M.D., Koening, J.A., & Elliot, S.W. (Eds) (2008). Assessing accomplished teaching: Advanced-level certification programs. Washington, DC. National Academies Press.

Hanushek, E., & Wossermann, L. (2007). Education quality and economic growth. World Bank. Washington, DC.

Hanushek, E. (1979). Conceptual and empirical issues in the estimation of educational production functions. *The Journal of Human Resources, 14*(3), 351-388.

Hanushek, E. (1986). The Economics of Schooling: Production and Efficiency in Public Schools. *Journal of Economic Literature*, 24, 1141-1177.

Hanushek, E. (1989). The impact of differential expenditures on school performance.

Hanushek, E. (1997). Assessing the effects of school resources on student performance: An update.

Hanushek, E, & Rivkin, S (2007). "Pay, Working Conditions, and Teacher Quality". *The Future of Children, 17*(1), 69-86.

Hanushek, E., & Wossermann, L. (2007). Education quality and economic growth. The World Bank. Washington, DC.

Hoxby, C. (2000). The effect of class size on student achievement. New evidence from population variation. *Quarterly Journal of Economics, 115*(4), 1239-85.

Hoxby, C., & Murarka, S. (2009). Charter school in New York City: Who enrolls and how they affect their students' achievement. NBER Working Paper 14852.

Koretz, D. (2008). Measuring up. What educational testing really tells us. Harvard University Press. Cambridge, Massachusetts.

Krueger, A. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics*, 114, No.2 (May):497-532.

Krueger, A., & Schanzenbach, D.W. (2002). Would Smaller Classes Help Close the Black-White Achievement Gap? In J. Chubb, J., & T. Loveless (Eds.), *Bridging the Achievement Gap*. Washington, DC: Brookings Institution Press.

Lavy, V. (2010). Do differences on school's instruction time explain international achievement gaps in Math, Science and Reading? Evidence from developed and developing countries. NBER Working paper series No. 16227.

Loeb, S. (2001). Teacher Quality. Its enhancement and potential for improving pupil achievement. In D. H. Monk, H. J. Walderg, & M. C. Wang (Eds.), *Improving Educational Productivity.* Greenwich, CT: Information Age.

McEwan, P. (2008). Evaluating multigrade school reform in Latin America. *Comparative Education*, *44*(4), 465-483.

Murnane, R. (1975). The impact of school resources on the learning of inner city children. Cambridge, Massachussets: Balinger Publishing Company.

OECD (2013). PISA 2012 Results: What students know and can do (Volume 1): Student performance in Mathematics, Reading, and Science, OECD Publishing.

OECD (2010). PISA 2009 Results: What students know and can do. Student performance in Reading, Mathematics and Science, OECD Publishing.

Ome, A. (2013). El estatuto de profesionalización docente: una primera evaluación. Cuadernos Fedesarrollo No. 43.

Patrinos, H., & Psacharopoulos, G. (1995). Educational performance and child labor in Paraguay. *International Journal of Educational Development, 15*(1), 47-60.

Patrinos, H., Barrera-Osorio, F., & Guaqueta, J. (2009). The Role and Impact of Public-Private Partnerships in Education. Washington, DC: The World Bank.

Pelayo, I., & Brewer, D.J. (2010). Teacher Quality in Education Production. In D. Brewer, & P. McEwan (Eds.), *Economics of education* (178-182). Amsterdam, Holland: Elsevier.

Rice, J. K., & Schwartz, A. (2008) Towards an Understanding of Productivity in Education. In H. Ladd & E. Fiske (Eds.), *Handbook of Research in Education Finance and Policy* (131-145). New York, NY: Routledge.

Rivkin,S., Hanushek, E., & Kain, J. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 4147-58.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay and student achievement. *The Quarterly Journal of Economics, 125*(1), 175-214.

Saavedra, J.E., Bettinger,E., Kremer, M., & Kugler, M. (2012).Collegiate and Labor Market effects of Vouchers for Private schooling in Colombia (RAND mimeo).

Schutz, G. (2009). Does the quality of pre-primary education pay off in secondary school? An international comparison using PISA 2003 (Working paper). Institute for Economic Research at the University of Munich.

Schanzenbach, D.W. (2010). The Economics of Class Size. In D. Brewer & P. McEwan (Eds.) (2010). *Economics of education*. Amsterdam, Holland: Elsevier.

Skoufias, E. (2005). Progresa and its impacts on the welfare of rural households in Mexico (Research Report 139). Washington D.C.: International Food Policy Research Institute.

Summers, A., & Wolfe, B. (1977). Do Schools Make a Difference? *American Economic Review, 67* (4), 639-652.

Vegas, E., & Petrow, J. (2007). Raising student achievement in Latin America: the challenge for the 21st century. Washington: DC, The World Bank.

Woessmann, L., & Fuchs, T. (2005). Families, schools and primary-school learning: evidence for Argentina and Colombia in an international perspective. *Applied Economics, 42,* 2645–2665.

World Bank (2008). Quality of education in Colombia. An analysis and options for a policy agenda (Report No. 43906-CO).

Zimmer, R., Gill, B., Booker, K., Lavertu, S., Sass, T., & Witte, J. (2009). Charter school in eight states: Effects on achievement, attainment, integration and competition. RAND.

Essay 2

Do longer school days improve student achievement? Evidence from Colombia

Abstract

This paper analyzes the impact of longer school days on student achievement in Colombia, where primary and secondary students attend schools that have either a complete (7-hour) or a half-day (4-hour) schedule. Using test score data from 5th and 9th graders in 2002, 2005, and 2009, along with school administrative data, this study identifies the effect of longer school days by implementing a school fixed effects model. The main model compares variation in average test scores across cohorts for schools that switched from a complete schedule to a half schedule and vice versa. I find that among schools that switch schedules between 2002 and 2009, the cohorts exposed to complete schedules have test scores that are about one tenth of a standard deviation higher than cohorts that attended half schedules. The impact of a complete schedule is larger for math test scores than for language test scores, and it is larger for 9th grade test scores than for 5th grade test scores. Effects are largest among the poorest schools in the sample and those in rural areas. The results suggest that lengthening the school day may be an effective policy for increasing student achievement, particularly for the lowest-income students in Colombia and other developing countries.

Keywords: quality of education, student achievement, instructional time, school schedule.

JEL codes: I21, I28, H43.

2.1 Introduction

There are large differences in the time students spend in school, both across and within countries. Are these variations in schooling contributing to student achievement gaps and education inequalities? It seems to be a widely held belief that the more time students spend in school, the more they will learn, and so the number of hours (length of the school day) and the number of school days (length of the school year) are directly and positively related to student learning and therefore, to student achievement. While there is a large literature that provides evidence about the effect of school inputs in the education production function, there is only limited evidence on the effects of instructional time on student outcomes, and even less evidence showing that interventions that increase the length of the school year or the school day will improve learning outcomes. Overall, research supports the idea that additional instructional time has a positive impact on student achievement and other student outcomes. [10] However, there are methodological difficulties in isolating the impact of instructional time that raise questions about the strength of much of this evidence.

This study contributes to the literature by analyzing the causal link between longer school days and student achievement in Colombia, using a school fixed effects identification strategy that mitigates some of the most critical selection and endogeneity problems found in the literature.

In Colombia, some schools offer a full 7-hour school day, known as a "complete" schedule, while others offer two separate 4 or 5 hour shifts, known as "half" schedules.[11] The additional time in the former is usually devoted to extra hours of instruction. Cross-

[10] I present a detailed review of the literature in section 2.3.
[11] In half schedule schools, two different groups of students attend the same school, with one group attending the morning shift, and the other one the afternoon shift.

sectional analyses comparing student test scores in schools with complete and half schedules will likely be biased since schools that have a complete schedule may have unobservable characteristics that are related to student achievement (e.g., more educated parents, or more motivated principals and teachers). However, in Colombia, schedules are determined at the municipal level according to policy and projected enrollment, and thousands of schools switched from half to complete (and vice versa) in the time period I study. I can therefore compare the test scores of cohorts of students in the same school before and after the switch to (or from) a complete schedule. This strategy can effectively control for all unobserved time-invariant characteristics of the school that may bias cross-sectional estimates, although time-varying unobservables may remain.[12]

I draw on data from the 2002, 2005, and 2009 SABER to implement this approach. The SABER is a nationwide standardized test administered every three years in all primary and secondary schools in Colombia. It evaluates 5th and 9th grade students' skills in both mathematics and language. I merge these data with administrative data on all primary and secondary schools in Colombia.

Policies aimed at increasing the length of the school day continue to be widely debated in Colombia, as well as in other countries. Arguing that extending the school schedule might have a positive impact on student outcomes, some countries, such as Chile and Uruguay, have implemented a complete schedule in all public schools that previously had a half schedule. Colombia implemented this same reform in 1994, but it was not enforced and was later rescinded, mainly because of the increasing demand for school spots that required having two half-day schedules in some schools. However, in

[12] Family mobility may be a particular concern if students switch schools after the change in schedule. However, few families in Colombia move because of schools and fewer still move before their children reach the 5th and 9th grades. I address these concerns in section 2.4.

recent years, some cities have starting piloting interventions to increase the length of the school schedule. As many countries and cities continue to experiment with this type of programs and policies, this study aims to contribute to the debate about the effectiveness of increasing the length of the school day.

This paper contributes to the literature on instructional time by analyzing the impact of time in school given by the length of the school day rather than the length of the school year, where a substantial portion of the literature has focused. It also helps fill the gap in the literature on the effect of instructional time in a developing country, and especially in Colombia, where evidence is particularly scarce.

To the best of my knowledge, this is one of only three studies that estimates the impact of longer school days in Colombia, and the first one to look at the impact on student achievement for 5^{th} grade and 9^{th} grade students in the country: Garcia, Fernandez, and Weiss (2012) look at the impact on dropout and grade repetition of 1^{st} and 2^{nd} grade students, while Bonilla-Mejia (2011) looks at the impact on student achievement in a high school exit exam (ICFES). This paper contributes to the literature on longer school days in Colombia in three main ways. First, because there might be different impacts of having longer schools days for students of different ages or grades, this study analyzes the impact on students attending different grades than those that have been studied before. Second, by analyzing 5^{th} grade and 9^{th} grade students in the same study, using the same type of data and methodology, this study allows comparing the differences in the magnitude and statistical significance of the impact of longer schools days between students attending primary and secondary grades. Third, by using a school fixed effects methodology that compares different cohorts of students within a school,

this study is able to mitigate many of the selection concerns that are common when comparing differences between schools.

I find that among schools that switch schedules between 2002 and 2009, the cohorts exposed to complete schedules have test scores that are about one-tenth of a standard deviation higher than cohorts that attended half schedules. The impact of a complete schedule is larger for math test scores than for language test scores (e.g., 0.138 v. 0.110 of a standard deviation respectively, for 9th graders) and it is larger for 9th grade test scores than for 5th grade test scores (e.g., 0.138 v. 0.082 of a standard deviation respectively, for math test scores).

Using different subsamples to estimate the impact of a complete schedule, I find that the effect of attending a complete schedule is largest among rural schools and among the very poorest schools in the country. Overall, the results suggest that lengthening the school day may be an effective policy for increasing student achievement, particularly for 9^{th} grade students, and for the lowest-income students in Colombia.

The rest of the paper is organized as follows: Section 2.2 describes the background on school schedules in Colombia. Section 2.3 provides a brief conceptual framework and reviews the most relevant literature on instructional time. Section 2.4 discusses the empirical strategies. Section 2.5 describes the data. Section 2.6 presents results, and Section 2.7 concludes.

2.2 Background

Over the last 20 years policymakers in Colombia have continually debated the issue of the appropriate length of the school day. In 1994, the government of Colombia passed the General Education Law (Law 115), which established that all public schools should have a single complete (7-hour) schedule. However, plans to implement this law did not begin until years later and were completely abandoned by 2002. This happened partly due to a substantial increase in demand for public schooling that was difficult to meet because of low physical capacity (fewer schools than the ones needed to meet the demand) and scarce resources (low capacity to hire enough administrators and teachers for a full day).

More recently, there have been proposals to establish a longer schedule in all public schools in the country, coming both from academia (Barrera-Osorio, Maldonado, & Rodriguez, 2012), and from policymakers, particularly in two of the largest cities in the country, Bogotá and Cali, where pilot programs to increase the length of the school day are being implemented. In April 2012, Bogotá city announced a pilot to turn 25 public schools into complete schedule schools, and Cali's Mayor announced that 2 schools in the city (encompassing approximately 1,000 students) will start having a complete schedule rather than the previous half-day schedules, and that this new standard will be implemented in the entire city during his term. In both cases, however, the additional time would be devoted to civics, sports, and cultural activities, rather than additional instructional time.[13]

[13] Although it is not possible to know what students are really doing during the additional hours in school from the data, in the current situation it is very likely that students attending schools with a complete schedule are receiving additional instructional time.

The argument for implementing a longer school day is that more time spent in school will translate into better outcomes for students. However, except for a few recent studies that have found a positive impact on student achievement (Bonilla-Mejia, 2011) and other student outcomes (Garcia, Fernandez, and Weiss, 2012), there is not a lot of evidence suggesting a causal link between the length of the school day and student outcomes in Colombia.

Establishing the impact of having a complete schedule on student outcomes is methodologically difficult. If students were randomly assigned to different school schedules, in principle one could analyze the impact of attending a complete schedule by estimating the difference in student achievement between students attending a complete schedule and students attending a half-day schedule. However, a student's application and allocation to a school, and the decision process on which type of schedule a school has, is neither simple nor random.

First, the Secretary of Education of each municipality assesses the expected demand for school spots for the following year, and the capacity that schools in that municipality have to offer school spots (i.e., supply). Depending on that information, he or she determines whether it might be necessary to have two half-day schedules (a morning and an afternoon shift) instead of a complete schedule in certain schools, in order to increase the supply of school spots (Ministerio de Educación Nacional, 2006). As a result, the process of deciding the schedule for the municipality's schools is a function of many variables. These include the number of people who are at school age in the municipality, the actual demand for school spots (i.e.; the number of students already enrolled in the education system, and the applications filled out by parents for each

school), the number of available places in each grade in each school, and the amount of education resources in the municipality (e.g., to hire more teachers). In some cases, the Secretary of Education might decide to implement half-day schedules in schools with a higher demand in order to be able to accept more students in those schools, and to implement complete schedules in schools with a lower demand. If the demand for each school is related to its quality (e.g., higher demand for relatively higher quality schools, and lower demand for relatively lower quality schools), this could possibly cause a downward bias in Ordinary Least Squares (OLS) estimates of the impact of a complete schedule, because schools with a complete schedule will have other unobserved characteristics related to lower test scores.

Second, in Colombia, parents cannot directly choose a public school for their children: When they want to enroll their children for the first time into a school, or want to transfer their children into a different school because they moved to a new area, they need to complete an application with the municipality's Secretary of Education. In this application, parents are able to express their preference for two schools, but they cannot express their preference for a particular schedule. The Secretary of Education then makes the decision on where to assign the new or transferred students whose parents have applied for a school spot, based on the demand and supply of schools spots for each grade. Ideally, this decision matches the parents' expressed preferences, but it is possible that children are assigned to a school that is not in their parents' preferences if there are not enough schools spots in the parents' preferred schools.[14]

[14] If that is the case, the municipality will try to allocate the students to the school closest to their place of residence. However, it might be possible that they are assigned to a school in the other side of the city if that is the only school where there are available spots, although this is not frequently the case.

2.3 Conceptual framework and literature review

i. Impact of the length of the school year

The literature on the impact of instructional time can be classified in two main categories: One that analyzes the impact of differences in the length of the school year, and one that looks at the impact of longer school days or more instructional time in a specific subject per day.

Most of the literature has focused on analyzing the impact of the school year's length on student achievement or other outcomes. For example, analyzing the effect of schooling on labor market outcomes, Margo (1994) found evidence that historical differences in the length of the school year accounted for a large fraction of the differences in earnings between black and white workers at the end of the 20th century. Lee and Barro (2001) compared differences in the length of school year across countries and found that more time in school improved math and science test scores, but it seemed to lower reading scores. Eren and Milliment (2007) studied the variation in the length of the school year and the average duration of classes across states in the United States. The results indicated that the school year length, and the number and average duration of classes, affected student achievement. However, the direction and magnitude of the effects were not homogeneous across the test score distribution: test scores in the upper tail of the distribution benefitted from a shorter school year, while a longer school year increased test scores in the lower tail.

Some authors have taken advantage of natural experiments to analyze the impact of time spent in school and student achievement. For example, Marcotte (2007) and

Hansen (2008) exploited year-to-year differences in the length of the school year that occur as a result of the weather. They compared how schools in two states in the United States performed in years when there were frequent cancellations due to snow days, to how they performed in years when the winter was relatively mild. They found that an additional 10 days of instruction resulted in an increase in student performance on state math tests of roughly 0.2 standard deviations.

Related to the analysis of the length of the school year, a strand of the literature has analyzed the *summer learning loss*. This concept is based on the idea that the summer break might have a negative impact on student learning, because of a lack of cognitive stimulation, a break in the rhythm of instruction, and forgetting that requires a significant amount of review when students return for the start of the academic year (Cooper et al., 2010). For example, a meta-analysis of the impact of summer vacation on test scores found that summer learning loss was equivalent to at least one month of instruction (Cooper et al., 1996). An alternative to mitigate the impact of the summer learning loss is extending the length of the school year, or switching to year-round calendars with the same number of instructional days but spreading the days of instruction evenly across the year. The impact of this type of interventions is not clear. For example, McMullen and Rouse (2007) found that for Wake County, North Carolina, a switch to a year-round calendar had essentially no impact on academic achievement. And for Colombia, Zuluaga (2013) estimated that 10^{th} and 11^{th} grade students exposed to an increase in school breaks had on average 0.22 standard deviations lower test scores.

ii. Impact of the length of the school day or more instructional time per day

Most of the literature analyzing the length of the school day has focused on developed countries, and particularly in the United States. A large proportion of these studies have looked at the impact of extending the length of kindergarten. In a meta-analysis of the literature, Cooper, Batts Allen, Patall, and Dent (2010) found that attending a full-day kindergarten was positively associated with academic achievement, compared to a half-day kindergarten, but the effect generally seemed to fade-out by third grade. They also found a significant correlation between attending a full day and the child's self-confidence and ability to work and play with others. However, they found that children attending a full-day kindergarten did not have as positive an attitude toward school, and had more behavioral problems. Rathburn (2010) explored the relationships between full-day kindergarten program factors and public school children's gains in reading test scores. She found evidence that children who attended kindergarten programs that devoted a larger portion of the school day to academic instruction (particularly reading instruction) made greater gains in reading over the school year than children who spend less time in such instruction. However, the effects of a full-day kindergarten seem to fade out beyond the kindergarten year, something that might be related to summer learning loss (Redd et al., 2012).

Although there is a large literature on the impact of school time in early education programs, little is known about the impact on K-12 education. This is partly due to the small variation in the length of the school day across schools districts (in the US), or municipalities (in other countries), or to variations in length that are directly related to

resources, which makes it difficult to estimate the impact of longer school days or longer school years. Some of the studies that have used identification strategies that allow them to establish a causal impact of school time on student outcomes have shown that spending more time in school could improve academic achievement and could benefits students, families and societies by reducing teen pregnancies (Kruger & Berthelon, 2009) and crime rates (Jacob & Lefgren, 2003).

There are a few papers that have analyzed the impact on test scores of variations in the length of instructional time across countries. Using data from the 2006 Programme for International Student Assessment (PISA)[15] for over 50 countries, Lavy (2010) found evidence that instructional time consistently had a positive and significant effect on test scores both in developed and developing countries. However, the estimated effect for developing countries was much lower than the effect size in developed countries. More recently, Rivkin and Schiman (2013) looked at the effect of instructional time on the 2009 PISA test scores results for 72 countries. They also found similar that achievement increases with instructional time, and that the increase varies by the amount of time and classroom environment.

In Latin America, some studies have exploited natural experiments given by the implementation of a full-day schedule in schools that previously had half-day schedules. Valenzuela (2005) did one of the first impact evaluations of the Chilean full school day program. Implemented in 1997, the goal of this program was to increase by 30% the number of daily hours in public and voucher schools in the country. Taking advantage of

[15] The PISA is an international study jointly developed by the Organization for Economic Cooperation and Development (OECD) countries to evaluate education systems worldwide. Its main component comprehends testing the mathematics, science, and reading skills and knowledge of 15-year-old students in schools in participating countries or economies.

the differences in the time in which schools implemented the program, the author implemented a Difference-in-Difference model at the school and county levels. He found that the program had a significant, positive, and robust effect on schooling outcomes, but the impact was different depending on socioeconomic characteristics of the treated schools and also among subjects. At the county level, the results also showed a positive impact of the program.

Also analyzing the Chilean full-day schedule program, Bellei (2009) found that it had positive effects on students' academic achievement in both mathematics and language, and these effects were constant over time. He also suggested that the program had larger positive effects on rural students, students who attended public schools, and students in the upper part of the achievement distribution. Pires and Urzua (2011) analyzed the effect of time spent in school on schooling attainment, cognitive test scores, socio-emotional variables, labor market outcomes, and social behavior in Chile. They concluded that enrollment into a full-day school had positive effects on academic outcomes and cognitive test scores, and reduced adolescent motherhood, but there was no effect on employment or wages. Cerdan-Infantes and Vermeersch (2007) analyzed the impact of a program that lengthened the school day from a half day to a full day in poor urban schools in Uruguay. They found that students in very disadvantaged schools improved test scores by 0.07 of a standard deviation in math, and 0.04 of a standard deviation in language.

Llach et al. (2009) analyzed the long-term impact of attending longer school days in Buenos Aires, Argentina. They found that students that attended full-day primary schools had a secondary school graduation rate 21 percent higher than those that attended

half-day primary schools, and this was mostly driven by students coming from low socioeconomic backgrounds. However, they did not find enduring effects on income and employment.

For Colombia, Camacho and Mejia (2013) analyzed if the conditionality on school attendance imposed by a Conditional Cash Transfer program *(Familias en Acción),* lead to reductions in crime by impeding teenagers from getting involved in criminal activities. They found that the program does not seem to have an effect on crime through this "incapacitation" mechanism cause by school attendance.

To the best of my knowledge, there are only two studies that have evaluated the impact of longer schools days on student outcomes in Colombia. In the first one, Bonilla-Mejia (2011) evaluated the impact of attending a complete schedule school on student performance in a high school exit exam (SABER 11, previously known as "ICFES"). He used the number of students enrolled in complete schedule schools to instrument the probability that a student has of attending a complete schedule school. He found that students attending complete schedules have average test scores that are 2.5 percentage points larger than those attending half-day schedules. Additionally, they found that the impact was larger when compared with students attending only an afternoon shift. In the second study, Garcia et al. (2012) used family fixed effects to analyze the impact of changing from a half-day to a complete schedule on student outcomes. They found that a complete schedule reduces the probability of early dropout and grade repetition.

iii. Heterogeneous effects

The literature has found that the impact of more instructional time on student outcomes tends to be larger for students from poorer backgrounds or the most at-risk children. For example, Olsen and Zigler (1989) found that in general, extended-day programs in kindergarten seemed to increase standardized test scores in the short term, particularly for disadvantaged, bilingual, or "least-ready" for school children. Cooper et al. (2010) compared evaluations of full-day kindergarten programs in urban versus nonurban settings, and concluded that these programs had a significantly stronger association with higher academic achievement for children attending in an urban setting than those in non-urban communities. They suggested that children attending in urban settings were more likely to come from poorer backgrounds, so this could be taken as indirect evidence of a potentially greater impact of full-day kindergarten for poorer children. Harn, Linan-Thompson, and Roberts (2008) analyzed the impact of intensifying instructional time on early literacy skills for the most at-risk first graders. They concluded that the students receiving more intensive interventions (more instructional time) made significantly more progress across a range of early reading measures.

Additionally, there might also be heterogeneous effects on different types of students, because of variations across schools in the quality of instruction and of the content taught during the extra hours. For example, Llach et al. (2009) show that for Buenos Aires, the content of the additional hours was probably more important for academic achievement than increasing the number of hours. The impact on academic results could be very different depending on whether the extra hours are just an extension

of the current curriculum, or whether they allow the disadvantaged students to develop their skills and abilities through instruction in the areas in which their more advantaged schoolmates usually learn and practice.

iv. Hypotheses

Given this framework, the impact of longer school days on student achievement in Colombia is uncertain. There are three arguments for why longer schools days could benefit students: First, by spending more time in school they will devote more time to learning. Second, they will spend less time alone at home or outside home doing educationally unproductive activities. And third, they will spend less time in the streets, exposing themselves to different risks (Kruger & Berthelon, 2009; Jacob & Lefgren, 2003). Therefore, having a complete schedule might have a positive impact on student achievement if the quality of the additional hours is good (i.e. stimulating or educationally productive), because students might be able to learn more things and improve their skills, and this could be reflected in higher student achievement.

On the other hand, longer school days could have a negative or null impact on student achievement if instruction during those extra hours is of poor quality. For example, if less prepared teachers are brought to schools in order to be able to extend the number of school hours, then the additional time spent in school will not necessarily translate into more learning for the students attending complete schedules. Alternatively, it may be the case that higher-quality schools have greater demand and are therefore "split" into a morning and afternoon shift, as mentioned previously. In this case, schools with complete schedules might then have lower test scores than half-schedule schools.

Additionally, the length of the school day could have no impact if the content of the extra hours in the complete schedule schools is largely irrelevant. For example, if they devote the additional time to sports, or more playing time, then there might not be a significant difference in student achievement between schools or cohorts that have a complete schedule and those that have a half-day schedule. Finally, the additional hours could have no impact if those attending a half-day schedule systematically come from better-off families or they can make up for the fewer hours spent in school.

The impact of longer school days might be heterogeneous. First, it might be larger for the poorest or most vulnerable children. The reason for this is that, assuming the quality of the additional hours of schooling is the same across all students and schools, they will benefit by more because that extra time in school will be relatively better for them than the home environment, compared to better-off children. For example, if not in school for those additional hours, students from higher-income families might be attending extra-curricular courses, playing with their parents, doing homework, or reading books. In contrast, students from lower-income families might spend relatively more time outside of school and outside their homes, increasing their exposure to risk factors associated with crime, teen pregnancies or drugs. The impact of longer school days might also be larger for the older students (i.e. 9^{th} grade), because they are more likely to gain more from not being exposed to the risk factors associated with being outside the home, compared to the younger students, who are more likely to be at home if not in school for those extra hours. For example, older students who are not in school would most likely be in the streets, exposing themselves to crime and teen pregnancies,

while younger students would most likely be at home or somewhere else, but in general less exposed to those risks.

2.4 Empirical strategy

The main goal of this study is to analyze the impact of longer school days (i.e., complete schedule) on student achievement. As described in section 2.2, there are methodological difficulties in isolating this impact, because there might be systematic selection of students to schools with a particular type of schedule, or of schools with particular characteristics into a type of schedule.

The following sub-sections describe several estimation strategies and the extent to which they can address endogeneity and selection concerns. The dependent variable is the average SABER test score for each grade, subject, and year for each school. I standardize the schools' test scores by grade, subject, and year, so that each of these samples of school test scores has a mean of zero and a standard deviation of one. This allows me to interpret the coefficients in the regressions as standard deviations from the mean. The key independent variable in this study is an indicator for the type of schedule that a school has in each year: COMPLETE[16] is equal to one if the school has a complete schedule in a year, or zero if it does not. Not having a complete schedule is equivalent to saying that the school has a half schedule (morning OR afternoon schedule) or double shifts (morning AND afternoon schedules).[17]

[16] From now on, "COMPLETE" in capital letters, will refer to the variable, and "complete" in lower case will continue to be used as a noun that describes the type of schedule.
[17] In a relatively few cases, a school reports information for two different schedules, and describes one as complete and the other as a morning or afternoon schedule. This seems unlikely to happen, and it is more

i. Estimation Strategy 1: Cross-sectional OLS

I first run a traditional cross-sectional OLS model. This model compares schools that have a complete schedule with those that have a half-day schedule in year *t*, controlling for observable characteristics and for school performance in the previous period. The use of previous period test scores requires dropping data for 2002 (because this year does not have a previous period test score). Therefore, this model uses a sample which contains only data for 2005 and 2009.

The following specification is run separately for each grade and subject:

$$\mathrm{SABER_{st}} = \beta_0 + \beta_1 \mathrm{COMPLETE_{st}} + \beta_2 \mathrm{SABER_{s(t-1)}} + \beta_3 \mathrm{X_{st}} + \mathrm{d_m} + \varepsilon_s \quad (1)$$

Where $SABER_{st}$ is the average test score in SABER for school *s* in year *t*; $COMPLETE_s$ is a dummy variable equal to 1 if the school has a complete schedule and 0 if it does not (i.e. if it is has a morning or afternoon schedule); $SABER_{s(t-1)}$ is the average test score in the previous period; X_{st} is a vector of school control variables. d_m are municipality fixed effects to control for observed and time-invariant unobserved characteristics that are particular to each municipality.

The main school control variables included are: a dummy for location in a urban area; a dummy for public schools; total school enrollment; teacher-student ratio; the percent of) teachers with a professional education; and the percent of teachers that have specific pedagogic training;[18] and school socioeconomic status (SES), a variable that takes values from 1 to 4, where 1 corresponds to schools with the lowest socioeconomic

---

likely a mistake that someone makes when filling out the C600 forms. My main sample drops observations for these schools (1,140 observations, or 0.6% of the C600 sample).

[18] The teacher-student ratio, the percent of teachers with a professional education, and the percent of teachers that have specific pedagogic training, are computed for primary (for 5th grade students) and secondary (for 9th grade students).

status, and 4 to the ones with the highest socioeconomic status. The SES is an index that was calculated by the ICFES for each school, using a combination of variables with the proportion of students in the school with each of following characteristics: socioeconomic strata,[19] household resources (e.g., computer, television, and cell phone), household income, dwelling's floor material, parents' education, parents' occupation, number of people in the household, and household overcrowding.

The cross-sectional OLS strategy will yield unbiased estimates of the effect of a complete schedule if the Secretary of Education assigns school schedules and student spots randomly, or based only on observed characteristics. Because the Secretary of Education is supposed to take into consideration a lot of variables when assigning a school schedule and when allocating student spots in each school, the cross-sectional OLS is likely to produce biased estimates of the impact of COMPLETE. Therefore, I also implement panel OLS and school fixed-effect models to control for additional unobservable characteristics, and mitigate some of the selection issues.

ii. Estimation Strategy 2: Panel with municipality fixed effects

In this model I exploit the panel nature of my data by including indicators for each year to control for any observed or unobserved policy changes or events that affected all schools equally in each year. I also include municipality fixed effects. Because I no longer use lagged tests scores, I am able to include observations from the three years. The model is specified as follows:

$$\mathrm{SABER}_{st} = \beta_0 + \beta_1 \mathrm{COMPLETE}_{st} + \beta_2 \mathrm{X}_{st} + \mathrm{d}_t + \mathrm{d}_m + \varepsilon_{st} \qquad (2)$$

[19] Colombia classifies all dwellings in six socioeconomic strata according to their geographic location. "Strata 1" corresponds to areas with the lowest socioeconomic characteristics, and "Strata 6" to the areas with the highest socioeconomic characteristics.

Where $SABER_{st}$ is the average test scores in SABER for school *s* in year *t*; $COMPLETE_{st}$ is a dummy variable equal to 1 if the school has a complete schedule and 0 if it does not (i.e. if it is a half-day schedule); d_t are year fixed effects; d_m are municipality fixed effects; and all other notation represents the same variables as in (1).

iii. Estimation Strategy 3: Panel with school fixed effects

The problem with the cross-sectional OLS and the panel with municipality fixed effects strategies, is that it is possible that the observed differences between the schools with different schedules might simply reflect pre-existing differences between the schools or their students, or reflect the effect of unobserved characteristics that are correlated with having both a complete schedule and higher test scores. Consequently, this strategy may provide a biased estimate of the impact of attending a complete schedule.

In order to mitigate this potential problem, I use the panel dataset to implement a school fixed-effects strategy, by exploiting the fact that some schools switched from having a half-day schedule to a complete schedule, or vice versa. The school fixed effects focuses on analyzing within-school variation in mean school test scores across different cohorts (time), allowing me to break the correlation between unobserved school characteristics and having a complete schedule, and therefore eliminating the bias created by schools systematically selecting into a type of schedule (Bifulco, Fletcher, and Ross, 2011; Hoxby, 2000). To implement this strategy I estimate the following specification:

$$\mathrm{SABER}_{st} = \beta_0 + \beta_1 \mathrm{COMPLETE}_{st} + \beta_2 X_{st} + d_t + d_s + \varepsilon_{st} \qquad (3)$$

The dependent variable is, again, the SABER test score for the school in each period. It is a function of time-varying school characteristics (X), the school schedule in each period (COMPLETE), and time (dt) and school (ds) fixed effects.

The identifying assumption for the school fixed effect model is that schools that switched from having a complete schedule to a single shift, or vice versa, did so for reasons exogenous to the test scores. That is, the reason for the switch must be unrelated to test scores or to unobservable characteristics that are correlated with test scores. I argue that because switching into a different type of schedule is a municipality-level decision typically based on policy or on the demand and supply for school spots, it is exogenous from the point of view of the school, the parents and the students, and unlikely to be correlated with year to year (i.e., across cohort) variations in test scores. In section 2.6, I use a balancing test to check if the switch to a complete schedule is in fact uncorrelated with school characteristics.

Additionally, given that the Secretary of Education mandates that in the process of allocating spots each year, priority should be given to students that are already enrolled in the schools, parents of students in 5^{th} grade and 9^{th} grade are not likely to request a change to a school with a particular type of schedule because it was switching into a particular schedule, given that they would risk losing their school spot. In impact evaluation terminology, this means that 5^{th} graders and 9^{th} graders did not seek to be treated, or alter their condition (enrollment in their current school) in order to get the treatment (enrollment in a school with a complete schedule).

This is my preferred specification and, in my view, the best estimation of the causal impact of having a complete schedule. By controlling for all observed and unobserved time-invariant school characteristics, this specification decreases the possible bias caused by characteristics that do not change over time within the school, although some endogeneity may remain from time-varying characteristics. For example, there might be some selection issues if the Secretary of Education chooses school schedules based on school attributes, and particularly if these school attributes are related to school performance in the SABER. For example, if very motivated principals or Parents' Associations systematically push and convince the Secretary of Education to implement a complete schedule in their schools, the estimate of the impact of having a complete schedule might be biased, because this type of schools (with more motivated principals or Parents' Associations) are also more likely to have a complete schedule and to perform better in SABER. There might also be selection issues if the Secretary of Education assigns students to schools with a certain type of schedule, based on parents' or students' socioeconomic characteristics, for example, if all relatively poor students are assigned to half-day schools.

There might also be endogeneity from time-varying characteristics if students systematically switch schools after the change in schedule. For example, if more educated parents systematically apply for schools that have switched to a complete schedule, or if they move to an area where most schools have a complete schedule. However, this might not be the case, because in Colombia parents do not have much choice of schools or schedules, and in contrast to the US, they do not tend to move because of the schools in the area. Although this is an empirical question that I am not able to test directly from my

data, other authors have cited evidence from the 2007 Quality of Life Survey in Bogota, suggesting that of the 21% of households that move within a 2 year period, just 4% of those cite education or health considerations as the reason for the move (Bonilla-Angel, 2011).

2.5 Data

To implement the strategies described above, this study exploits an extensive and relatively new panel dataset that contains test score results from SABER 5th and 9th for three periods: 2002, 2005, and 2009.[20] SABER 5th and 9th are nationwide standardized assessments administered every three years to all students in 5th grade and 9th grade in Colombia.[21] According to the Ministry of Education, the main goal of these tests is "to contribute to the improvement of the quality of Colombian education, by carrying out periodic measurements of the development of competencies in students enrolled in basic education, as an indicator of the quality of education."[22] SABER tests are the best available assessments of primary and secondary student achievement in the country because of their scientific rigor and nationwide coverage. The micro data from these tests has only been made available to researchers in the last few years, making this one of the best and newest available sources to study questions related to student achievement in primary and secondary education in Colombia.

[20] More precisely, the SABER 2002 was administered in different moments in 2002-2003, but most tests for 5th grade and 9th grade were conducted in 2002, so I refer to this test as SABER 2002. In the same way, the SABER 2005 was administered in different moments in 2005-2006, but I refer to this test as SABER 2005. All SABER 2009 tests were administered in 2009.

[21] The 2012 application of SABER was extended to evaluate third grade students, but data for this round is not included in this dataset.

[22] Author's translation from the Ministry of Education's website. Retrieved from: http://www.mineducacion.gov.co/1621/w3-article-244735.html, on June 1st, 2013.

The SABER tests are administered by the Instituto Colombiano para la Evaluación de la Educación (ICFES, the Colombian agency for the evaluation of education) every 3 years. Along with the test, some socioeconomic information of the students and schools is also collected. The SABER 5th and 9th are designed to evaluate competencies in language (Spanish), math, and science[23] at the end of each basic education cycle. Therefore, they are administered to students in 5th grade, which is the end of the "basic primary cycle," and 9th grade, which is the end of the "basic secondary cycle." SABER evaluates student achievement in all schools in the country, both private and public, but private schools are excluded from this analysis because they are not affected my most public policy decisions, including those regarding the length of the school day.

The SABER panel dataset used in this study comes from the ICFES' databases. The SABER 2002 used different evaluation instruments, and the methodology for implementing the tests changed in the three rounds. Therefore, initially it was not possible to find information about performance in SABER that was comparable through time. However, since 2008, the ICFES, together with the World Bank, has worked to reclassify the tests administered in 2002 and 2005, to make them fully comparable to those administered in 2009. As a result of this process, they constructed a dataset with information from student achievement in the 2002, 2005, and 2009 rounds of SABER (they refer to this dataset as "historic SABER"). This study draws on this dataset because it contains comparable test scores across time, allowing me to compare variations in academic achievement over this period.

[23] The science assessment was introduced in 2012, so it is also excluded from this dataset and the analysis.

The SABER panel contains test scores at the individual level, and some basic identification information, but it does not contain information on school characteristics. Therefore, in order to implement the strategies described above, I merge the SABER panel dataset with school data from the C600, a survey used annually by the Departamento Administrativo Nacional de Estadística (DANE, the national statistics agency) to collect information on basic school characteristics. This dataset contains information on the schools' type (public or private), location (urban or rural), and school characteristics such as enrollment, information on number of transfers, dropouts, grade repetition and promotion in the previous academic year, number of groups or classes by grade, administrative staff, and some information about teachers (e.g., grade they teach, and their educational level). There are two types of C600 forms: C600 A collects the information for each school, and C600 B collects information for each school schedule. That is, if a school has more than one schedule, it fills out one C600 A form, and one C600 B form for each of its schedules or shifts.

The C600 data was initially fragmented in different datasets for each year and for each section or chapter of the C600 form, so I first merge all these parts for each year, and then I create a panel with information from the C600 from 2002, 2005, and 2009.[24] Even though the C600 is collected at the school-schedule level, the data from SABER does not allow identifying the type of schedule that each student attends. Therefore, in order to match these two datasets I have to aggregate the C600 data at the school level. Because schools that do not have a complete, morning, or afternoon schedule are very uncommon (mostly night and weekend schedules, and about 5.5% of all schedule-level

[24] For administrative and technical reasons, the C600 was not collected in 2003 and 2004.

observations), and their student characteristics might be so different than those of students attending traditional "week-day day schools," before aggregating at the school level I drop night and weekend schedule observations from the dataset, and exclude them from the analysis.

The SABER dataset contains information for each student that took the test, but given that the data is not representative at the student level, and that I only have additional data on school level characteristics, I first aggregate the SABER data at the school level. With both the SABER and C600 panels at the school level, I use the school identification codes to merge them and create a school-level panel. SABER test scores cannot be compared across grades, so I create separate samples for 5^{th} and 9^{th} grade tests, including only schools with non-missing math and language scores in the given grade. . Each of these samples contains the school's average SABER math and language test scores for that grade and its corresponding school characteristics from the C600.[25]

To be part of the panel, a school has to have information from both SABER and C600 for a given year, but it does not necessarily have to have information for the three periods. That is, it is an unbalanced panel.[26] For every year, there are a significant number of schools in the C600 that cannot be matched with any school in the SABER dataset. In some cases this is due to problems matching the school identification codes in the two datasets (because of mistakes when typing-in the data, or because of changes in the codes). Additionally, when the ICFES and the World Bank constructed the historic SABER with data from schools in each of the periods, a large number of school

[25] Separating schools into four sub-samples by grade and subject (and therefore including schools that are missing one subject test per grade) does not change the results (available on request).
[26] As part of the robustness checks, I also run my models using a balanced panel with schools that appear in the three periods in the database.

observations were not matched across years, and were not included in the dataset. Therefore, they could not be matched to C600 either.

There are three main reasons why schools could not be matched across time, and are therefore not included in the "historic" SABER dataset that I use. First, although the reclassification exercise made most of the SABER tests comparable across time, a substantial number of the SABER 2002 tests could not be made comparable and were not included in the historic SABER dataset. Second, some school identification codes could not be matched across time because some small schools were merged into larger "education institutions" with a single administration, changing their identification codes. Finally, there was evidence of cheating in some schools, especially in 2002 and 2005, which made impossible to make these schools' test scores comparable across time (ICFES, 2011).

Table 2.1 contains descriptive statistics of the school panel for the 5th grade and 9th grade samples. In the fifth grade sample, 18.8% of the schools in the sample are located in an urban area. This percentage increases to 25.4% percent in 2005, and goes down to 21.1% in 2009. The percentage of urban schools is much larger for the 9th grade sample, were between 51.1% (2009) and 60% (2005) of schools in the sample are urban. In the 5th grade sample, the average socioeconomic status (between 1.38 in 2002 and 1.41 in 2005) and the high percentage of schools with the lowest socioeconomic status (almost three-quarters in each year) indicates a high concentration of the poorest schools in the school panel. In the 9th grade sample, although there is also a large percentage of schools with the lowest socioeconomic status, the average SES is higher (around 1.8), and the concentration of poorest schools is lower than in the 5th grade sample (less than half the

schools in 2002 and 2005, and 52.4% in 2009). In the 5th grade sample, around 27 percent of schools have a complete schedule in 2002 and 2009, and 34.2 percent of schools have a complete schedule in 2005. In the 9th grade sample there is a similar trend: In 2002, 21.5 percent of schools have a complete schedule, and this number goes up to 23.9 percent in 2005, and goes down again in 2009 to 19.9 percent.

Table 2.2 presents the unconditional mean and standard deviation of test scores, and test scores at different percentiles of the distribution, for schools in the 5th grade and 9th grade samples.[27] In general, the unconditional mean of all test scores goes down in time. The unconditional mean in both grades and both subjects got worse between 2002 and 2005, but the largest decrease was for the 5th grade math test scores, where there was a tenfold decrease in the mean test score (from -0.01, to -0.10). The smallest decrease was in 9th grade math test scores, with test scores going from -0.09 to -0.12 of a standard deviation. The average 9th grade test scores got even worse in 2009: the language test scores for this grade went from -0.22 in 2005, to -0.28 in 2009, and the math test scores went from -0.12 in 2005, to -0.26 in 2009. In contrast, the average language test scores improved in the 5th grade language sample: it went from -0.16 in 2005, to -0.12 in 2009. The standard deviation narrowed down in time in the four samples.

Since my identification strategy relies on schools that switched schedules, Table 2.3 shows the number of schools that switched by type of change (from half to complete schedule, and from complete to half schedule), and classified by the year in which the

[27] Test scores were standardized by grade, subject, and year, in the complete SABER dataset (i.e., including private schools and schools with schedules different from complete, morning or afternoon, all of which were excluded from the final panel; and also before losing some observations when merging the SABER dataset with the C600 school data). For this reason, the unconditional means in this table are different from zero, and the standard deviations are different from one.

switch takes place. In total, 2,262 schools in the 5th grade sample and 464 schools in the 9th grade sample switched schedule at some point between 2002 and 2009, which corresponds to 12 percent and 10 percent respectively, of the schools that exist in at least 2 periods in the panel.[28] The number of schools that switched scheduled in each sample, corresponds to the schools that are used in the school fixed effect model to identify the effect of COMPLETE.

2.6 Results

i. Main Results

Table 2.4 presents the results of the cross-sectional OLS model (1) for the 4 samples. This model exploits the variation of schedule and other variables across schools to estimate how much of the variation between school test scores can be explained by having a complete schedule, controlling for other observable school characteristics, previous performance on the SABER exam, and municipality. For both of the 5th grade samples, the coefficients on COMPLETE are very small and not statistically significant. In contrast, the results for the 9th grade sample show that schools with a complete schedule have language test scores that are 0.126 of a standard deviation higher (column 3), and math test scores that are 0.114 of a standard deviation higher (column 4), than schools without a complete schedule. Most of the school characteristics are also

[28] There are 237 schools in the 5th grade sample, and 74 schools in the 9th grade sample, that switched from half to complete in 2005 and then back to half in 2009, or that switched from complete to half in 2005, and then back to complete in 2009. These schools are excluded from the calculations in this table, but they are included in the sample used in the regressions

statistically significantly associated with higher test scores, particularly the previous year test scores (associated with about one tenth of a standard deviation higher test scores), the dummy for urban schools, and the school socioeconomic status. The primary teacher student ratio and the percentage of teachers with a pedagogical training are also statistically significant, but only for the 5th grade samples. However, the coefficients from these across-school regressions are likely biased because there might be unobserved characteristics that are correlated with test scores and the likelihood of having a complete schedule.

Table 2.5 presents the main findings for 5th grade test scores by subject, using the panel specifications (2) and (3) described in section 2.4.[29] The municipality fixed effects specification is reported in columns 1 and 3, and the school fixed effects specification is reported in columns 2 and 4. As mentioned above, the latter is my preferred specification, because most of the characteristics and confounding factors that could bias the estimates of COMPLETE originate from differences between schools with and without a complete schedule. Therefore, by comparing variations in test scores within a school, across cohorts (before and after a scheduling switch), many of the selection and endogeneity concerns are mitigated.[30]

The first row of Table 2.5 presents the estimates for COMPLETE, the impact of having a complete schedule on test scores. COMPLETE is statistically significant at the 1

[29] I ran regressions of a base model that only includes COMPLETE, and year and municipality fixed effects as explanatory variables. The results for this model tell us the test scores differences between school with a complete schedule and schools with a half schedule: Schools with a complete schedule have language test scores that are on average 0.285 of a standard deviation higher, and math test scores that are 0.274 of a standard deviation higher, than schools with a half schedule.

[30] Because the school fixed effects only captures the changes in test scores that can be explained by time-varying variables, time-invariant variables are excluded from this regression.

percent level for the panel model for 5th grade language, and across the two specifications for 5th grade math. Looking at the language results first, the estimate of COMPLETE in the municipality fixed effects specification shows that having a complete schedule increases language test scores by 0.055 of a standard deviation, compared to schools without a complete schedule in the same municipality and year (column 1). In the school fixed effects specification (column 2), the estimate of COMPLETE (0.04) is only slightly lower than in the panel specification, but it is no longer statistically significant.

Columns 3 and 4 in Table 2.5 show the set of findings for 5th grade math. The coefficient for COMPLETE in both the municipality and the school fixed effects specifications is 0.082 and statistically significant, and in both cases, it is larger than the respective specification for language (e.g., 0.082 v. 0.055, using the municipality fixed effects specification). In the school fixed effects model (column 4), the estimate for COMPLETE can be interpreted as the within-school impact of switching to a complete schedule on math test scores. In this case, among schools that switched schedules, the cohorts exposed to a complete schedule have test scores that are 0.082 of a standard deviation higher than cohorts that attended half schedules. The estimated coefficients are very similar across specifications for 5th grade math (0.082), and they do not change much across specifications for the 5th grade language (0.055 in the municipality fixed effects, and 0.044 in the school fixed effects).The fact that the school fixed effects do not substantially alter the estimates, suggest that most of the selection occurs at the municipality level. The coefficients in all of the specifications are larger than those in the cross-sectional OLS model (Table 2.4, columns 1 and 2). This is consistent with a possible downward bias in the OLS estimates caused by a systematic selection of lower-

quality, and therefore lower-demand, schools into a complete schedule, as noted in sections 2.2 and 2.3.

Table 2.6 presents the findings for 9th grade test scores by subject. The first row shows that there is a statistically significant (at the 1% level) difference in test scores between schools that have a complete schedule and schools that have a half schedule. The municipality fixed effects results reveal test score differences of 0.162 for language (column 1) and 0.137 for math (column 3). The school fixed effects results show that among schools that switched schedules, there is also a statistically significant difference in test scores between cohorts exposed to a complete schedule and cohorts that attended half schedules The estimated effect is 0.110 for language (columns 2) and 0.138 for math (column 4), suggesting stronger effects in math, as in the 5th grade sample.

Most notably, all the coefficients for COMPLETE in the 9th grade sample are larger than the respective coefficients for 5th grade, suggesting that having a complete schedule has a larger positive impact on 9th graders than on 5th graders. The 9th grade estimates are more than double the 5th grade estimates in language, and about 50 percent larger in math. The estimates of COMPLETE are also larger than those in the cross-sectional OLS model (Table 2.3, columns 3 and 4), except in the school fixed specification for language test scores. Again, this might be reflecting a downward bias in the OLS estimates caused by systematic selection of lower-quality schools into complete schedules.

In sum, the most reliable estimates of the impact of a complete schedule, which come from the school fixed effects specifications, show that among schools that switched

schedules, the cohorts exposed to complete schedules have higher test scores than cohorts that attended half schedules, and the impact ranges from 0.082 of a standard deviation for 5th grade math test scores, to 0.138 of a standard deviation for 9th grade math test scores.

ii. Balancing Tests

Throughout this paper I argue that the school fixed effect approach provides the most reliable estimates of the impact of attending a school with a complete a schedule. As mentioned in section 2.4, the identifying assumption for this approach is that switching to a complete schedule must be uncorrelated with the characteristics of the school. To test this assumption, I estimate separate regressions of school characteristics as a function of COMPLETE, cohort (time) fixed effects, and school fixed effects. Table 2.7 shows the estimated coefficient for COMPLETE for each of these balancing tests. These coefficients tell us if there is a statistically significant difference in each of the school characteristics in column 1, between cohorts exposed to a complete schedule and cohorts exposed to half schedules, in the same school.

As expected, there is a statistically significant difference in school enrollment between schools exposed to complete and half schedules, given that schools with double shifts are likely to have a larger enrollment. There is also a statistically significant difference in the percentage of teachers with a professional degree in the 9th grade sample. The negative coefficient suggests that schools that switch to complete schedules have a lower proportion of professional teachers. This negative selection would likely bias estimates down. Further, the absence of statistically significant estimates for the other schools characteristics suggests that the observable characteristics of the school are

not highly correlated with the scheduling switch within schools across cohorts. Although there might still be some correlation between switching to a complete schedule and unobserved characteristics of the school, these results provide supporting evidence for the internal validity of the school fixed effects identification strategy.

iii. Heterogeneous effects and alternate samples

In Table 2.8, I test for heterogeneous effects by estimating the school fixed effects specification using different subsamples: only rural schools, only urban schools, and only schools classified in each of the four socioeconomic statuses (SES). Columns 1 to 4 present the results for each grade and subject, while each row represents the coefficient of COMPLETE for the specific subsample. Although the smaller samples reduce the power to detect effects, I find a statistically significant impact of COMPLETE for the rural schools in the 5^{th} grade math sample: cohorts exposed to complete schedules in rural schools have test scores that are almost one tenth of a standard deviation higher than cohorts that attended half schedules in rural schools. There is also a marginally statistically significant (10%) impact of being exposed to a complete schedule in schools in urban areas for the 9^{th} grade language sample.

Splitting the sample by socioeconomic status in the lower rows of Table 2.8, reveals that the impact of having a complete schedule is statistically significant only for the subsample of schools with the lowest socioeconomic status (SES=1), and it is not statistically significant for any of the subsamples containing schools with higher socioeconomic statuses (SES=2, 3, or 4). Among the poorest schools, cohorts exposed to a complete schedule have test scores that are 0.067 of a standard deviation higher in 5^{th}

grade language, and 0.093 of a standard deviation higher in 5th grade math, than the cohorts that attended half schedules. There is an even larger impact of attending a complete schedule for the poorest schools in the 9th grade math sample: cohorts exposed to a complete schedule have test scores that are one fifth of a standard deviation higher that cohorts exposed to half schedules. This provides more evidence supporting what has been found in previous studies: having longer school days has heterogeneous impacts on student achievement, and the impact is generally larger for the poorest schools.

Table 2.9 shows the results of estimating the school fixed effects specification using another set of subsamples. In the first two rows, I run the model using a subsample with only schools that switched from half to complete (first row), and a subsample with only schools that switched from complete to half (second row), in order to analyze if there are symmetric effects of switching from half to complete, and from complete to half. COMPLETE is only statistically significant for the subsample of schools that went from complete to half: for these schools, the cohorts exposed to a complete schedule have 0.214 of a standard deviation higher test scores in 5th grade math, and 0.306 of a standard deviation higher test scores in 9th grade math. The regressions reported in the third row use only the schools that exist throughout the whole period (i.e., using a balanced school panel). In this case, COMPLETE is only statistically significant for the 9th grade language sample (at the 10% level), and the impact is smaller than for the complete sample of schools (0.089 v. 0.11 (Table 2.6, column 2)).

Finally, the bottom two rows of Table 2.9 examine "treatment" intensity. I ask whether the impact of attending a complete schedule is different for schools that were exposed to a complete schedule for longer/shorter lengths of time. The fourth row

presents the results for regressions that used a subsample of only schools that switched from a half to complete schedule between 2002 and 2005, and kept the complete schedule in 2009. These schools could be considered "early adopters" and were exposed for longer to the "treatment" (i.e., complete schedule). In this case, at least two cohorts within the school were exposed to a complete schedule. It could also mean that students in these schools were exposed to a complete schedule for more than one grade, although it is not possible to know for sure if students changed schools over this period. Despite the very small sample size (444 schools), the results show that cohorts exposed to a complete schedule in schools that were treated for longer have test scores that are 0.274 of a standard deviation higher in 9^{th} grade language, and 0.292 of a standard deviation higher in 9^{th} grade math, than cohorts that attended half schedule.

The bottom row reports the regressions results using a subsample of only schools that switched from half to complete after 2005. That is, it contains schools that had a shorter duration of a complete schedule and a longer duration of a half schedule in the time period studied. Essentially, students in these schools were "late adopters," and were exposed to the treatment for a shorter period. . There is no statistically significant impact of being exposed to a complete schedule for the schools in this subsample.

2.7 Conclusions and policy implications

This study provides some of the first evidence that longer school days have an impact on 5^{th} grade and 9^{th} grade academic achievement in Colombia. By using a school fixed effect model that exploits within-school changes in the type of schedule, I am able to control for observed and unobserved time-invariant characteristics of schools,

mitigating some of the most critical selection and endogeneity problems that commonly occur when comparing different types of schools. I find that there is a positive impact of having a complete schedule (approximately 2-3 additional hours) on school achievement in 5th grade math, and 9th grade math and language SABER test scores. Results from the school fixed effects model show that among schools that switch schedules between 2002 and 2009, the cohorts exposed to complete schedules have test scores that are about one tenth of a standard deviation higher than cohorts that attended half schedules. This corresponds to approximately a 2.6 percent increase in test scores with respect to the mean for each grade, subject and year.

To put the magnitude of this effect in perspective, the impact of longer school days is smaller than other popular education interventions in Colombia, namely the PACES voucher program, and the contractual schools in Bogota. The PACES voucher program, which has received a lot of attention in the literature because of the use of a lottery to allocate vouchers, had an impact of about 0.2 standard deviations on student test scores relative to students who did not win the lottery (Angrist et al., 2002). While contractual schools in Bogota, which were traditional public schools whose administration was contracted out to reputed, not-for-profit private schools and universities, had an impact of 0.6 and 0.2 standard deviations in math and verbal tests, respectively, relative to traditional public schools (Bonilla-Angel, 2011). However, the magnitude of the impact estimated in this study is similar to other school intervention such as reducing class size by 4 students (Krueger, 1999), or the impact of charter schools on reading test scores (Angrist et al., 2010), and considering that two-tenths of a standard deviation is roughly the score gain associated with one additional school year

(Cole et al., 1993), it is a sizable effect. Overall, the results suggest that lengthening the school day may be an effective policy for increasing student achievement.

Further, I find that the impact of having a complete schedule is larger for math test scores than for language test scores in both 5th and 9th grade (e.g., 0.138 v. 0.11 respectively, for 9th graders). This result is not surprising, since, as other research has found, schools may play a larger role in teaching math relative to language, since language skills are often shaped by the home environment.

More notable, is that the positive effects of a longer school day are stronger for 9th grade students than for 5th grade students (e.g., 0.138 v. 0.082 respectively, for math test scores). This makes sense since adolescents may be more likely to engage in risky behaviors outside of school than younger children. Even if they are not engaging in academic endeavors during the extra school hours, 9th graders are certainly spending less time exposed to risk factors outside of school, which ultimately translates into better academic achievement.

Finally, my results suggest that the effects of complete schedules are heterogeneous. Like many other interventions in developing countries, the impact of complete schedules on test scores is largest among the poorest schools and those in rural areas.

Overall, my findings complement those of previous studies in Colombia, which have found a positive impact of attending a complete schedule on dropout and grade repetition in primary (Garcia et al., 2012), and a positive impact of attending a complete schedule on graduation-exit tests (Bonilla-Mejia, 2011). Together, all these findings

suggest that longer schools days have a positive impact on academic achievement and other student outcomes in Colombia. Therefore, lengthening the school day may be an effective policy for increasing student achievement, particularly for the lowest-income students in Colombia and other developing countries.

One argument against increasing the length of the school day is that the personnel costs will almost double because of the need to hire new teachers for the additional shift or the additional hours. However, in the Colombian context this might not be the case, because right now different groups of teachers serve different school shifts, so a change to a complete schedule would imply a salary increase for current teachers, but will not require hiring new teachers (Garcia et al., 2012). Similarly, it might be argued that implementing a complete schedule in all public schools would require high investments in infrastructure, and possibly constructing many more schools. However, right now there a large number of schools that have only a morning or an afternoon shift (Bonilla-Mejia, 2011). This means that there is an opportunity for increasing the length of the school day at least in those schools, without high investments in new infrastructure or constructing new schools.

A caveat with the main empirical strategy used in this paper, the school fixed effects model, is that it is not able to control for time-variant characteristics that could be correlated with both the change in schedule and test scores. Although I argue that these are not likely to be large, future research should consider expanding the empirical work to include other quasi-experimental estimation strategies or, if possible, a randomized controlled trial to more definitively assess the impact of longer school days on student achievement.

2.8 References

Angrist, J., Dynarski, S., Kane, T., Pathak, P., & Walters, C. (2010). Who Benefits from KIPP? NBER, Working Paper 15740.

Angrist, J., Bettinger, E., Bloom, E., King, E., & Kremer, M. (2002). Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment. *American Economic Review*, *92*(5), 1535-58.

Barrera-Osorio, F., Maldonado, D., & Rodríguez, C. (2012). Calidad de la Educación Básica y Media en Colombia: Diagnóstico y Propuestas. *Documento CEDE No. 41*.

Berliner, D. (1990). What's all the fuss about instructional time? In Ben-Peretz, M. & Bromme, R. (Eds.), *The nature of time in schools* (3-35). New York: Teachers College Press.

Bellei, C. (2009). Does Lengthening the School Day Increase Students' Academic Achievement? Results from a Natural Experiment in Chile. *Economics of Education Review, 28*(5*),* pp. 29-40.

Bifulco, R., Fletcher, J., & Ross, S. (2011). The effect of classmate characteristics on post-secondary outcomes: Evidence from the Add Health. *American Economic Journal: economic policy* 3, 25-63.

Bonilla-Angel, J.D. (2011). Contracting out public schools and academic performance. Evidence from Colombia (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses.

Bonilla-Mejia, L. (2011). Doble jornada escolar y calidad de la educación in Colombia. Documentos de trabajo sobre Economía Regional, No. 143.

Camacho, A., & Mejía, D. (2013). The Externalities of Conditional Cash Transfers on Crime: The case of Familias en Acción in Bogotá. Universidad de Los Andes, Documento CEDE 2013- 15.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64,* 723-733.

Cerdan-Infantes, P., & Vermeersch, C. (2007). More time is better: An evaluation of the full-time school program in Uruguay (Working Paper No. 4167). Retrieved from The World Bank website:

http://elibrary.worldbank.org/doi/book/10.1596/1813-9450-4167

Cole, N., Trent, E., & Wadell, D. C. (1993). La prueba de realización en Español: Teacher's guide and technical summary. Chicago: Riverside Publishing Company, 1993.

Cooper, H., Batts Allen, A., Patall, E., & Dent, A. (2010). Effects of full-day kindergarten on academic achievement and social development. *Review of Educational Research, 80* (1), 34-70.

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research, 66* (3), 227-268.

Patall, E., Cooper, H., & Batts Allen, A. (2010). Extending the school day or school year: As systematic review of research (1985-2009). *Review of Educational Research, 80* (3), 401-436.

Eren, O., & Millimet, D. (2007). Time to learn? The organizational structure of schools and student achievement. *Empirical Economics, 32*, 301–332.

Gaviria, A., & Barrientos, J. (2001) Determinantes de la calidad de la educación en Colombia. Archivos de Economía, No. 159.

Garcia, S., Fernandez, C., & Weiss, C. (2012). Does lengthening the school day reduce the likelihood of early school dropout and grade repetition: Evidence from Colombia. Paper presented at the Population Association of America 2012 Annual Meeting, San Francisco, CA, May 3-5, 2012.

Hanushek, E., & Wossermann, L. (2007). Education quality and economic growth. Washington, DC: The World Bank.

Hanushek, E. (1979). Conceptual and empirical issues in the estimation of educational production functions. *The Journal of Human Resources*, *14*(3), 351-388.

Harn, B., Linan-Thompson, S., & Roberts, G. (2008). Intensifying instruction: Does additional time make a difference for the most at-risk first graders? *Journal of Learning disabilities*, *41*(2), 115-125.

Hoxby, C. (2000b). Peer effects in the classroom: Learning from gender and race variation (Working Paper No. 7867). Retrieved from National Bureau of Economic Research website: http://www.nber.org/papers/w7867

ICFES (2011). SABER 5º y 9º 2009. Informe técnico de resultados históricos. Bogotá, D.C.

Jacob, B., & Lefgren, L. (2003). Are idle hands the devil's workshop? Incapacitation, concentration, and juvenile crime. *American Economic Review,* 93, pp. 1560-1577.

Krueger, A. (1999). Experimental estimates of education production function. *The Quarterly Journal of Economics, 114*(2), pp. 497-532.

Kruger, D. I., & Berthelon, M. (2009). Delaying the bell: The effects of longer school days on adolescent motherhood in Chile. *Institute for the Study of Labor (IZA) Discussion Paper, 4553*.

Lavy, V. (2010). Do differences on school's instruction time explain international achievement gaps in Math, Science and Reading? Evidence from developed and developing countries (Working paper series No. 16227). Retrieved from National Bureau of Economic Research website: http://www.nber.org/papers/w16227

Lee, J. & Barro, R. J. (2001), Schooling Quality in a Cross–Section of Countries. *Economica, 68*, 465–488

Llach, J., Adrogue, C., & Gigaglia, M. (2009). Do Longer School Days Have Enduring Educational, Occupational, or Income Effects? A Natural Experiment in Buenos Aires, Argentina. *Economía, 10* (1), 1-43.

Marcotte, D. (2007). Schooling and Test Scores: A Mother-Natural Experiment. *Economics of Education Review, 26*(3), 629-40.

Marcotte, D., & Hansen, B. (2010). Time for School. *Education Next, 10* (1), 52-59.

Margo (1994). Explaining Black-White Wage Convergence, 1940-1950. *Industrial and Labor Relations Review, 48* (3) (Apr.1995), 470-481

McMullen, S., & Rouse, K. (2012). The impact of year-round schooling on academic achievement: Evidence from mandatory school calendar conversions. *American Economic Journal: Economic Policy, 4* (4), 230-252.

Ministerio de Educación Nacional (2013). Pruebas SABER. Retrieved from: http://www.mineducacion.gov.co/1621/w3-article-244735.html, on June 1st, 2013.

Ministerio de Educación Nacional (2006). Cartilla 1: Proceso de matrícula. Retrieved from: http://www.mineducacion.gov.co/cvn/1665/article-99324.html on November 14th, 2013.

Olsen, D., & Zigler, E. (1989). An assessment of the all-day kindergarten movement. *Early Childhood Research Quarterly, 4*, 167–186.

Pires, T., & Urzua, S. (2011). Longer schools days, better outcomes? (Working Paper). Northwestern University.

Rathburn, A. (2010). Making the most of extra time: Relationship between full-day kindergarten instructional environments and reading achievement. American Institute for Research, Research Brief.

Redd, Z., Boccanfuso, C., Walker, K., Princiotta, D., Knewstub, D., & Moore, K. (2012). Expanding time for learning both inside and outside of the classroom: A review of the evidence base (Child Trends #2012-20).

Rivkin, S., & Schiman, J. (2013). Instruction time, classroom quality, and academic achievement. NBER Working Paper No. 19464.

Rivkin, S., Hanushek, E., & Kain, J. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 4147-58.

Valenzuela, J.P. (2005). Essays in economics of education. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses.

World Bank (2008). Quality of education in Colombia. An analysis and options for a policy agenda. Report No. 43906-CO

Table 2.1. Descriptive statistics of school panel

	5th grade			9th grade		
	2002	**2005**	**2009**	**2002**	**2005**	**2009**
Total schools	14419	15006	24767	2970	4214	6428
% urban	18.8%	25.4%	21.1%	56.4%	60.0%	51.1%
Average socioeconomic status (SES)	1.38	1.41	1.39	1.80	1.89	1.77
% of poorest schools (SES=1)	73.5%	72.8%	72.9%	49.7%	46.0%	52.4%
% of richest schools (SES=4)	2.2%	2.8%	2.4%	7.4%	8.2%	7.0%
Average school enrollment	195	258	205	609	766	673
Average 5th/9th grade enrollment	22	30	24	62	82	73
Average primary/secondary enrollment	129	153	118	321	387	342
Primary/secondary teacher-student ratio	0.04	0.04	0.05	0.07	0.06	0.06
% primary/secondary teachers with professional degree	59.4%	68.3%	67.0%	86.5%	91.7%	90.0%
% primary/secondary teachers with pedagogic training	60.5%	67.8%	67.0%	83.8%	86.9%	83.5%
% schools with complete schedule	27.0%	34.2%	27.1%	21.5%	23.9%	19.9%
% schools that switched schedule	14.1%	10.4%	9.8%	15.3%	11.4%	8.2%

Note: School socioeconomic status (SES) is an index that goes from 1 (poorest schools) to 4 (richest schools).

Table 2.2. Distribution of test scores by year

	Language			Mathematics		
	2002	2005	2009	2002	2005	2009
Percentile			**5th grade**			
5%	-1.53	-1.63	-1.49	-1.48	-1.53	-1.45
25%	-0.76	-0.78	-0.74	-0.84	-0.79	-0.77
Median	-0.14	-0.21	-0.21	-0.10	-0.24	-0.23
75%	0.64	0.38	0.39	0.73	0.46	0.40
95%	1.75	1.61	1.64	1.77	1.80	1.71
Mean	**-0.04**	**-0.16**	**-0.12**	**-0.01**	**-0.10**	**-0.11**
St. deviation	**1.01**	**0.96**	**0.94**	**1.02**	**1.00**	**0.95**
Percentile			**9th grade**			
5%	-1.37	-1.63	-1.59	-1.12	-1.06	-1.43
25%	-0.73	-0.75	-0.78	-0.73	-0.61	-0.77
Median	-0.26	-0.24	-0.24	-0.38	-0.34	-0.29
75%	0.30	0.28	0.21	0.19	0.06	0.16
95%	1.73	1.28	0.95	2.16	1.78	1.08
Mean	**-0.13**	**-0.22**	**-0.28**	**-0.09**	**-0.12**	**-0.26**
St. deviation	**0.94**	**0.89**	**0.80**	**0.99**	**0.94**	**0.80**

Note: Test scores are standardized by grade, subject and year.

Table 2.3. Number of schools that switched schedule by type of change and year

	5th grade	9th grade
Total number of schools*	19546	4570
Number of schools that switched schedule	**2,262**	**464**
Percent of schools that switched schedule	12%	10%
From half to complete schedule	**1398**	**250**
half in 2002, complete in 2005 and 2009	409	148
half in 2002 and 2005, complete 2009	120	50
half in 2002, complete in 2005**	47	8
half in 2002, complete in 2009**	723	25
half in 2005, complete in 2009**	99	19
From complete to half schedule	**864**	**214**
complete in 2002, half in 2005 and 2009	173	75
complete in 2002 and 2005, half in 2009	89	38
complete in 2002, half in 2005**	14	3
complete in 2002, half in 2009**	220	33
complete in 2005, half in 2009**	368	65

Notes: Schools that switched from half to complete in 2005 and then back to half in 2009, or that switched from complete to half in 2005, and then back to complete in 2009 are excluded from the total number of schools that switched schedule (237 schools in the 5th grade sample, and 74 schools in the 9th grade sample).*Total number of schools that exists in 2 or 3 periods. **These schools only exist in 2 periods.

Table 2.4. Cross-sectional OLS regressions. Dependent variable: School test scores

	5th grade		9th grade	
	Language	Mathematics	Language	Mathematics
	(1)	(2)	(3)	(4)
Complete	0.006	0.040	0.126***	0.114**
	[0.033]	[0.036]	[0.041]	[0.047]
Test score in previous period	0.087***	0.108***	0.115***	0.084***
	[0.008]	[0.008]	[0.015]	[0.018]
Dummy for urban schools	-0.097***	-0.127***	-0.049	-0.031
	[0.028]	[0.027]	[0.031]	[0.032]
School socioeconomic status	0.137***	0.111***	0.297***	0.205***
	[0.020]	[0.018]	[0.018]	[0.020]
School enrollment	0.001	-0.005**	0.001	-0.001
	[0.002]	[0.002]	[0.002]	[0.002]
Primary teacher-student ratio	1.270***	0.490	0.093	0.323
	[0.346]	[0.348]	[0.328]	[0.312]
% teachers with professional education in primary	0.048	0.049	0.006	-0.087
	[0.031]	[0.032]	[0.071]	[0.075]
% teachers with pedagogical training in primary	0.064**	0.062**	0.024	-0.061
	[0.031]	[0.031]	[0.065]	[0.076]
Observations	27,578	27,578	6,984	6,984
R-squared	0.209	0.228	0.440	0.312
Number of municipalities	1035	1035	1012	1012

Standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1

Notes: All regressions include municipal fixed effects. Complete is a dummy variable equal to 1 when the school has a complete schedule. School socioeconomic status (SES) is an index that goes from 1 (poorest schools) to 4 (richest schools).

Table 2.5. Panel regression results. Dependent variable: 5th grade test scores.

	Language		Mathematics	
	Municipality fixed effects	School fixed effects	Municipality fixed effects	School fixed effects
	(1)	(2)	(3)	(4)
Complete	0.055**	0.044	0.082***	0.082**
	[0.024]	[0.034]	[0.027]	[0.033]
Dummy for urban schools	-0.101***	--	-0.181***	--
	[0.022]		[0.024]	
School socioeconomic status	0.115***	--	0.078***	--
	[0.016]		[0.014]	
School enrollment	-0.002	0.007	-0.009***	0.004
	[0.001]	[0.006]	[0.003]	[0.006]
Primary teacher-student ratio	1.271***	0.509	0.791***	-0.123
	[0.277]	[0.468]	[0.264]	[0.470]
% teachers with professional education in primary	0.033	-0.036	0.039	-0.016
	[0.026]	[0.047]	[0.025]	[0.046]
% teachers with pedagogical training in primary	0.067***	0.048	0.052**	0.041
	[0.026]	[0.046]	[0.024]	[0.045]
Observations	54,192	54,192	54,192	54,192
R-squared	0.161	0.611	0.185	0.618
Year fixed effects	YES	YES	YES	YES
Municipal fixed effects	1093	NO	1093	NO
School fixed effects	NO	26919	NO	26919

Standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1

Notes: Complete is a dummy variable equal to 1 when the school has a complete schedule. School socioeconomic status (SES) is an index that goes from 1 (poorest schools) to 4 (richest schools). All standard errors are robust.

Table 2.6. Panel regression results. Dependent variable: 9th grade test scores.

	Language		Mathematics	
	Municipality fixed effects	School fixed effects	Municipality fixed effects	School fixed effects
	(1)	(2)	(3)	(4)
Complete	0.162***	0.110**	0.137***	0.138**
	[0.033]	[0.055]	[0.034]	[0.063]
Dummy for urban schools	-0.040	--	-0.074***	--
	[0.029]		[0.025]	
School socioeconomic status	0.269***	--	0.168***	--
	[0.021]		[0.017]	
School enrollment	0.001	-0.016***	-0.001	-0.008
	[0.002]	[0.006]	[0.002]	[0.007]
Secondary teacher-student ratio	0.230	0.032	0.548*	0.451
	[0.296]	[0.573]	[0.319]	[0.503]
% teachers with professional education in secondary	0.036	-0.028	-0.014	-0.076
	[0.052]	[0.091]	[0.052]	[0.097]
% teachers with pedagogical training in secondary	-0.005	-0.103	-0.034	-0.056
	[0.047]	[0.090]	[0.051]	[0.097]
Observations	13,612	13,612	13,612	13,612
R-squared	0.287	0.661	0.212	0.611
Year fixed effects	YES	YES	YES	YES
Municipal fixed effects	1,083	NO	1,083	NO
School fixed effects	NO	6725	NO	6725

Standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1

Notes: Complete is a dummy variable equal to 1 when the school has a complete schedule. School socioeconomic status (SES) is an index that goes from 1 (poorest schools) to 4 (richest schools). All standard errors are robust.

Table 2.7. Balancing tests for COMPLETE

	5th grade	9th grade
	(1)	(3)
Enrollment	-0.143***	-0.454***
	[0.018]	[0.091]
Teacher student ratio	0.001	-0.000
	[0.001]	[0.003]
% teachers with professional degree	0.001	-0.025**
	[0.010]	[0.011]
% teachers with pedagogic training	-0.005	-0.014
	[0.010]	[0.013]

Standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1

Note: All regressions include school and year fixed effects. Each column reports the coefficient for COMPLETE, which shows the differences in school characteristics (the dependent variable in the first column) between complete and half schedule schools, controlling for school and year.

Table 2.8. Heterogeneous effects: School fixed effects model using alternate samples

	5th grade		9th grade	
	Language	Mathematics	Language	Mathematics
	(1)	(2)	(3)	(4)
Rural schools	0.056	0.097***	0.116	0.175*
	[0.037]	[0.036]	[0.092]	[0.103]
Observations	*42,446*	*42,446*	*6,121*	*6,121*
Urban schools	0.005	0.009	0.104*	0.106
	[0.068]	[0.077]	[0.063]	[0.073]
Observations	*11,746*	*11,746*	*7,491*	*7,491*
Poorest schools (SES=1)	0.067*	0.093**	0.138	0.200**
	[0.039]	[0.038]	[0.084]	[0.093]
Observations	39,571	39,571	6,785	6,785
Middle-to-poor (SES=2)	0.006	0.087	0.103	0.102
	[0.077]	[0.075]	[0.081]	[0.094]
Observations	*9,298*	*9,298*	*3,564*	*6,785*
Middle-to-rich (SES=3)	0.072	-0.028	0.027	0.001
	[0.134]	[0.143]	[0.152]	[0.162]
Observations	*4,012*	*4,012*	*2,251*	*3,564*
Richest schools (SES=4)	-0.169	-0.132	0.121	-0.029
	[0.142]	[0.177]	[0.170]	[0.330]
Observations	*1,311*	*1,311*	*1,012*	*2,251*

Standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1

Notes: Complete is a dummy variable equal to 1 when the school has a complete schedule. School socioeconomic status (SES) is an index that goes from 1 (poorest schools) to 4 (richest schools). Only the coefficient for COMPLETE is reported. All regressions control for school enrollment, teacher student ratio, percentage of teachers with a professional degree, percentage of teachers with pedagogic training, and include school and year fixed effects. All standard errors are robust.

Table 2.9. School fixed effects model using alternate samples

	5th grade		9th grade	
	Language	Mathematics	Language	Mathematics
	(1)	(2)	(3)	(4)
Switchers from half to complete	0.085	0.055	0.017	-0.263
	[0.114]	[0.113]	[0.155]	[0.180]
Observations	*3,325*	*3,325*	*698*	*698*
Switchers from complete to half	0.054	0.214*	0.079	0.306*
	[0.128]	[0.126]	[0.167]	[0.179]
Observations	*1,990*	*1,990*	*541*	*541*
Schools in 3 periods (balanced panel)	-0.024	-0.017	0.089*	0.096
	[0.041]	[0.039]	[0.053]	[0.063]
Observations	*23,181*	*23,181*	*6,951*	*6,951*
Switchers with longer treatment	0.101	-0.055	0.274***	0.292***
	[0.076]	[0.077]	[0.103]	[0.105]
Observations	*1,227*	*1,227*	*444*	*444*
Switchers with shorter treatment	0.020	0.134	-0.036	-0.030
	[0.121]	[0.134]	[0.207]	[0.182]
Observations	360	360	150	150

Standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1

Notes: Complete is a dummy variable equal to 1 when the school has a complete schedule. School socioeconomic status (SES) is an index that goes from 1 (poorest schools) to 4 (richest schools). Only the coefficient for COMPLETE is reported. All regressions control for school enrollment, teacher student ratio, percentage of teachers with a professional degree, percentage of teachers with pedagogic training, and include school and year fixed effects. Switchers with longer treatment are schools that were half in 2002, switched to complete in 2005 and kept being complete in 2009. Switchers with shorter treatment are schools that were half in 2002 and 2005, and switched to complete only in 2009. All standard errors are robust.

Essay 3

The effectiveness of multigrade classrooms:

Evidence from Colombia's New School model

Abstract

This paper analyzes the impact of the "Escuela Nueva" model (New School or EN) on student achievement, using test score data from SABER 2002 and 2005, a national standardized test administered to 5th and 9th graders in Colombia. EN is an educational model originally designed to improve the effectiveness of rural schools. It is characterized by multigrade classrooms (i.e., one instructor teaches students in various grades in the same classroom), a child-centered curriculum, flexible systems of grading and promotion, intensive teacher training, and parental involvement. To mitigate the concerns about systematic selection of schools into EN that might bias the estimations of the EN impact, this study implements a school fixed effects model that controls for time-invariant characteristics within the school. Results show that among schools that switched models between 2002 and 2005, the cohorts of 5th grade students exposed to EN have on average 0.135 of a standard deviation higher language test scores than cohorts exposed to other models, while there is no statistically significant impact on switching to EN for 9th graders. The impact of EN is largest among rural schools and the poorest schools in the sample. The paper also implements propensity score matching methods to identify a control group of schools that have similar probabilities of implementing EN.

Keywords: quality of education, student achievement, education policy.

JEL codes: I20, I21, I28.

3.1 Introduction

Escuela Nueva ("New School", or EN) is an educational model that originated in Colombia in the 1970s, "to improve the quality, relevance and effectiveness of the schools in the country (FEN, 2014)." It was originally targeted at rural multigrade schools where one or two teachers simultaneously teach all grades. These schools were very common in rural areas in Colombia, and in other developing countries, where there is a low population density, and generally fewer resources (FEN, 2014). The EN model is known for encouraging multigrade classrooms, flexible grading and promotion, intensive teacher training, and parental involvement. With support from the central government, the *Escuela Nueva model* (EN) began expanding nationally in the 1980s, and was later adapted to urban contexts and to secondary grades (6^{th} to 9^{th}) in rural areas (known as, Rural Post-primary).

The EN model has often been cited as a best-practice in rural school reform (Schiefelbein, 1991; PREAL, 2003), and has been widely publicized as a successful model by its creators, the Escuela Nueva Foundation (FEN, for its acronym in Spanish),[31] the government, and international development organizations. However, despite its expansion at the national level, and its apparent success, there is relatively little evidence about the impact of the program in the country. Most of the existing evidence has been descriptive in nature, or based on comparing the average outcomes of students attending EN schools with those of students attending traditional schools. This type of evaluation provides weak causal evidence of the impact of EN because of possible endogeneity and

[31] The Escuela Nueva Foundation is the creator and main promoter of the EN model.

selection problems, as discussed in section 3.4. Additionally, there have not been recent impact evaluations of the EN model.

This study estimates the impact of the Escuela Nueva model on 5th grade and 9th grade student achievement, using test score data from a national assessment (SABER) from 2002 and 2005. It contributes to the literature by using three identification strategies that mitigate some of the main selection issues found in previous studies, and allows making causal statements about the impact of EN on student achievement with a higher level of confidence.

Comparing simple average outcomes between EN and traditional schools is likely to reflect differences in other characteristics that also affect student achievement. Therefore, estimating the impact of EN using a traditional OLS model might provide biased estimates, since schools that implement EN may have unobservable characteristics that are related to student achievement. To address these issues, this study first implements a Panel OLS model, which includes municipality fixed effects and a dummy variable for 2005 that controls for anything that affected all schools equally in that year.[32] Second, it implements a school fixed effect strategy that exploits within-school changes in the type of model implemented (to/from an EN model). By comparing test scores of cohorts of students in the same school before and after they implement the EN model, I am able to control for all unobserved time-invariant characteristics of the school that may bias cross-sectional estimates, although some bias from time-varying unobservable characteristics might remain. Finally, I implement a Propensity Score Matching (PSM) strategy, which uses several matching methods to construct a treatment and control

[32] The omitted or reference variable is a dummy variable for 2002.

group. These groups are very similar in observable characteristics, and have very similar probabilities of implementing the EN model, although only those in the treatment group actually implement EN. The PSM estimates the average treatment effect on the treated (ATT), by comparing test scores between schools in the treatment group with schools in the control group.

I also examine heterogeneity in the impact between younger (5^{th} grade) and older students (9^{th} grade). The impact of EN might be larger for the younger students, because they are more likely to gain more from the elements of the EN model. For example, younger students are more likely to benefit from a child-centered curriculum, where they can make progress at their own pace, and where teachers give special attention to each student's learning process. Additionally, I look at heterogeneity in the impact of EN by school socioeconomic status and location (urban/rural) by applying the school fixed effect strategy to different subsamples of each type of school. Additionally, I analyze heterogeneity by fidelity of implementation. The EN model comprises many features, and requires the use of many inputs. However, inputs are not equally provided in all EN schools and classrooms, so schools that claim to implement EN are not necessarily fully implementing the model. I look at heterogeneity by fidelity of implementation by comparing schools in four departments have had traditionally well-established EN models, with schools in the other departments.

The results of this study show that among schools that switched to or from the EN model between 2002 and 2005, the cohorts of 5^{th} grade students exposed to EN have on average 0.135 of a standard deviation higher language test scores than cohorts exposed to other models, while there is no statistically significant impact on switching to EN for 9^{th}

grade test scores. The impact of EN is largest among rural schools and the poorest schools in the. Using Propensity score matching, the estimated average treatment effect of EN is about one-fifth of a standard deviation, but only for 5th grade language test scores.

The rest of the paper is organized as follows: Section 3.2 describes the background on the *Escuela Nueva* model in Colombia. Section 3.3 reviews the relevant literature. Section 3.4 discusses the empirical strategies. Section 3.5 describes the data. Section 3.6 presents the results, and Section 3.7 concludes.

3.2 The Escuela Nueva model

The EN was created based on the unitary schools model promoted by the United Nations Educational, Scientific, and Cultural Organization (UNESCO) in the 1960s. It was first implemented on a large scale in Colombia in 1975. The model was originally targeted to rural schools, which at that point were very inefficient and provided low-quality education: there were high grade repetition and drop-out rates, schools were geographically isolated, there was a disconnect between the schools and the community, teacher motivation and opportunities for training were low, the learning contents were not very relevant nor appropriate, and the methods of instruction were passive and based on rote learning.[33] Additionally, because of low population density, scarce resources and low infrastructure, these schools were frequently organized as multigrade schools, with one or two teachers simultaneously teaching all of the primary grades (1st-5th grade) in one single space or classroom.

[33] See for example: Colbert, V. (1999); Forero, Escobar and Molina (2006).

EN was initially implemented in three of the country's departments in 1976, and then expanded to another five by the following year. In 1987, as part of the "Plans for the Promotion and Universalization of Education," the Ministry of Education adopted EN as a national public policy, implementing the EN model in approximately 24,000 schools in the country in that year, and gradually expanding to different departments (FEN, 2014).[34] The EN model became a central part of the government's project to improve the quality and access to primary education, as well as the main strategy for achieving universal primary education in rural areas. Even though the importance given to rural education in general, and to EN in particular, has changed through time with shifts in the political context (McEwan & Benveniste, 2001), EN still seems to be a central part of the national education policy. Between 2002 and 2005, EN was expanded to 28 territorial entities (mainly departments or municipalities). In this period, 3,162 teachers were trained, and 200,000 students participated in EN in 592 municipalities. In the same period, 34,900 new schools spots were created through the EN schools. In 2005, 862,000 students were officially enrolled in EN schools (FEN, 2006). And by 2006, approximately 20,000 of the 29,896 rural public schools claimed to follow the EN model (Forero et al., 2006).

The main goal of the EN model is to facilitate the learning process in multigrade classrooms by promoting active, participatory, and cooperative learning among school students. In a typical EN school, children of different ages and grades share a classroom and teacher, but the classroom space, the learning guides, and teachers are organized to facilitate multigrade schooling. EN it not simply a name given to multigrade schools, but a model that exploits this characteristic to promote a type of classroom pedagogy that is

[34] FEN (2014). Retrieved from: http://www.escuelanueva.org/portal/es/modelo-escuela-nueva/historia.html

child-focused, where students learn at their own pace, and where there are flexible schedules, and systems of grading and promotion.

EN also has a strong focus on teachers, and teacher training. Teachers attend workshops in which they learn how to implement the EN model in their school and their communities, and they are trained on teaching strategies in the context of multigrade classrooms. For example, they are trained in the use and adaptation of individual study guides, and learn to apply the guides in the specific context of their schools. They are also trained in group management and the simultaneous use of curricula for children of different ages, and to organize and use the school and classroom libraries in the particular context of multigrade classrooms. Teachers also attend a sequence of workshops that ensures the training and follow-up of teachers. These workshops have become known as "microcenters," where teachers learn through the evaluation process, analyze individual problems, and construct solutions in a participatory learning process.

EN also has a very strong civic component: it promotes a model of student government, which aims to promote civic values and peaceful social interaction. Also, it works on strengthening the relationship between the school and the community.

3.3 Extant literature

Several studies have analyzed the EN model. However, most of them are purely descriptive, and few actually estimate the impact of the model. For example, describing the EN, a World Bank (2008) report states: "Colombia's world-renowned and internationally replicated *Escuela Nueva* (EN) program has improved student achievement in rural areas by enabling students to progress through a flexible curriculum,

by engaging them with active pedagogy supported by teacher training, and by adapting to local needs through democratic decision making and community engagement (p. xi)." And according to the Minister of Education:

> A very preliminary analysis of fifth grade student achievement indicates that in all localities and in all subjects, the average achievement of schools that use EN model is higher than the average achievement in the locality. But an in-depth analysis is needed. There is some preliminary evidence that for 2005, the difference in student achievement between rural and urban areas is small, and rural areas are ahead of urban areas in civic competences of 9th graders (Author's translation based on FEN, 2006, p.22-23).[35]

Three studies are perhaps the most comprehensive comparative evaluations of EN and traditional schools. Rojas and Castillo (1988) used a dataset collected in 1987 in 11 departments to compare student achievement in math, Spanish, civic behavior, self-esteem, and creativity in 168 EN schools with student achievement in 60 traditional schools. They concluded that EN schools had significantly higher test scores than traditional schools in civic behavior, social self-concept, 3rd grade math, and 3rd and 5th grade language. The study also included some qualitative data that showed that EN schools had higher levels of participation in community activities and high levels of teacher satisfaction with the methodology, training courses, and self-instructional learning guides. However, the department and schools in the study were not selected randomly: departments were selected where the EN was relatively more developed, and schools were selected where the program had been in place for at least three years. Therefore, their results reflect more the impact of the "ideal" EN on educational outputs, and cannot be generalized to all EN schools (McEwan, 1998).

[35] FEN (2006).This was part of a presentation given by the Minister of Education at that time,+ Cecilia Maria Velez, at the Second International Congress of New Schools in 2006.

Psacharopoulos, Rojas, and Velez (1993) used the same data as Rojas and Castillo (1988) to estimate the impact of the EN model. They compared means and found a statistically significant difference between student achievement in EN and traditional schools in 3rd grade Spanish and math, creativity, civics, and self-esteem, and 5th grade Spanish. They also used ordinary least squares regressions to estimate a production function for 3rd and 5th grade students in EN and traditional schools. With this approach, they found that EN had a positive and statistically significant difference on Spanish and math tests in 3rd grade, Spanish in 5th grade, and civic behavior for the pooled group of students. These authors also used a probit model to estimate the probability of dropping out, and found that fifth graders in EN schools had a lower probability of dropping out than students in traditional schools.

McEwan (1998) criticized the Psacharopoulos et al. (1993) paper by arguing that they left important inputs out of the estimation of the production function. In particular, they did not take into account variables that reflect the availability of EN inputs such as libraries and textbooks. Neither had they considered that there are different degrees of implementation of the EN models, or that some traditional schools might apply some features of the EN. In contrast, McEwan (1998) measured inputs that are particular to the EN model, or at least are typically associated with EN, and included them in his estimation of the education production function. The study used data from a 1992 survey of 3rd and 5th grade students in 52 randomly chosen schools (24 EN schools and 28 traditional schools), in 3 departments of the country: Cauca, Nariño, and Valle. Students in these schools were tested in Spanish and math, and, additional to the EN variable, other inputs that are particular to the EN model were also included in the regressions,

such as the presence of a school library, textbook availability, frequency of group work in the classroom, and supervisory visits to the school in the past year. McEwan found that student achievement in 3rd and 5th grade math and Spanish tests was statistically higher in EN schools. The impact was larger for 3rd graders than for 5th graders, and larger for Spanish tests than for math tests.

Finally, Forero et al. (2006) analyzed the impact of EN but on a totally different outcome: the peaceful interaction of children in Colombia ("convivencia"). The authors used hierarchical multilevel models to estimate the impact of student, classroom, school, and municipality variables on the peaceful interaction of students. In contrast to previous studies, this one accounted for the differences in the level of implementation of the EN model by constructing an implementation index, which at the same time was composed by classroom and teacher implementation indexes. The classroom implementation index was measured by the existence and practice of physical and organizational EN features, like (group) desk organization, the use of personal study guides, the frequency of group activities, the existence and use of classroom libraries and learning corners, and curricular flexibility. The teacher implementation index measured the number and level of training workshops attended and the level of micro-center activities. The authors concluded that the use of EN methodologies had a significant impact on the peaceful social interaction of children. They also found significant positive effect of EN on some familiar behavior related to home educational practices and to the influence of the school in parent participation in the community.

In sum, there is some evidence that the EN model has a positive impact on student achievement and other outcomes. However, as briefly mentioned above for each of these

studies, all of them have weaknesses because of the data or methods used. McEwan's study was based on a random sample of schools and included data on important EN inputs, but given that his sample comes from 3 departments, his results may not be generalizable to the whole EN school population.

This study aims to improve knowledge about the effectiveness of EN by using a much larger sample of EN schools that covers most departments of the country. Additionally, it contributes to the literature in various other ways. First, it analyzes the impact that the EN model has had on student achievement in the last decade, providing an updated estimate of the impact and analysis about the relevance of the policy during this period. The second contribution is to estimate the impact of EN using a new, large dataset that has not been used in impact evaluations of the EN before, and that has been developed by the ICFES (the national agency for the quality of education) in recent years. The only recent study (hired by FEN, but still not publicly available) analyzes the impact of the program on student achievement using the SABER tests from 2009, but it uses a probabilistic sample of schools in only one department of the country (Boyacá). Therefore, it has low external validity, and its findings cannot provide a picture of the overall impact of the EN as a national public policy. This study expands the evidence on the impact of EN by taking into consideration student achievement of all 5^{th} and 9^{th} graders in a large sample of schools from most departments in the country.[36] Finally, by comparing the impact of EN in departments with well-established EN models, with the impact in departments with less-established EN models, this study provides some

[36] If there are not enough EN schools in urban areas in the dataset, the sample might be restricted to rural areas.

suggestion on the consequences of the different levels of implementation of the EN model.

3.4 Empirical Strategy

To estimate the impact of the EN model on 5th grade and 9th grade student achievement, ideally one would assign some schools to a "treatment group," which receives the intervention (in this case the EN model), and to a "control group," which does not receive the intervention, and which is used as a comparison group. If this were the case, because schools are randomly assigned to each group, the characteristics of both groups should not be systematically different. In other words, the two groups will be equal in expectation (Murnane & Willett, 2011). Therefore, any difference in the outcomes between the two groups after the intervention can be attributed to the intervention. However, the EN is not a randomly assigned intervention.

The EN model was essentially created to address the problems of rural basic education in Colombia, so it is mainly present in schools in rural areas of the country. Also, since the 1990s, the country has been decentralizing the administration and provision of education. As a result of this process, nowadays each department and municipality has the autonomy to make most decisions in terms of the models and the allocation of resources to their schools.[37] Therefore, when a municipality is interested in implementing EN, it elaborates an assessment of its current situation and needs, and then

[37] Most municipalities have their own resources, and receive resources from the central government that they managed. They also make most of the decisions of their education system. So for simplicity, I will from now on refer to municipalities and their Secretary of Education as the main decision-makers. However, in some cases, small municipalities with a low institutional capacity are "managed" by the administration of its respective department. In those cases, it's the Secretary of Education of the department who manages the resources and makes most of the decisions.

decides on the coverage of the model, and in which schools it will be implemented. In urban areas, schools might also manifest their interest in applying the model, and must elaborate the same type of diagnosis and present it to the Secretary of Education, who makes the final decision. When the Secretary of Education has decided on the implementation of EN in a school, he or she has to provide the necessary training and inputs or materials needed for implementing the EN model. As mentioned before, the amount of EN inputs is not the same in all schools, but perhaps the most important thing that the government provides is teacher training in the EN model, and the individual learning EN guides. It could also provide infrastructure, such as round tables, or classroom libraries, but this varies depending on the municipalities' available resources.

In this context, if one compares average student achievement in SABER in schools that implemented the EN model to average student achievement in schools that did not implement the EN model (traditional schools), it is possible that there is a selection problem. I do not expect to have selection based on students' or parents' preferences, since most EN schools are located in rural isolated areas, where there is practically no school choice (McEwan, 1998). However, it is possible that there is a selection problem at the school level. For example, if EN schools are systematically different from traditional schools (e.g., poorer, or with fewer resources), the differences in student achievement between EN and traditional schools might reflect all these characteristics and not only the impact of the EN model. If selection is based only on observable characteristics, an OLS regression model would allow estimating the average differences in student achievement, while controlling for those characteristics. However, there might also be a selection problem based on unobservable characteristics, for

example, if EN schools have more motivated principals and teachers than traditional schools. To address these selection and endogeneity problems I implement the following estimation strategies.

i. Estimation Strategy I: Panel OLS

If selection into the EN model in a year is only based on observable characteristics, one could control for most of the observable characteristics that are causing the selection problem by running an OLS model with controls. The following regression is run by subject (language and math) and grade (5^{th} grade and 9^{th} grade).[38]

$$\text{SABER}_{st} = \beta_0 + \beta_1 \text{EN}_{st} + \beta_2 \text{X}_{st} + \text{d}_t + \text{d}_m + \varepsilon_{st} \quad (1)$$

Where *SABER* is the average test score for school *s* in year *t*, *EN* is a dummy variable equal to 1 if the school implemented the EN model and 0 otherwise, X is a vector of school characteristics, d_t are year fixed effects[39] (i.e., a 2005 dummy variable), d_m are municipality fixed effects, and ε is the error term.

ii. Estimation Strategy II: School fixed effects model

The OLS model (1) allows controlling for observable characteristics that could be causing selection into the EN schools. However, schools could be selecting into the EN model based on unobservable characteristics. For example, it could that be that schools

[38] The EN model for secondary education, Rural post-primary, has expanded to only some departments, so there might be fewer observations and variability for the 9^{th} grade sample than for the 5^{th} grade sample.
[39] The 2005 dummy variable controls for any observed or unobserved policy changes or events that affected all schools equally in 2005.

with the most motivated principals are the ones that implement the EN model. If there is selection on unobservable characteristics, like motivation, the differences in outcomes might reflect these characteristics, and therefore, estimations of the impact of EN in model (1) might be biased.

To address the possible endogeneity problems caused by the selection of schools, I implement a school fixed effects model, which allows controlling for any time invariant characteristics of the school that could be related to the implementation of EN. Basically, this is within-school, across-cohort design: it compares different cohorts of students that were exposed to different models (EN) within the same school. The following is the school fixed effects specification, to be run for each subject and grade:

$$SABER_{st} = \beta_0 + \beta_1 EN_{st} + \beta_3 X_{st} + d_s + d_t + \varepsilon_{st} \quad (2)$$

Where d_s are the school fixed effects and d_t are year fixed effects, and the other notation represents the same variables as in (1).

The main identifying assumption is that parents and students are not selecting into or out of school because they know that the school is going to implement EN. In impact evaluation terminology, this means that 5^{th} graders and 9^{th} graders did not seek to be treated, or alter their condition (enrollment in one school), in order to get the treatment (EN).

There are some implications of omitting particular EN inputs, such as the use of the EN learning guides, from models (1) and (2). If EN inputs are applied as a package, and one assumes that they have a positive impact on the outcome, leaving them out is likely to bias upward the EN dummy coefficient. However, if some traditional schools

implement EN elements, and some EN schools only partially implement the EN model, the coefficient will reflect an average effect of the EN model (McEwan, 1998). Because I do not have data on which of the inputs or elements of the EN model are actually implemented in each school, the coefficients reflect the impact of the EN model as it was actually implemented compared to other schools (that may implement some EN elements), rather than the impact of the full implementation of the program where there is 100% compliance with all the features of the model and no use of EN features in control schools.

Given that there is variation in the level of implementation of EN, I implement the school fixed effects model using a subsample of schools that are located in departments that have had a traditionally well-established EN model, and therefore one could expect a better implementation of the EN. This allows me to evaluate the heterogeneity in impact by fidelity of implementation. According to the FEN, these departments are the ones located in what is known as the "Coffee Region," comprised of the departments of Antioquia, Caldas, Armenia, and Quindío. The assessment of the "quality" of implementation is subjective,[40] but at least it sheds some light on the type of differences given by different levels of implementation of the model. However, the results of the model using this subsample might be more biased than the pooled model, due to higher likelihood of selection bias. For example, if parents know that these departments have a "well-established EN model," they are more likely to select the EN schools.

[40] This assessment, as well as the suggestion to look at these departments, was discussed in a video conference with staff from the Research Department at the Escuela Nueva Foundation, in May 2012.

iii. Estimation Strategy III: Propensity Score Matching

An alternative strategy that might help mitigating the selection bias problem is to implement a Propensity Score Matching (PSM) approach (Rosenbaum and Rubin, 1983; Heckman, Ichimura & Todd, 1998). Implementing PSM could provide a better controlled model than that in the Panel OLS model (1) and the school fixed effects model (2) described above, because one compares schools that are very similar in observable characteristics, and that given these characteristics, also have very similar probabilities (propensities) of implementing EN. However, some of these schools actually implemented EN (these constitute the treatment group) while others did not (these form the control or comparison group). In this approach, systematic differences between the treatment and comparison schools, with the same values for observable characteristics, can be attributed to the EN model, under the assumption that EN treatment satisfies some form of exogeneity (Caliendo & Kopeinig, 2008). This is usually referred to in the literature as the unconfoundedness assumption (Rosenbaum and Rubin (1983), selection on observables (Heckman and Robb, 1985), or the conditional independence assumption (Lechner, 1999).

Estimating the propensity scores

To use this methodology, implementing the EN model is seen as a type of intervention in a school. EN schools are taken as the treatment group, and a comparison group is "built" using observed schools' characteristics. This group is comprised of schools that do not implement EN (do not receive the treatment), but that have the same

(or close to the same) probability of implementing EN (receiving treatment) as any of the schools in the treatment group.

The first step of the PSM is estimating the propensity score for each school, that is, the probability of implementing EN for each school. In principle, any discrete choice model can be used, but there is usually a preference for logit or probit models (Caliendo & Kopeinig, 2008). I use a probit model. In this model, the school variables that could affect the probability of implementing EN are included as covariates or independent variables, as shown in the following equation:

$$\Pr(t_s) = \lambda_1 X_s + \varepsilon$$

Where t_s takes the value of one if the school implements EN in 2005, and zero if does not, and X is a vector of school variables that affect the probability of implementing EN in 2005, namely: previous year test scores, the location of the school (rural or urban), the socioeconomic status of the school (SES), total school enrollment, the teacher student ratio, the percentage of teachers with a professional degree, the percentage of teachers with a pedagogical degree, and if the school has a complete schedule. This strategy uses test scores and other school variables from 2002 to estimate the probability of implementing Escuela Nueva in 2005. Therefore, for this particular strategy, I use the 2005 cross-section, rather than the 2002-2005 panel.

Matching methods or algorithms

The probability model (probit) described above provides the school's probability of being treated (a propensity score), and with these propensity scores it is possible to

match schools that have the same (or similar) probability of implementing EN, but that differ from the treatment group only by the intervention being evaluated (EN). The schools that are matched are part of the "common support." There are different ways to construct these treatment and control groups. I use three of the most common matching methods used in the literature: nearest neighbors (NN) with replacement, caliper, and kernel matching.

Nearest neighbor

The nearest neighbor (NN) matching is the most straightforward method of matching. In this case, a school in the comparison group is chosen as a matching partner for a school in the treatment group that is closest in terms of the propensity score. NN matching with replacement means that schools in the comparison group can be used more than once as a match for schools in the treatment group.[41] By allowing replacement, the average quality of the match increases (because there are more controls observations to match) and the expected bias decreases (Caliendo and Kopeinig, 2008), but there is a loss in efficiency. I use 1, 5, and 10 NN, which means that 1, 5 and 10 matching partners respectively are chosen for each treated school (in the case of 5 and 10 NN, this is referred to as "oversampling"). The larger the number of neighbors used, the smaller the variance, resulting from using more observations to construct the counterfactual for each school in the treatment group, and the larger the bias, resulting from on average poorer matches.

[41] Another option is to do NN matching (one-to-one) without replacement, meaning that an untreated or control school can be considered only once as a match. If the number of schools in the treatment group is larger than the number of schools in the control group, it is possible to start with the comparison schools and match to them schools that are in the treatment group.

Caliper matching

If the closest neighbor is far away, then using NN might cause some bad matches. To avoid this, one can impose a tolerance level on the maximum propensity score distance (caliper). This means that a school in the control group is chosen as a matching partner for a treated school that lies within the caliper (propensity range) and is closest in terms of the propensity score (Caliendo and Kopeinig, 2008). This is a way of imposing a common support condition, because not all participant schools are matched. By imposing a caliper, the bias decreases, but there is a loss in efficiency. Smith and Todd (2005) note that is if difficult to know a priori what a choice for the tolerance level is reasonable. I try using three different levels for the caliper: 0.01, 0.05, and 0.1 (Following Mueser et al., 2007).

Kernel matching

Both the nearest neighbor and Caliper matching use only a few observations from the comparison group are used to construct the counterfactual outcome of the treated schools. The Kernel matching (KM) is a nonparametric estimator that uses weighted average of almost all schools in the control group to construct the counterfactual outcomes, achieving in this way a lower variance because more information is used. However, the problem with this method is that it might use observations that are bad matches. Therefore, this method results in an increase in the potential bias of the estimation (Caliendo and Kopeinig, 2008).

Stratification and Interval matching

Another alternative is to divide the common support of the propensity score into a set of intervals (or strata), and calculate the impact by comparing the mean difference in outcomes between the treatment and control observations within each interval. Although it is not clear how many intervals should be used. Caliendo and Kopeinig (2008) suggest checking if the propensity score is balanced within an interval. If they are not, that means that the intervals are too large, and need to be split into more intervals.

PSM estimator

The matching methods above are used to construct treatment and control groups that have similar probabilities of implementing EN. Having the common support with treated schools and control schools, it is possible to find the effect that EN has on test scores by using the PSM estimator, which compares the outcome of treated schools with the outcome from the control group members. This effect is generally measured as the difference in the means in the outcome (SABER 2005 test scores) between the treatment and control groups, through a parameter known in the literature as the average treatment on the treated (ATT). The ATT shows the difference in test scores of those schools that implement EN with regards to their test scores if they had not implemented EN. This difference is attributed to the treatment, in this case, to the implementation of EN.

3.5 Data

This study uses data from an extensive panel dataset that contains test score results from the 2002-2003 and 2005-2006 rounds of SABER 5th and 9th, which are national standardized language and math assessments administered every three years to students in 5th grade and 9th grade in private and public schools nationwide.[42] The main goal of these tests is "to contribute to improving of the quality of Colombian education, by carrying out periodic measurement of the development of competences in students enrolled in basic education, as an indicator of the quality of education (MEN, 2013).[43] The SABER 5th and 9th are designed to evaluate competencies or abilities in language (Spanish), mathematics, and science, at the end of each basic education cycle: they are administered to students in 5th grade, which is the end of "basic primary cycle," and 9th grade, which is the end of the "basic secondary cycle." SABER tests are the only national achievement tests administered to students in these grades and that have nationwide coverage,[44] they are therefore one of the best available assessment of primary and secondary student achievement in the country.

The SABER panel dataset used in this study comes from the ICFES, the national agency responsible for the quality of education. Initially test score results were not

[42] The SABER 2002 was administered in different moments in 2002-2003, but most tests for 5th grade and 9th grade were conducted in 2002, so I refer to this test as SABER 2002. In the same way, the SABER 2005 was administered in different moments in 2005-2006, but I refer to this test as SABER 2005.
[43] Author's translation from the Ministry of Education's website. Retrieved from: http://www.mineducacion.gov.co/1621/w3-article-244735.html, on June 1st, 2013.
[44] The other national standardized test administered to basic education students corresponds to the high school exit exam called SABER 11 (previously known as "ICFES"). The SABER 11 is administered annually to all students in 11th grade, the last grade of high school. Students are required to take these tests in order to graduate, and the test score results are a common requirement in order to apply to most universities or higher level institutions. Additionally, some international tests are administered to a sample of primary and secondary students. For example, the Organization for Economic Cooperation and Development's (OECD) Programme for International Student Assessment (PISA), which test 15-year old students, and the U.S.'s Trends in International Mathematics and Science Study (TIMSS), which tests students in 4th grade and 8th grade.

comparable across time, because the 2002 SABER used different evaluation instruments, and the methodology for implementing the tests changed in 2005. However, in 2008 the ICFES united with The World Bank to make the test scores fully comparable across the different rounds. This resulted in a dataset that contains comparable student achievement from 2002, 2005, and 2009. This study draws on this dataset because it allows comparing changes in performance in SABER over time.

The SABER panel contains test scores and some basic school identification information, but it does not contain information on school characteristics. Because it is important control for school characteristics in order to implement the identification strategies, I merge the SABER dataset with data coming from the DANE C-600 school survey. The C600 is an annual survey administered by the National Administrative Department of Statistics (DANE, for its acronym in Spanish) to all schools in the country, in order to collect information on basic school characteristics. For example, the survey collects information on the school's type (public or private), location (urban/rural, municipality and department), total enrollment, number of students in each grade, number of teachers, teachers' educational level, type of school schedule (e.g., complete, morning, or afternoon. These variables are used as control variables for all regression models in this study.

The C600 also contains a question regarding the methodology or model that the school implements. For primary grades, one of the models that a school can implement is *Escuela Nueva*, and for secondary grades, Post-primary, which is an adaption of Escuela Nueva for secondary education. Therefore, I use this information to construct the main variable of interest in this study: EN, a dummy variable equal to 1 if the school

implements the Escuela Nueva model in primary grades, or the "Post-primary" model in secondary grades,[45] or zero if it implements any other model. The variable that identifies the school model does not appear in the 2009 dataset of the C600, which requires dropping all data for 2009 in this study.

In order to create a school panel with data from SABER and the C600 datasets, I first aggregate each of these at the school level, and then merge them to each other using the school identification code.[46] To be part of the panel, a school has to have information from both SABER and C600 for a given year, but it does not necessarily need to have information for the two periods of time (i.e., it is an unbalanced panel).

Table 3.1 contains the descriptive statistics of the school panel for the 5th grade and 9th samples. The average language and math test scores went down between 2002 and 2005 in both the 5th grade and 9th grade samples. The largest decrease was for 5th grade math test scores, where there was a tenfold decrease in the mean test score (from -0.01, to -0.10), while the 9th grade math test scores had the smallest decrease (from -0.09 to -0.12). The language test scores also went down from -0.04 to -0.15 for 5th grade, and from -0.13 to -0.23 for 9th grade. Most schools in the 5th grade sample are located in rural areas (81.2% in 2002 and 74.8% in 2005), although around 40% of schools in the 9th grade sample were located in rural areas. The socioeconomic status of schools, is, on average, lower in the 5th grade than in the 9th grade sample (1.4 v. 1.8-1.9), but it does not change significantly between 2002 and 2005 in any of the samples. Approximately 73

[45] Given that the Post-primary model was created as an extension of the EN model to secondary grades, I refer to this as the EN model for secondary grades.

[46] The SABER dataset contains test scores at the individual level, while the C600 contains information at the school-schedule level. For schools that have only one schedule (e.g., a complete schedule), the data from the C600 variables would not change when aggregating by school. However, when a school has two shifts (e.g., a morning and afternoon schedule), the aggregated school level variables will be either the sum or the average between the two shifts.

percent of schools in the 5th grade samples are schools with the lowest socioeconomic status (SES=1), while less than half of the schools in the 9th grade sample are classified in that same socioeconomic status group. Similarly, the percentage of schools with the highest socioeconomic status is very low in the 5th grade sample (2.2 percent in 2002 and 2.8 percent in 2005). That same percentage is higher in the 9th grade sample, but still only about 7-8 percent of schools in this sample come from the highest socioeconomic status. Average enrollment increases over time in both samples, while the teacher student ratio remains steady in time. The percentage of teachers with a professional degree and with pedagogic training increases between 2002 and 2005 in both samples, although both increases are larger in the 5th grade sample. Finally, the percentage of schools with a complete schedule increases in both samples: it goes from 27 percent to 34.5 percent in the 5th grade sample, and from 21.6 percent to 24.2 percent in the 9th grade sample.

The bottom rows in Table 3.1 refer to the percentage of schools in each sample that implement the EN model, and the amount of schools that switched from or to EN between 2002 and 2005. In the 5th grade sample, about 63 percent of the schools in 2002, and 56 percent of the schools in 2005 implemented EN. About 2.5 percent of the schools in this sample switched to EN, while 4 percent of schools switched from EN to a different model. The number of schools that implement EN in 9th grade is much smaller than in 5th grade, and corresponds to about 10 percent of the schools in the 9th grade sample. For this sample, there is a small percentage of schools that switched to EN (0.7 percent), and from EN to another model (1.2 percent). The small percentage of schools that switched models within the period of study reduces the power to detect effects using the schools fixed effects strategy, and threatens the external validity of the results, particularly if these

schools are systematically different from the rest of the sample. I analyze the characteristics of these schools, and they are not systematically different from the rest of the schools,[47] although unobservable characteristics may remain.

One of the most important endogeneity concerns is that schools with certain characteristics might systematically select to implement EN. Table 3.2 and Table 3.3 look at this issue by presenting descriptive statistics for the 5th grade and 9th grade samples by type of model, the difference between the two types of model, and t-statistics evaluating the statistical significance of the difference. Most of the differences are statistically significant (bolded in the table), indicating that schools that implement EN seem to be different than those that do not, in terms of these characteristics. Table 3.2 compares the descriptive statistics for the 5th grade language sample. Two things that stand out in this table are that most EN schools are located in rural areas, while only around half of the other (non-EN) schools are located in rural areas, and that the socioeconomic status is lower in EN schools. EN schools are also much smaller on average than non-EN schools, and tend to have a smaller percentage of teachers with professional or pedagogic training. There is also a larger percentage of EN schools that have a complete schedule. The teacher student ratio is very similar between EN and non-EN schools, although the difference is also statistically significant.

Table 3.3 shows the descriptive statistics by type of model for the 9th grade language sample. Similar to the differences by type of model for the 5th grade sample (Table 3.2), compared to non-EN schools, the EN schools have lower socioeconomic status, higher enrollment, a smaller percentage of teachers with professional education and pedagogic training, and a larger percentage of schools with a complete schedule. All

[47] Available on request.

the differences between EN and non-EN schools are statistically significant, except for the average math test scores.

3.6 Results

i. Panel and school fixed effects models

Table 3.4 presents the regression results for the panel model (Estimation strategy 1) and the school fixed effects model (Estimation strategy 2), for the 5th grade samples. The results using the panel model show that schools that implement EN achieve statistically significantly higher test scores than non-EN schools in language (0.152) and math (0.206), controlling for municipality and for anything that could affect all schools equally in 2005.

The Panel regressions (Columns 1 and 3) also show that, as expected, the school socioeconomic status is positively and statistically significantly associated with higher test scores, although the estimate is more than twice as large for language as for math (0.094 v. 0.043, respectively). The primary teacher-student ratio coefficient is also statistically significant, and positively associated with 5th grade test scores: increasing the primary teacher student ratio increases language test scores by about one full standard deviation (column1), and increases math test scores by 0.69 of a standard deviation. However, given that the mean teacher-student ratio for this sample is 0.04, the one point increase associated with such an increase in test scores is unlikely. The coefficient for the rural school dummy is statistically significant, and positively associated with higher math test scores, although it is not significant for language test scores. A larger school

enrollment seems to harm math test scores, although only marginally (-0.01). Finally, schools that have a complete schedule (7 hours of instruction) have math test scores that are 0.064 of a standard deviation higher than schools that do not have a complete schedule (typically 4-5 hours of instruction). In contrast with previous estimates (Hincapié, 2014), the complete schedule coefficient is not statistically significantly associated with language test scores, and although it is statistically significantly correlated with math test scores as in the previous study, the magnitude is slightly smaller (0.064 v. 0.082).

The school fixed effects model (columns 2 and 4) assesses how changes in school characteristics across time impact school performance in SABER test scores. The school fixed effects models show that among schools that switched models between 2002 and 2005, the cohorts exposed to EN have language test scores that are 0.135 of a standard deviation higher than cohorts exposed to non-EN models. Although this coefficient is only marginally statistically significant (10%). Cohorts exposed to EN do not seem to have higher math test scores than those exposed to non-EN models, although the point estimates remains positive, around one-tenth of a standard deviation.

The school fixed effects estimates are slightly smaller than those estimated with the panel specification (column 1). One way of explaining this is that implementing EN has a higher role in explaining test scores variation between schools (given by the Panel OLS estimates), than test score variation within school (given by the school fixed effects estimates). However, the fact that the school fixed effects is able to mitigate some of the most important endogeneity concerns, particularly those caused by selection of schools,

means that the school fixed effects are likely to provide better, less biased estimates of the impact of implementing EN.

Table 3.5 presents the panel and school fixed effects models for 9th grade language and math test scores. In contrast to the 5th grade results, implementing EN does not have a statistically significant impact on 9th grade test scores, and does not have a significant role in explaining either between school variation in test scores (panel models in Columns 1 and 3), or within school variation (school fixed effects model in Column 2 and 4). This suggests that the EN model might not be as successful in increasing student achievement in 9th grade, as it is in 5th grade. Although I am not able to directly test for the reasons that this model is not successful in increasing 9th grade test scores, I hypothesize that this might be explained by the shorter implementation of the EN model in secondary grades. This model (also known as Post-primary), was created in 1999 (more than 20 years after the start of the EN model) as an extension of the EN model (MEN, 2014). Therefore, the model has only been partially implemented in secondary grades.[48] It might be possible that more time is need to consolidate the model before it is possible to see any significant impact on test scores.

In terms of the covariates for the 9th grade regressions, the panel model results (Columns 1 and 3) show that location in a rural area and a higher socioeconomic status are both statistically significant associated with higher language and math test scores. For example, rural schools have test scores that are 0.125 of a standard deviation higher in language, and 0.134 of a standard deviation higher in math. The impact of a higher school socioeconomic status is larger for 9th grade language (Column 1): moving one "level" up

[48] Although the EN model has been partially implemented in some primary schools as well, the model is even less consolidated in secondary schools.

the socioeconomic status (e.g., from SES=1 to SES=2) is associated with 0.245 of a standard deviation higher test scores. School enrollment is only marginally significant for math test scores (Column 3) using both the panel (column 3) and school fixed effects model (column 4).

ii. Heterogeneous effects

To test if the EN models have heterogeneous impacts for schools with different characteristics, Table 3.6 reports the results of the school fixed effects model using different subsamples: a subsample with only rural schools, only urban schools, only the poorest schools (SES=1), only the schools with a middle socioeconomic status (SES=2; SES=3), and only the relatively richest schools (SES=4). Columns 1 and 2 show the results for the 5^{th} grade language and math samples respectively. Although the smaller samples reduce the power to detect effects, I find a statistically significant impact of EN for the rural schools in the 5^{th} grade language sample (although only marginally significant). In this case, among the rural schools that switched model, the cohorts exposed to EN had an average of 0.132 standard deviation higher test scores, than cohorts exposed to non-EN models (Column 1). The results also show that among schools with the lowest socioeconomic status (SES=1) that switched models, the cohorts exposed to EN have on average 0.14 of a standard deviation higher 5^{th} grade language test scores than cohorts exposed to other models (Column 1).

The last row in Table 3.6 shows the school fixed effect results for a subsample of only schools that are located in four departments that, according to the FEN, have traditionally well-established EN models: Antioquia, Caldas, Quindío, and Risaralda.

Although this a subjective assessment, and I am not able to directly test for the fidelity or quality of implementation of the EN model in these departments, running the schools fixed effects model for this subsample might provide some indication of the impact of EN under different levels of implementation. However, the results show that for this subsample, the impact of EN is not statistically significant in any of the samples.

Regardless of the subsample used, the effect on EN is non-existent for the language and math 9th grade samples. Although this could be due to the small sample of schools, and the even smaller sample of schools that switched in each of this subsamples, which reduces the power to detect effects. The subsample for schools with a middle socioeconomic status (SES=3) show a statistically significant coefficient for the 9th grade language sample (column 3). However, only 3 secondary schools switched models in this particular subsample, and that is driving that coefficient up.

iii. Propensity Score Matching

Propensity Score Matching relies on estimating the probability of implementing EN (the propensity scores) based on the available control variables for each school in the sample, and then estimating the impact of EN by comparing matched schools that either implemented or not the EN model (treatment and comparison groups), but which have similar probabilities of implementing EN. The advantage of using PSM is that it allows better modeling of the selection process of schools into EN, because the treatment and comparison groups are created based on the estimated probabilities of implementing EN. In other words, the probability model (and its estimated propensity scores) captures the

selection process of schools into EN based on observable characteristics that influence the decision to implement EN.

In order for the PSM to be successful, two important conditions must be met. First, the model must include variables that are able to predict the probability of implementing EN. Second, the matching procedure has to be able to balance the distribution of the relevant variables in both the control and treatment group (i.e., the differences between these groups must be statistically insignificant) (Caliendo & Kopeinig, 2008). In Table 3.7 and 3.8, I check if these conditions are successfully met.

Table 3.7 presents the probit model used to generate the propensity scores. The dependent variable is a dummy equal to 1 if the school implemented EN in 2005. The variables included in the regressions are the school characteristics that have been included as control variables in all previous models, but for the previous period (2002). That is, the characteristics of the school in 2002 are used to predict the probability of implementing EN in 2005. They show that most of the variables included in each probit regression are statistically significantly associated with the probability of implementing EN in 2005.

Table 3.8 presents the results of balancing tests for each variable. Each balancing test shows if there is a statistically significant difference in the relevant variables between the treatment and control groups generated through the matching process. The results presented in this table are based on the treatment and control groups created using nearest neighbor matching with 5 neighbors. I ran similar tests for each of the other matching algorithms used, but the treatment and control groups are so similar that the results do not change significantly. However, as Table 3.8 shows, after conditioning on the propensity

score there are still statistically significant differences in all variables between the treatment and control groups, which suggest that matching on the score was not completely successful. In the case of unsuccessful matches, Caliendo and Kopeinig (2008) suggest trying including higher order terms and interaction variables, or different covariates, until the treatment and controls groups appear to be balanced. Therefore, I ran the PSM models using different variables, and including higher order terms and interactions. I also trimmed the distribution of the propensity scores before matching (i.e., the lower and upper 5% of values of the variable were dropped). However, in all of these cases the treatment and comparison groups remain unbalanced, which may indicate a fundamental lack of comparability between the two groups (Caliendo & Kopeinig, 2008).

The results from PSM are reported in Table 3.9, although given that it has not been possible to balance the treatment and comparison groups, there is a threat to the internal validity of the results generated by the matching method. Each column contains the results for each of the grade and subject samples, while each row represents the estimates of the EN impact, and more specifically, the average treatment effect on the treated (ATT), using different propensity score matching methods. I use Stata's "psmatch2" command to generate these results. The standard errors estimated using propensity score matching do not take into account that the propensity score is estimated. Therefore, the standard errors are estimated using a bootstrap procedure with 100 replications.

The first three rows present the ATT when using nearest neighbor matching, using one, five, and ten nearest neighbors (NN). All the nearest neighbor matchings are done with replacement. The ATT estimated using NN is only significant for the 5th grade

language samples, when using 1 and 5 neighbors. In both cases, the ATT is one fifth of a standard deviation in student test scores, a larger impact than the one estimated using the panel and school fixed effects model. The impact of EN is also significant using NN matching with 1 neighbor in the 9th grade language sample: in that case, the ATT is 0.33, which is the largest impact estimated with any model in this study.

Some of the matches using nearest neighbors might be bad matches if the closest neighbor is far away. Therefore, rows 4-6 show the ATT when using 5 nearest neighbors and caliper matching, that is, imposing a maximum propensity score distance. When imposing even a small caliper (0.01) the sample size does not change, suggesting that the matches used with the nearest neighbor matching are generally good (Mueser et al., 2007). However, the only matches that provide statistically significant ATT in this case, are the ones using 0.01, 0.05, and 0.1 calipers for the 5th grade language sample. In this case, the ATT is 0.2 standard deviations in test scores. Given that the 5th grade language sample shows similar results using different types of matching, it seems like the choice of the type of matching might be unimportant in this case.

The PSM results support some the main findings from the panel model: in both cases, there is a positive impact of EN on 5th grade language test scores, although the impact is larger using PSM (0.2 v. 0.152 in the panel model). However, in contrast to the panel results, none of the PSM models for the 5th grade math sample show a statistically significant impact of EN. The 9th grade PSM results are slightly larger in magnitude than the other models, but again, these estimates are rarely significant.

3.7 Conclusions and policy implications

The Escuela Nueva model that originated in Colombia in the 1970s has been widely promoted as a successful model by a wide variety of actors (e.g., government, NGOs, development organizations). But despite its scale and apparent success, most of the evidence on its impact comes from descriptive studies, or from studies with weak identification strategies that do not allow isolating the impact of EN on test score from other possible confounding factors. This study is the only recent impact evaluation study of EN, and the only one to use several identification strategies that allow mitigating most of the endogeneity and selection issues.

The results of the school fixed effect strategy show that the within school impact of switching to EN is of 0.135 of a standard deviation in language test scores. In contrast to the 5^{th} grade results, the 9^{th} grade results show that implementing EN does not have a statistically significant impact on test scores. This suggests that the EN model might not be as successful in increasing student achievement in 9^{th} grade, as it is in 5^{th} grade. The main reason for this difference is that the EN model has been implemented for a shorter period of time in secondary grades. This model (also known as Post-primary), was created more than 20 years after the original EN model, which was meant for primary schooling, was created. Therefore, the model has only been partially implemented in secondary grades. The EN model has also been partially implemented in some primary schools, but the model is even less consolidated in secondary schools. It is also possible that students in 9^{th} grade have not been exposed to EN for long (e.g., only one or two years). It might be possible that more time is needed to consolidate the model before it is possible to see any significant impact on test scores.

The heterogeneity analysis shows that there is a statistically significant impact of EN for the rural schools in the 5th grade language sample. This result has two important implications: first, given that the EN was originally designed specifically with the aim to improve the provision of education in rural areas, one would expect that its larger impact would be concentrated on these areas. It seems that despite the effort to adapt the EN model to urban areas (also known as *Escuela Activa Urbana*), this model is not as effective as in rural areas, at least in terms of its effect on student test scores. Another way of looking at this is that the elements that compose the EN model are in fact more effective in a rural context. For example, the flexibility to adapt to different learning paces is probably more relevant for students in rural areas. Second, there might be particular characteristics in urban schools that might affect the implementation of EN. For example, the larger number of students that attend urban schools might make very difficult for students to benefit from the child-center curriculum, cooperative learning, and in general, the type of learning process promoted by EN.

The results also show that among schools with the lowest socioeconomic status (SES=1) that switched models, the cohorts exposed to EN have on average 0.14 of a standard deviation higher 5th grade language test scores than cohorts exposed to other models. This result is consistent with many other school interventions in developing countries, where it is common to find that the students, schools or families with lowest socioeconomic status tend to benefit more from this type of interventions.

Overall, the evidence presented in this study suggest that the *Escuela Nueva* model might be a useful intervention to improve the quality of education, particularly in

rural areas and in schools with low socioeconomic status, where the characteristics of this flexible education model might be more useful.

3.8 References

Abadie, A., & Imbens, G. (2006a). Large sample properties of matching estimators for average treatment effects. *Econometrica, 74* (1), 235-267.

Abadie, A., & Imbens, G. (2006b). On the failure of the bootstrap for matching estimators. Unpublished paper. John F. Kennedy School of Government. Harvard University. Mimeo.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22(1),* 31-72.

Forero, C., Escobar, D., & Molina, D. (2006). Escuelas Nueva's impact on the peaceful interaction of children in Colombia. In A. Little (Ed), *Education for all and multigrade teaching. Challenges and opportunities*. Dordretch, The Netherlands: Springer.

FEN (2012). Escuela Nueva Foundation website: http://www.escuelanueva.org/portal/en.html

FEN (2006). Memorias del Segundo Congreso Internacional de Escuelas Nuevas. Fundación Escuela Nueva. Medellín, Marzo de 2006.

Becker, S. O., & Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *The Stata Journal, 2*(4), 358–377.

Colbert, V. (1999). Mejorando el acceso y la calidad de la educación para el sector rural pobre. El caso de Escuela Nueva en Colombia. Available at: http://www.rieoei.org/rie20a04.htm

Fundación Escuela Nueva (FEN) (2014). Official website: http://www.escuelanueva.org/portal/

Heckman, J. Ichimura, H., & Todd, P., (1998). Matching as an Econometric Evaluation Estimator. *Review of Economic Studies, 65*(2), 261-294.

Heckman, J. & Robb, R. (1985). Alternative models for evaluating the impact of interventions. In J. Heckman and B. Singer (Eds.), *Longitudinal Analysis of Labor Market Data* (pp. 156–245). Cambridge: Cambridge University Press.

Lechner, M. (1999) Earnings and employment effects of continuous off-the-job training in East Germany after unification. *Journal of Business Economic Statistics 17*(1), 74–90.

McEwan, P. (1998). The effectiveness of multigrade schools in Colombia. *International Journal of Educational Development, 18*(6), 435-452.

McEwan, P. & Benveniste, L. (2001). The politics of rural school reform: Escuela Nueva in Colombia. *Journal of education policy, 16* (6), 547–559.

Ministerio de Educación Nacional (MEN) (2014). Official website: http://www.mineducacion.gov.co/1621/w3-channel.html.

Mueser, P., Troske, K., & Gorislavsky, A. (2007). Using state administrative data to measure program performance. *The review of economics and statistics, 89* (4), 761-783.

Murnane, R. & Willett, J. (2011). Methods matter. Improving causal inference in educational and social science research. Oxford University Press.

Programa de Promoción de la Reforma Educativa en América Latina y el Caribe (PREAL) (2003). Desarrollo de la educación en sectores rurales. Serie Mejores Prácticas. Santiago: PREAL.

Psacharopoulos, G., Rojas, C., & Velez, E. (1993). Achievement evaluation of Colombia's

Escuela Nueva: is multigrade the answer? *Comparative Education Review*, *37*(3), 263- 276.

Rojas, C., & Castillo, Z. (1988). Evaluación del Programa Escuela Nueva. IFT-133. Instituto SER de Investigación, Bogotá.

Rosenbaum, P. R., & Rubin, D.B., (1983). The central role of the propensity score in Observational Studies for causal effects. *Biometrika*, *70*(1), 41-55.

Schiefelbein, E. (1991). In search of the school of the XXI century: is the Colombian Escuela Nueva the right pathfinder? Santiago: UNESCO-UNICEF.

World Bank (2008). Quality of education in Colombia. An analysis and options for a policy agenda. Report No. 43906-CO.

Table 3.1. Descriptive statistics of school panel

	5th grade		9th grade	
	2002	**2005**	**2002**	**2005**
Total schools	14413	14848	2,968	4,092
Average standardized language test score	-0.04	-0.15	-0.13	-0.23
Standard deviation of test score	1.01	0.96	0.94	0.90
Average standardized math test score	-0.01	-0.10	-0.09	-0.12
Standard deviation of test score	1.02	1.00	0.99	0.94
% rural	81.2%	74.8%	43.56%	40.15%
Average SES	1.4	1.4	1.8	1.9
% of poorest schools (SES=1)	73.5%	72.9%	49.7%	46.3%
% of richest schools (SES=4)	2.2%	2.8%	7.4%	8.3%
Average school enrollment	195.3	256.7	609.4	769.1
Average 5th/9th grade enrollment	21.9	29.6	61.9	81.8
Average primary/secondary enrollment	128.9	152.4	321.6	386.9
Primary teacher/student	0.04	0.04	0.07	0.06
% primary teachers with professional degree	59.4%	68.3%	86.5%	91.6%
% primary teachers with pedagogic training	60.5%	67.8%	83.9%	86.9%
% with complete schedule	27.0%	34.5%	21.6%	24.4%
% Escuela Nueva schools	62.9%	56.1%	10.0%	9.5%
Schools that switched to Escuela Nueva	-	2.5%	-	0.7%
	-	(365)	-	(30)
Schools that switched from Escuela Nueva	-	4.0%	-	1.2%
	-	(593)	-	(48)

Notes: School socioeconomic status (SES) is an index that goes from 1 (poorest schools) to 4 (richest schools).

Table 3.2. Descriptive statistics of school panel by type of model. Sample: 5th grade.

	2002			2005		
	Escuela Nueva (SD)	**Non-EN (SD)**	**Diff. (t-stat)**	**Escuela Nueva (SD)**	**Non-EN (SD)**	**Diff. (t-stat)**
Average language standardized test score	-0.02	-0.06	**0.04**	-0.14	-0.18	**0.04**
	(1.08)	(0.90)	**(-2.41)**	(1.08)	(0.78)	**(-2.53)**
Average math standardized test score	0.07	-0.16	**0.23**	-0.01	-0.22	**0.21**
	(1.06)	(0.93)	**(-13.88)**	(1.14)	(0.77)	**(-13.68)**
% of rural schools	99%	51%	**0.48**	99%	44%	**0.56**
	(0.10)	(0.50)	**(-69.46)**	(0.09)	(0.50)	**(-89.65)**
Mean school SES	1.14	1.77	**-0.62**	1.10	1.80	**-0.70**
	(0.40)	(0.91)	**(47.60)**	(0.34)	(0.94)	**(57.07)**
Mean school Enrollment	59	426	**-368**	60	509	**-449**
	(68.5)	(475.7)	**(56.2)**	(81.53)	(538.09)	**(66.79)**
Mean teacher student ratio	0.04	0.04	**0.01**	0.05	0.04	**0.01**
	(0.02)	(0.03)	**(-15.14)**	(0.02)	(0.01)	**(-37.51)**
% of teachers w/professional education	55%	66%	**-11%**	63%	75%	**-13%**
	(0.45)	(0.33)	**(16.55)**	(0.44)	(0.29)	**(21.03)**
% of teachers w/pedagogic training	57%	67%	**-10%**	62%	75%	**-12%**
	(0.45)	(0.33)	**(15.12)**	(0.44)	(0.29)	**(20.07)**
% of schools that have a complete schedule	37%	11%	**26%**	50%	15%	**35%**
	(0.48)	(0.31)	**(-39.75)**	(0.50)	(0.35)	**(-50.17)**
Observations (No. of schools)	9062	5351		8331	6517	

Note: Standard deviations and t-statistic for the difference in parentheses. Statistiscally significant differences are bolded (5%). School socioeconomic status (SES) is an index that goes from 1 (poorest schools) to 4 (richest schools).

Table 3.3. Descriptive statistics of school panel by type of model. Sample: 9th grade.

	2002			2005		
	Escuela Nueva (SD)	**Non-EN (SD)**	**Diff. (t-stat)**	**Escuela Nueva (SD)**	**Non-EN (SD)**	**Diff. (t-stat)**
Average language standardized test score	-0.24	-0.11	**-0.13**	-0.05	-0.24	**0.19**
	(1.04)	(0.93)	**(2.08)**	(1.22)	(0.86)	**(-2.99)**
Average math standardized test score	-0.07	0.98	0.03	-0.01	-0.14	0.12
	(1.10)	(0.10)	(-0.39)	(1.40)	(0.88)	(-1.67)
% of rural schools	100.0%	37.3%	**62.7%**	100.0%	33.9%	**66%**
	-	(0.48)	**(-67.01)**	-	(0.47)	**(-84.98)**
Mean school SES	1.10	1.88	**-0.78**	1.05	1.98	**-0.92**
	(0.35)	(0.02)	**(28.40)**	(0.24)	(0.99)	**(45.49)**
Mean school Enrollment	170	658	**-488**	159	833	**-674**
	(122.54)	(534.73)	**(38.88)**	(87.80)	(660.71)	**(57.44)**
Mean teacher student ratio	0.08	0.07	**0.00**	0.06	0.06	**0.01**
	(0.04)	(0.03)	**(-1.96)**	(0.03)	(0.03)	**(-4.56)**
% of teachers w/professional education	77%	88%	**-11%**	85%	92%	**-7%**
	(0.31)	(0.18)	**(5.75)**	(0.26)	(0.16)	**(5.27)**
% of teachers w/pedagogic training	76%	85%	**-9%**	81%	88%	**-7%**
	(0.32)	(0.21)	**(4.91)**	(0.29)	(0.20)	**(4.64)**
% of schools that have a complete schedule	50%	18%	**32%**	58%	21%	**38%**
	(0.50)	(0.39)	**(-10.59)**	(0.49)	(0.41)	**(-14.45)**
Observations (No. of schools)	297	2671		387	3705	

Note: Standard deviations and t-statistic for the difference in parentheses. Statistiscally significant differences are bolded (5%). School socioeconomic status (SES) is an index that goes from 1 (poorest schools) to 4 (richest schools).

Table 3.4. Regression results. Dependent variable: 5th grade test scores.

	Language		Mathematics	
	Panel	School fixed effects	Panel	School fixed effects
	(1)	(2)	(3)	(4)
Escuela Nueva	0.152***	0.135*	0.206***	0.101
	[0.020]	[0.075]	[0.022]	[0.073]
Dummy for rural schools	0.006	--	0.117***	--
	[0.026]		[0.028]	
School socioeconomic status	0.094***	--	0.043***	--
	[0.016]		[0.015]	
School enrollment	-0.000	0.020	-0.010***	0.009
	[0.002]	[0.017]	[0.003]	[0.016]
Primary teacher-student ratio	1.023**	-0.384	0.651*	-0.706
	[0.414]	[0.825]	[0.342]	[0.804]
% teachers with professional education in primary	0.042	-0.070	0.066*	0.085
	[0.035]	[0.120]	[0.035]	[0.116]
% teachers with pedagogical training in primary	0.068**	-0.025	0.043	-0.078
	[0.034]	[0.114]	[0.035]	[0.111]
Dummy for complete schedule	0.034	-0.055	0.064**	-0.042
	[0.030]	[0.081]	[0.032]	[0.079]
Observations	29,261	29,261	29,261	29,261
R-squared (overall)	0.206	0.779	0.241	0.793
Year fixed effect	YES	YES	YES	YES
Municipal fixed effects	1041	NO	1041	NO
School fixed effects	NO	20949	NO	20949

Standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1

Notes: Escuela Nueva is a dummy variable equal to one when the school implements the Escuela Nueva model, and zero otherwise. School socioeconomic status is an index that goes from 1 (poorest schools) to 4 (richest schools). All standard errors are robust.

Table 3.5. Regression results. Dependent variable: 9th grade test scores.

	Language		Mathematics	
	Panel	School fixed effects	Panel	School fixed effects
	(1)	(2)	(3)	(4)
Escuela Nueva	0.039	0.161	0.092	0.242
	[0.067]	[0.275]	[0.067]	[0.310]
Dummy for rural schools	0.125***	--	0.134***	--
	[0.046]		[0.039]	
School socioeconomic status	0.245***	--	0.112***	--
	[0.023]		[0.022]	
School enrollment	-0.002	-0.023	-0.005*	-0.029*
	[0.003]	[0.015]	[0.003]	[0.016]
Secondary teacher-student ratio	-0.190	-1.762	0.310	-0.451
	[0.706]	[2.518]	[0.589]	[1.780]
% teachers with professional education in seconda	-0.058	-0.164	-0.034	-0.143
	[0.102]	[0.239]	[0.095]	[0.244]
% teachers with pedagogical training in secondary	0.025	-0.132	-0.037	0.071
	[0.087]	[0.207]	[0.083]	[0.217]
Dummy for complete schedule	0.141***	0.056	0.070	0.015
	[0.046]	[0.102]	[0.043]	[0.115]
Observations	7,060	7,060	7,060	7,060
R-squared (overall)	0.304	0.769	0.258	0.766
Year fixed effects	YES	YES	YES	YES
Municipal fixed effects	1006	NO	1006	NO
School fixed effects	NO	4651	NO	4651

Standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1

Notes: Escuela Nueva is a dummy variable equal to one when the school implement the Escuela Nueva model, and zero otherwise. School socioeconomic status is an index that goes from 1 (poorest schools) to 4 (richest schools). All standard errors are robust.

Table 3.6. Heterogeneous effects: School fixed effects model using alternate samples

	5th grade		9th grade	
	Language	Mathematics	Language	Mathematics
	(1)	(2)	(3)	(4)
Rural schools	0.132*	0.109	0.144	0.223
	[0.080]	[0.078]	[0.274]	[0.315]
Observations	*22,815*	*22,815*	*2,936*	*2,936*
Urban schools	0.065	-0.073	-	-
	[0.203]	[0.224]		
Observations	6,446	6,446		
Poorest schools (SES=1)	0.140*	0.109	0.210	0.253
	[0.083]	[0.081]	[0.297]	[0.345]
Observations	*21,421*	*21,421*	*3,368*	*3,368*
Middle-to-poor (SES=2)	0.055	0.079	-0.135	-0.060
	[0.188]	[0.192]	[0.797]	[0.443]
Observations	*4,931*	*4,931*	*1,918*	*1,918*
Middle-to-rich (SES=3)	0.228	-0.033	-0.968***	0.382
	[0.330]	[0.382]	[0.295]	[0.556]
Observations	*2,186*	*2,186*	*1,215*	*1,215*
Richest schools (SES=4)	-0.223	-0.328	-	-
	[0.728]	[0.953]		
Observations	*723*	*723*		
Well-established	0.068	0.123	0.609	-0.082
implementation	[0.204]	[0.175]	[0.480]	[0.508]
	6,451	*6,451*	*1,665*	*559*

Standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1.

Notes: Escuela Nueva is a dummy variable equal to one when the school implement the Escuela Nueva model, and zero otherwise. School socioeconomic status (SES) is an index that goes from 1 (poorest schools) to 4 (richest schools). Only the coefficients for EN are reported. All regressions control for school enrollment, teacher student ratio, percentage of teachers with a professional degree, percentage of teachers with pedagogic training, and include school and year fixed effects. All standard errors are robust.

Table 3.7. PSM Probit regressions. Dependent variable: Escuela Nueva in 2005

	5th grade		9th grade	
	Language	Mathematics	Language	Mathematics
	(1)	(2)	(3)	(4)
Test score in previous period	0.012	0.044***	0.005	-0.007
	[0.019]	[0.019]	[0.053]	[0.048]
Dummy for rural schools	1.026***	1.028***	-	-
	[0.101]	[0.101]		
School socioeconomic status	-0.185***	-0.179***	-0.090	-0.090
	[0.048]	[0.048]	[0.161]	[0.161]
School enrollment	-0.605***	-0.599***	-0.705***	-0.705***
	[0.019]	[0.019]	[0.064]	[0.065]
Mean teacher student ratio	-2.184*	-2.349*	-8.834***	-8.829***
	[1.295]	[1.296]	[1.542]	[1.542]
% of teachers w/professional education	-0.302***	-0.310***	-0.370	-0.369
	[0.112]	[0.112]	[0.295]	[0.295]
% of teachers w/pedagogic training	0.339***	0.345***	0.226	0.228
	[0.111]	[0.112]	[0.286]	[0.287]
% of schools that have a complete schedule	0.710***	0.717***	0.453***	0.452***
	[0.042]	[0.043]	[0.109]	[0.109]
Observations	8312	8312	1007	1007
R-squared	0.510	0.510	0.280	0.280

Standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1

Notes: All regressions include municipal fixed effects. Escuela Nueva is a dummy variable equal to one when the school implements the Escuela Nueva model, and zero otherwise. School socioeconomic status is an index from the SABER data that goes from 1 (poorest schools) to 4 (richest schools).

Table 3.8. Balancing tests for PSM treatment and control groups.

	5th grade			9th grade		
	Control group	**Treatment group**	**Diff. (t-stat)**	**Control group**	**Treatment group**	**Diff. (t-stat)**
Average language standardized test score	-0.14	-0.09	**-0.05**	-0.24	-0.18	-0.06
	(0.85)	(1.06)	**(-2.32)**	(1.00)	(1.03)	(-0.80)
Average math standardized test score	-0.25	0.02	**-0.27**	-0.12	-0.02	-0.10
	(0.88)	(1.05)	**(-12.54)**	(1.03)	(1.18)	(-1.11)
% of rural schools	46.2%	99.4%	**-53.2%**	-	-	
	(0.50)	(0.08)	**(-64.06)**			
Mean school SES	1.81	1.09	**0.72**	1.24	1.07	**0.17**
	(0.93)	(0.32)	**(44.85)**	(0.53)	(0.28)	**(6.22)**
Mean school Enrollment	493.3	63.0	**430.4**	355.4	153.1	**202.3**
	(495.8)	(73.40)	**(52.21)**	(287.79)	(84.90)	**(17.21)**
Mean teacher student ratio	0.03	0.04	**-0.01**	0.08	0.08	0.00
	(0.03)	(0.02)	**(-10.71)**	(0.03)	(0.04)	(-0.22)
% of teachers w/professional education	66.6%	52.4%	**14.1%**	82.4%	75.3%	**7.1%**
	(0.32)	(0.45)	**(16.76)**	(0.23)	(0.31)	**(3.10)**
% of teachers w/pedagogic training	66.6%	54.2%	**12.3%**	79.9%	75.1%	**4.7%**
	(0.32)	(0.45)	**(14.56)**	(0.24)	(0.31)	**(2.04)**
% of schools that have a complete schedule	12.0%	48.1%	**-36.2%**	24.4%	55.3%	**-30.9%**
	(0.32)	(0.50)	**(-39.68)**	(0.43)	(0.50)	**(-8.18)**
Observations (No. of schools)	3721	4581		799	208	

Note: Standard deviations and t-statistic for the difference in parentheses. Statistiscally significant differences are bolded. The treatment and control groups correspond to the ones created by using the Nearest Neighbor matching with 5 neighbors. All variables are from 2002, and were used to create the treatment and comparison groups.

Table 3.9. Estimates of program effect (ATT) on standardized test scores using propensity score matching

	5th grade		9th grade	
	Language	Mathematics	Language	Mathematics
	(1)	(2)	(3)	(4)
Nearest neighbor	**0.20**	0.05	**0.33**	0.14
(NN==1)	**(0.09)**	(0.10)	**(0.16)**	(0.21)
	[N=8,263]	[N=8,241]	[N=992]	[N=992]
Nearest neighbor	**0.20**	-0.10	0.20	0.07
(NN==5)	**(0.10)**	(0.09)	(0.16)	(0.15)
	[N=8,263]	[N=8,241]	[N=992]	[N=992]
Nearest neighbor	0.14	-0.13	0.13	0.02
(NN==10)	(0.10)	(0.07)	(0.14)	(0.17)
	[N=8,263]	[N=8,241]	[N=992]	[N=992]
Caliper (0.01)	**0.20**	-0.08	0.20	0.13
(NN=5)	**(0.09)**	(0.09)	(0.15)	(0.17)
	[N=8,263]	[N=8,241]	[N=992]	[N=992]
Caliper (0.05)	**0.20**	-0.10	0.23	0.13
(NN=5)	**(0.10)**	(0.10)	(0.16)	(0.15)
	[N=8,263]	[N=8,241]	[N=992]	[N=992]
Caliper (0.10)	**0.20**	-0.10	0.20	0.08
(NN=5)	**(0.10)**	(0.09)	(0.15)	(0.16)
	[N=8,263]	[N=8,241]	[N=992]	[N=992]
Kernel (bw=0.01)	0.13	-0.09	0.21	0.11
	(0.06)	(0.06)	(0.14)	(0.15)
	[N=8,263]	[N=8,241]	[N=992]	[N=992]

Note: Standard errors in parentheses. Standard errors are calculated by bootstrap methods based on 100 replications. Statistically significant differences are bolded. Number of observations in the common support in brackets.