THE COLOMBIAN ESCUELA NUEVA SCHOOL MODEL:
LINKING PROGRAM IMPLEMENTATION AND LEARNING OUTCOMES

AN ABSTRACT

SUBMITTED ON THE SIXTH DAY OF APRIL 2017

TO THE PAYSON PROGRAM IN GLOBAL DEVELOPMENT

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

OF THE SCHOOL OF LAW

OF TULANE UNIVERSITY

FOR THE DEGREE

OF

DOCTOR OF PHILOSOPHY

BY:

Katharina Hammler

APPROVED:

William E Bertrand, Ph.D. (Chair)

John Edwards, Ph.D.

Julie Hernandez, Ph.D.

Paul Hutchinson, Ph.D.

Eamon Kelly, Ph.D.

Abstract

This dissertation uses a mixed methods design to analyze how the Colombian student-centered school model Escuela Nueva affects learning outcomes, and how well the model is implemented. Primary data from 78 schools in the department Quindío show large variation in implementation across schools, both overall and with regard to the model elements. On average, schools implement only around 62% of the elements. While schools that are officially classified as Escuela Nueva tend to implement more elements than conventional schools, the difference is not large, and considerable variation exists within each group. Qualitative data confirms these heterogeneities, and suggests that differences across schools are even larger than captured by the quantitative data, given the different ways in which the program is being used or adapted in practice.

Learning outcomes are measured as scores on the national standardized test Pruebas SABER. Multilevel modeling techniques are used to analyze the scores from over 810,000 students in 21,235 schools across Colombia. The results show that students in schools that are officially classified as Escuela Nueva score significantly better, the difference amounting to 10.5 to 23.2 points (0.14 to 0.30 standard deviations). This effect is comparable to the effect of the difference of one socioeconomic level. Furthermore, Escuela Nueva tends to decrease the achievement gaps between socioeconomic levels and genders. The analysis also reveals large differences in the effect of the school model across municipalities and departments.

For the department Quindío, the effect of the school model is analyzed using an implementation index instead of the official classifier. Data is available for 1,068 students in 76 schools, representing half of the department's rural primary schools. Multilevel estimation generally shows no effect of program implementation, but cannot take into account the large relative sample size. Survey estimation techniques reveal a large effect of Escuela Nueva implementation for grade 3 mathematics and for civic competencies, where the difference in the expected score between a school with a low and one with a high implementation index is 140 to 220 points. The department-level analysis also confirms that the Escuela Nueva model helps to close gaps between socioeconomic levels.

THE COLOMBIAN ESCUELA NUEVA SCHOOL MODEL:
LINKING PROGRAM IMPLEMENTATION AND LEARNING OUTCOMES

A DISSERTATION

SUBMITTED ON THE SIXTH DAY OF APRIL 2017

TO THE PAYSON PROGRAM IN GLOBAL DEVELOPMENT

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

OF THE SCHOOL OF LAW

OF TULANE UNIVERSITY

FOR THE DEGREE

OF

DOCTOR OF PHILOSOPHY

BY:

Katharina Hammler

APPROVED:

William E Bertrand, Ph.D. (Chair)

John Edwards, Ph.D.

Julie Hernandez, Ph.D.

Paul Hutchinson, Ph.D.

Eamon Kelly, Ph.D.

Acknowledgements

I am deeply thankful to the many people without whom this dissertation would not have been possible. First and foremost, I would like to thank Dr. Bill Bertrand for his support throughout my time at Tulane. In German, a PhD supervisor is called *Doktorvater* (“doctoral father”), which does more justice than the English term to Dr. Bertrand’s role: All the opportunities to learn and grow and the guidance and encouragement that he has given me have shaped not only my academic interests, but also the person I am today. Thank you for that.

I also want to thank the rest of my doctoral committee for their help and guidance: Dr. John Edwards, Dr. Julie Hernandez, Dr. Paul Hutchinson, and Dr. Eamon Kelly: thank you for working with me through this project, and for your support and dedication.

There are many people outside of Tulane who contributed to this research project. I am sincerely grateful to Fabio Gomez at the Universidad del Quindío for his enthusiasm and invaluable help in organizing the fieldwork in Colombia. Without the dedicated work of my team of field workers—Sebastián Idarraga, Cindy Otálora, Andrea Contreras, and Andrea Marin—this research would not have been possible. Likewise, the support of Leonor Botero Arboleda from the Universidad de La Sabana was crucial for obtaining all required IRB permits. Vicky Colbert and Clarita Arboleda and their team at Fundación Escuela Nueva supported the development of the project and provided valuable input and feedback. Laura Ospina and Juan Diego Álvarez provided much more than a place to sleep on my trips to Colombia. I am also grateful to Martin Burt, Sara Hooper, and Kate Melman for connecting me with researchers from Colombia, editing and certifying Spanish documents, checking the accuracy of data entry sheets, and much more. I am thankful to all these people for their role in making this project possible.

Finally, I want to thank Dan for his moral support throughout the process, for discussing ideas and challenges, for editing drafts, and for looking out for my well-being in the most stressful times. I could not have done this without you.

Overview of Contents

Table of Contents

1 Introduction

In November of 2016, just weeks after Colombians voted “no” on the plebiscite about a peace deal with the guerilla organization FARC, hundreds of educators, researchers, policy makers, and other stakeholders gathered in Bogotá to discuss the present and future of a participatory school model that originated in Colombia’s coffee growing region four decades ago, Escuela Nueva. The panelists and audience members of the third Congreso Internacional de Escuelas Nuevas seemed to agree that providing access to quality education to all children remains a key challenge, and that more than ever schools also play an important role in promoting peace and mutual understanding and in building citizenship. There was a lot of optimism that the Escuela Nueva model might provide the key to achieving both of these.

Another message stuck out from the panels: Despite Escuela Nueva’s large reach and fame, the formal evidence base is largely outdated and generally scarce, and many of the perceived benefits of the program are based on feelings, rather than on empirical data. One author of one of the few formal studies on the model said on a panel that we may not have a lot of hard data on what works and why, but when one visits an Escuela Nueva school, one simply *feels* that the atmosphere is different, the students are different, and the teachers are different. As important as these feelings are to inspire teachers and change makers, they cannot replace the insights that can be gained from thorough studies of the model. In order to improve the model’s implementation and to make its alleged benefits accessible to many more children around the world, positive feelings need to be accompanied by statistical evidence that the model is working.

This dissertation contributes to the body of empirical evidence about Escuela Nueva's effect on learning outcomes. As it turns out, the first challenge arises when trying to define what constitutes an "Escuela Nueva"—maybe not in theory, but in practice, as the assumption that the program is properly implemented in all schools that auto-identify as an Escuela Nueva proves to be problematic. In order to really understand the effect of the school model, one needs to first know what is actually happening in the classrooms before one can try to correlate these classroom practices with learning outcomes.

The relevance of this project extends beyond Colombia, as the provision of quality education remains a central concern in the field of Development Studies across the world. The Escuela Nueva model has already expanded to many other countries and regions (mostly within Latin America, but also to Africa and Asia), yet with mixed success. Increasing the knowledge base on the model not only serves academic interests, but can help inform continuous education policy reforms both in Colombia and in other parts of the world.

The remainder of this dissertation is structured as follows. The rest of this introduction gives an overview on learning outcomes in Colombian primary schools, and presents the Colombian Escuela Nueva school model in detail. After discussing theoretical support for the school model, the introduction closes with formulating the problem that this research aims to address: How can learning outcomes in Colombia remain so poor, if an allegedly successful model has been widely used for decades?

Chapter 2 contains a review of the existing evidence of Escuela Nueva implementation and program outcomes. It suggests, on the one hand, that program implementation is irregular and incomplete, and, on the other hand, that the school model helps to improve learning outcomes as well as civic competencies. However, it also reveals methodological shortcomings of previous evaluation studies.

Chapter 3 describes the methods that the dissertation applies. It first presents research hypotheses and research questions, and then introduces a mixed methods approach to test and answer them. The chapter includes an extensive discussion of multilevel modeling techniques, which are argued to be the appropriate framework to analyze the effect of the Escuela Nueva model on learning outcomes. The chapter ends with a presentation of the database used for this study.

Chapter 4 contains the country-level analysis of learning outcomes, which tests whether students in schools that are officially classified as Escuela Nueva schools can expect to score higher on the standardized test *Pruebas SABER*. The results based on multilevel modeling techniques suggest that this is the case, and that the model additionally helps to close achievement gaps between socioeconomic levels and genders.

The data on program implementation in Quindío, a department in Colombia's coffee growing region, is analyzed in chapter 5. After describing the implementation index that was constructed from primary data collected in 78 rural primary schools, the chapter compares implementation scores between the schools. This is done based on the overall index and on its individual dimensions. The data suggests that the model is not being implemented very faithfully, and that there is considerable heterogeneity both across and within schools. Furthermore, the data indicates that the official Escuela Nueva classifier is not a precise way of identifying Escuela Nueva schools. Qualitative evidence confirms these quantitative results and provides some more insights into the way that the model elements are being used, or not used, in the department.

Using the index developed in chapter 5 to measure Escuela Nueva implementation, chapter 6 analyzes Pruebas SABER test score in Quindío. The results provide some evidence for the hypothesis that Escuela Nueva improves learning outcomes, but the results are not as clear as in the country-level study. Where there are statistically significant effects, they are large; but that is only the case for grade 3 mathematics and for civic competencies. The discussion of the results suggests that this may be due to the low statistical power of the analysis, rather than to the lack of an effect in the study population.

Finally, chapter 7 summarizes the findings and discusses limitations of the research as well as areas of future research. The chapter closes with a list of policy recommendations.

1.1 Learning outcomes in Colombia

Education has been a top priority on the International Development agenda for decades, and rightly so: the social and economic benefits of education are obvious and manifold, ranging from increased levels of productivity and economic opportunities to better health, lower fertility, higher self-esteem, improved civil engagement and political participation, and far beyond. In the past decade, the focus of research and policy interest has shifted from questions of mere access to issues of quality and equity. Bringing children from deprived backgrounds to school and keeping them there is still a challenge, but it is equally important that they acquire meaningful skills and knowledge while they are in class. Evidence is accumulating that the link between school attendance and learning outcomes is often not very strong, as will be shown in the following pages.

We all know from our own experience that being in school does not necessarily equate to learning, and in resource-deprived contexts this problem is often magnified. Can we really expect a chronically malnourished daughter of illiterate parents who is sharing the classroom with 50 other children and listens to a poorly trained and often-absent teacher lecturing in a language she barely understands to somehow be able to read, write, and do basic math upon graduation? Yet, this has been the working assumption in international development for a long time: It has been common practice to define and measure educational program and policy goals in terms of inputs or, at best, direct outputs, and not in terms of learning outcomes. Common program goals have been to increase the enrollment rate, decrease the student-teacher ratio, expand spending on education, increase the graduation rate, and so on.[1]

For a large part, a focus on inputs and outputs is the result of data constraints: Counting the number of students in a classroom is a lot easier than determining what they learned. However, the growing number of studies that try and focus on educational outcomes (instead of on inputs or outputs) show that the difference between these approaches cannot simply be dismissed as "noise" in the data. It turns out that the discrepancies are substantial.

Two of the first authors to point out these discrepancies were Nancy Birdsall and Jere Behrman, in a paper published as far back as 1983 (Behrman and Birdsall 1983). The authors show theoretically and empirically for the case of Brazil that estimates of returns to education are seriously biased if education is only measured by years of schooling (which is only a measure of attendance and not of learning outcomes). Using a very simple measure of education quality—namely, the teachers' education level—the authors formally incorporate schooling quality in the standard Mincerian framework.[2] Accounting for education quality cuts the estimated effect of

[1] For instance, the second Millennium Development Goal was to get all children enrolled in school; it did not say anything about desired learning outcomes

[2] Jacob Mincer (1958) and Gery S. Becker (1965), the founders of neoclassical human capital theory, set up a formal microeconomic model connecting investment in education to the level of income that individuals can expect over their lifetimes. The basic idea is that personal income is a function of human capital that can be accumulated through education; and individuals (or families) can chose to forgo current

years of schooling on expected income in half and explains a significant part of the variation in returns to schooling over space and among individuals. Recent data shows that Colombia is a case in point.

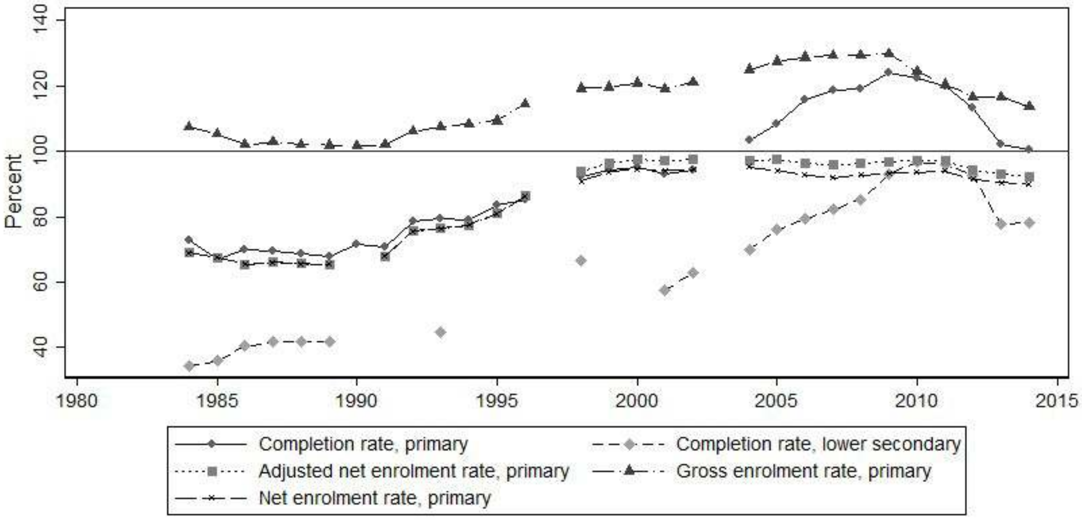


Figure 1 Education Participation in Colombia, 1984-2014. Data source: World Bank (2016)

Figure 1 shows participation and completion statistics for primary and lower secondary education for Colombia from 1984 to 2014. In 2014, the most recent year for which data is available, the country's primary gross enrollment rate was 113.7%, the adjusted primary net enrollment rate

consumption in order to invest in their (children's) education so that their future income increases. This simple idea had a considerable impact on the field of economics and related areas and triggered a wide range of theoretical and empirical work trying to refine and the model and to prove its validity and usefulness for the real world.

was 92.3%, the primary completion rate was 100.6%, and the lower secondary completion rate was 78.2% (World Bank 2016). Hence, the numbers seem to show that most of Colombia's children are in school and can expect to complete basic education.

However, a closer look at the apparently good numbers points towards some potential quality and equity issues. A first indication is the large difference between gross enrollment and adjusted net enrollment. The gross enrollment rate captures total enrollment in primary schools, regardless of age, expressed as a percentage of the population of official primary education age. A rate of 113.7% thus indicates that many of the children enrolled in primary school are above primary school age. The adjusted net enrollment rate, which looks only at pupils of the official primary school age group, is with 92.3% in 2014 significantly lower. This reflects a high over-age enrollment ratio of 21% in 2014, indicating that one out of every five students in Colombia's primary schools is over the official primary school age – a result of both late entrance and frequent grade repetition. Over-age primary school attendance is a particularly common problem in rural areas, where the ratio for the last available year, 2010, was as high as 29%. It is also a common problem among the poorest quintile, where the ratio reached 33%. Among male students from the poorest quintile living in rural areas, the over-age primary school attendance rate was a staggering 39% (UNESCO EFA 2013).

Second, and related, in 2013 the survival rate to the last grade of primary was only 83.5% (World Bank 2016). While the same indicator is not available on a disaggregated level, UNESCO EFA (2013) reports that the 2010 primary completion rate among the age group of 15-24 years was only 82% in rural areas, 78% among the poorest quintile, and 76% among rural males from the poorest quintile. This indicates that addressing the issue of primary school drop-out continues to be crucial when it comes to improving educational outcomes for disadvantaged students.

If indicators of educational quantity show some shortcomings, indicators of educational quality reveal true deficiencies. Results from international school achievement tests suggest that Colombia's schools do a bad job in teaching children skills and knowledge. For instance, 61% of Colombian 8th graders (i.e., children who actually are in school!) showed below low proficiency levels on the 2007 round of TIMSS (Trends in International Mathematics and Science Study). In other lower income countries, only 45.4% of students performed as poorly. Reversely, only 2% of Colombian 8th graders showed high levels of proficiency, and 0% showed an advanced level, compared to 6% and 1.1%, respectively, of 8th graders from other lower income countries. Only 11% of Colombian students performed at least at an intermediate level, compared to 25.5% in other lower income countries (World Bank 2010, xiv). Results from international reading performance tests are no less alarming. For instance, in the 2011 round of PIRLS (Progress in International Reading Literacy Study), 28% of Colombian 4th graders scored below the low international benchmark (international median: 5%), while only 10% reached the high international benchmark (international median: 44%) (Mullis et al. 2012, 68).

What does this tell us? Even though Colombia's participation statistics show a little room for improvement, they are not abysmal; most children attend school and can expect to graduate from primary school, and 8 out of 10 children even graduate from lower secondary school. However, only 1 out of 10 secondary students obtains the expected mathematics skills while in school, and only 7 out of 10 primary school students obtain basic reading skills. And, for what it is worth, this does not even tell us anything about problem-solving or life skills, but is simply about the capacity to read in the language of instruction.

It is revealing to look at some of these numbers on a more disaggregated level. Figure 2 depicts the results of the 2007 TIMSS round for primary education-aged children (4th graders). Children from the poorest quintile were far less likely to have learned basics in mathematics than their richer peers (27% in the poorest quintile versus 50% in the richest one; the country average is 38%). In both income groups—but especially so in the poorer one—fewer girls than boys had learned the basics, and children from rural areas typically performed worse than children from urban areas (except for poor males). Only 16% of poor girls from rural areas were able to solve basic mathematics problems after attending school for four years, while 59% of well-off boys (from both urban and rural areas) were able to do so.

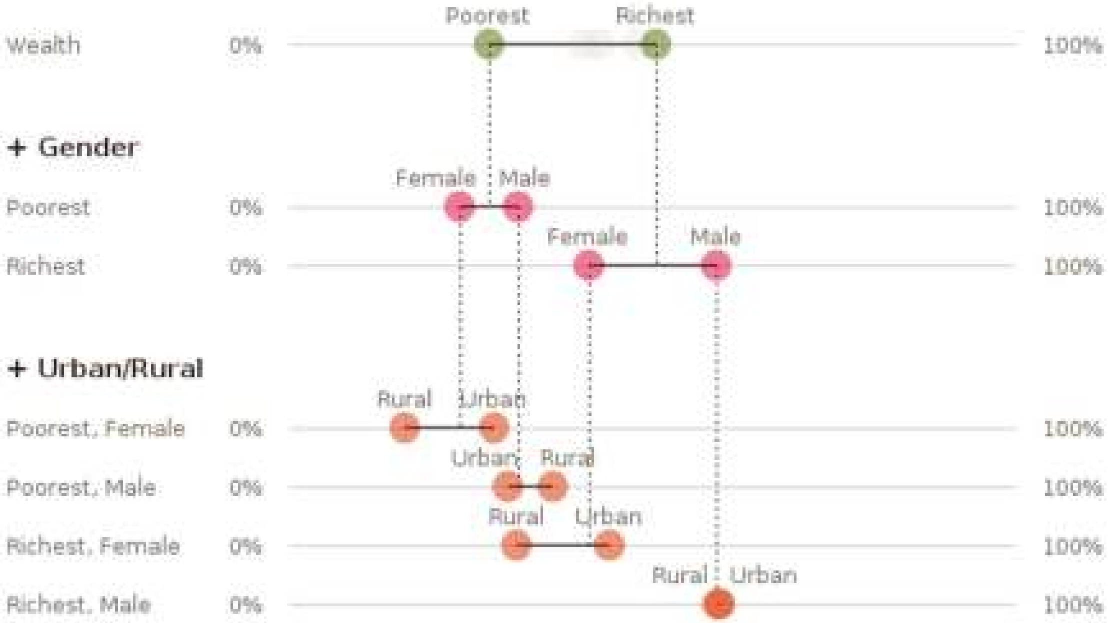


Figure 2 Percentage of primary students (4th graders) who learned basics in mathematics: Differences by wealth quintile, gender, and area (Source: UNESCO EFA 2013).

Hence, after successfully providing access to basic education for the vast majority of its children, Colombia faces another important challenge: How can the country make sure that *all* children benefit from their time in school? How can it offer them an education that is meaningful and successful in creating skills and competencies? And how can it make sure that the school addresses the specific needs of children from very different backgrounds, so that wealth, gender, or area of residence do not continue to predetermine a child's learning outcomes? For its primary schools, Colombia seems to have found an answer in an innovative school model, Escuela Nueva (*"New School"*). This school model will be introduced on the following pages.

1.2 The Escuela Nueva model

Based on UNESCO's multi-grade school reform strategy of the 1960s and 70s, Escuela Nueva (EN) started as a grassroots movement led by a group of rural teachers and educators (among them Vicky Colbert[3] who, up to this date, is spearheading the expansion of the model as Executive Director of the NGO Fundación Escuela Nueva). This group had the vision to create a school that would respond to the specific needs of children in an isolated poor area. Since then, EN has become national policy: EN schools are public schools that are financed by the government, but they use materials and approaches that often differ from "conventional" schools. As Torres (1992) puts it, this is perhaps the model's greatest merit and most promising aspect: EN is *not* an alternative to formal or state education, but an alternative *within* the formal and public education system. This clearly distinguishes EN from many other educational programs that target

[3] To this day, Vicky Colbert has played a central role in EN's development and implementation. After earning a Master's degree in Education from Stanford University, she returned to her home country Colombia to work in some of the poorest, most isolated rural schools. Since then, she developed, expanded, and sustained this social innovation in different roles: as Escuela Nueva National Coordinator, as Vice-Minister of Education of Colombia, as UNICEF Regional Education Adviser for the LAC Region, and as Director of Fundación Escuela Nueva (FEN), a Colombian NGO that she founded in 1987 to ensure the quality and sustainability of the model. Vicky Colbert has received numerous prestigious awards for her work, among them the Camilo Torres Medal in Education, awarded by the Colombian Ministry of Education in 2001; the Skoll Award for Social Entrepreneurship in 2006; the Clinton Global Citizenship Award in 2007; and the Wise Prize for Education in 2013, just to name a few.

disadvantaged children around the world. Today there are around 20,000 EN schools across Colombia, most of them in rural areas. Officially, about half of the country's primary schools have adopted the EN model (ICFES 2010).

Interestingly, there are no general, established rules on the adaption of the model. Schools and local authorities are free to choose from a range of educational models, the EN approach being one of them.[4] Some departments, such as Boyacá, actively promote the model and spend significant resources on the development of materials and on teacher training. Other departments, such as Quindío, have officially turned all of their rural schools into EN schools, yet there are few resources dedicated to the program's implementation. Still other departments seem to have no official policy strategy, leaving the decision entirely up to the individual municipalities or schools.

This variation in the official adaption of the model is at least partly due to the large level of autonomy that Colombian departments, municipalities, and schools have in the field of education. The Constitución de 1991, the Ley General de Educación de 1994, and the Ley 715 de 2001 transferred many responsibilities for the provision, funding, and supervision of primary and secondary schools to the departments and municipalities, and gave considerable autonomy to the individual schools with regards to curriculum and evaluation (OECD 2016, 42). Thus, the central government can no longer mandate the use of specific school models. Furthermore, the autonomy of municipalities and schools to promote or choose a specific school model is

[4] According to the Colombian Ministry of National Education (Ministerio de Educación Nacional n.d.), a school model is a defined set of strategies, pedagogical, and didactical principles in order to provide relevant and high-quality education to a target population with specific characteristics. In doing so, a model uses a defined set of educational materials and training processes. The ministry provides a catalogue of different schools models that the individual schools may or may not use. Note that for the purpose of this research project, the terms "school model" and "educational model" are used synonymously.

somewhat constrained by the departmental Secretaries of Education's budgetary decisions. In short, the choice of an educational model is a complex process, and the distribution of EN schools across Colombia is not standardized.

The EN model tries to address the specific needs of disadvantaged children in rural areas and, more recently, of children in poor urban areas. According to Vicky Colbert (2009), EN's core assumption is that the overall quality of education can only improve if creative changes throughout the school model are made. Furthermore, systematic change can be possible only if the model includes plans to go to scale. In accordance with these two core ideas—creative change throughout the system and replicability—a number of key characteristics emerged (Colbert 2009; McEwan 2008; Forero-Pineda, Escobar-Rodriguéz, and Molina 2006). They are systemized and visualized as a conceptual framework in Figure 3 and are summarized below.

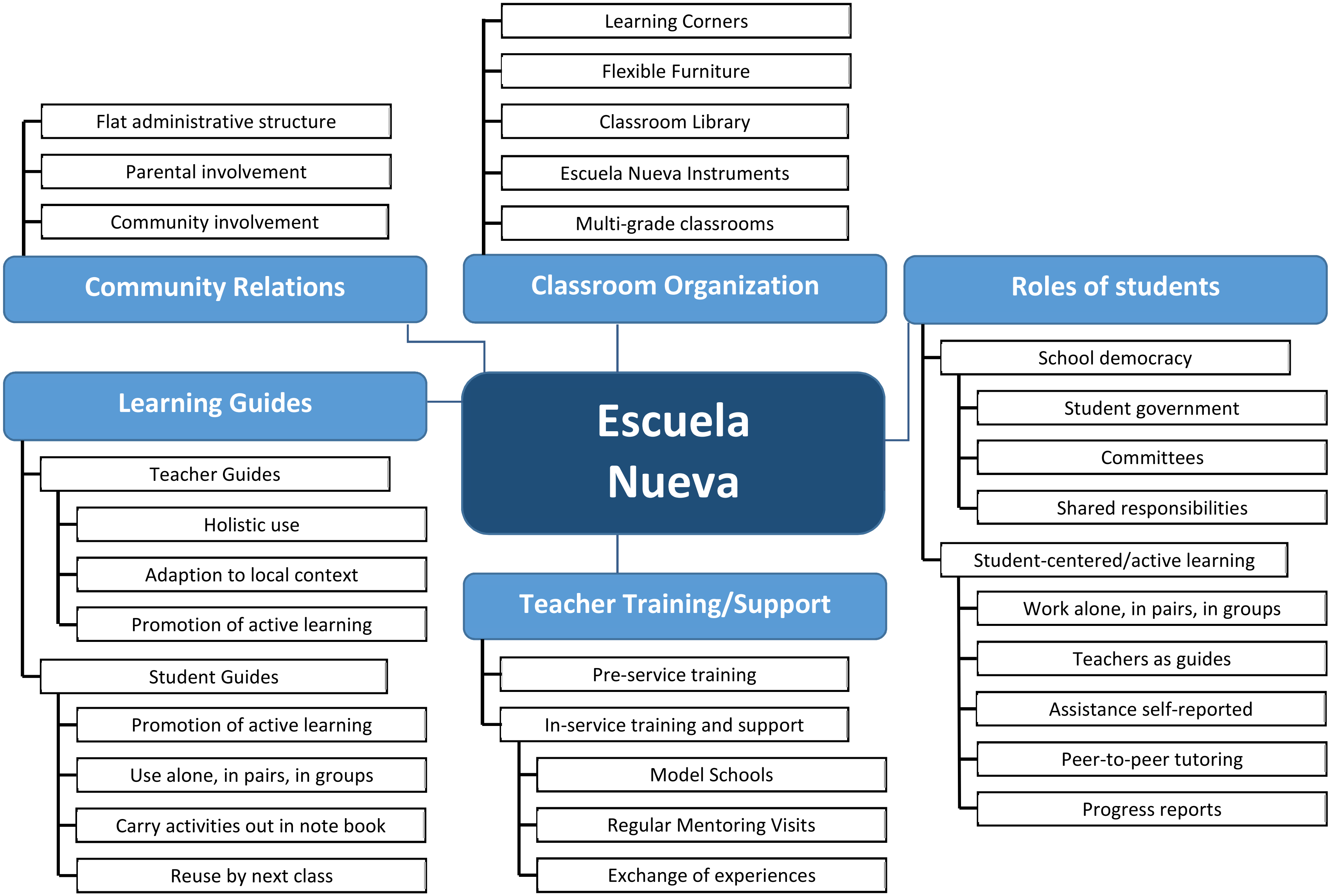


Figure 3: Conceptual Framework of Escuela Nueva Model. Source: Author's depiction

It may be telling that the conceptual model presented in Figure 3 is a depiction of the author. While descriptions of the model's key characteristics can be found in different publications on the EN model, to the best of the author's knowledge there is no official conceptual model or coherent catalogue of model elements describing the characteristics of each element together with its role and purpose in the model and its relationship to other elements. At least in part, this lack of detailed standardization is due to the intent to keep the model flexible and open for local adaptions – the more detailed and prescriptive the guidelines are, the less adaptable the model becomes. However, less detailed and less prescriptive guidelines can also cause problems when scaling up, requiring better training and preparation of the individual teachers. This tension between replicability and adaptability has not been fully resolved. That being said, the key dimensions of the model are clearly spelled out and described in several publications. What follows is a description of these key dimensions and some of their elements.

The first key characteristic is a new **role of students**, who play the leading part in their own education through active, participatory, and reflective learning. "Soft skills" such as cooperation, democratic attitudes, and an improved self-concept are explicit aims of the model. The focus lies on teaching children the ability to acquire and apply knowledge and skills. In that sense, knowledge is not "imposed" onto them, but they are guided to construct it. All of this is achieved by shifting from a focus on lectures to a student-centered school where students work alone, in pairs, and in groups. The role of the teacher shifts to that of a guide or mentor, and students are encouraged to help each other through peer-to-peer tutoring. Students are also encouraged to take responsibility for their own learning by self-reporting their assistance and by choosing when to be tested on their learning progress (flexible progress reports). The model uses flexible promotion, meaning that there is no grade repetition. Instead, a student advances to the next grade level whenever all grade requirements of a given content area are fulfilled. This allows children to advance at their own pace or to leave school temporarily, which is particularly important for poor families: It is very common for children from such families to be required to help with household or farming tasks for short or extended periods of time to contribute to the survival of the family. In conventional schools, this typically leads to high drop-out rates as it becomes hard or impossible to catch up with the class once the child returns to class.

The active learning approach extends beyond academic goals: School democracy is a guiding principle, with the aim of having students learn civic behavior through everyday experience. At the beginning of every year, students elect their student governments and organize themselves in committees (such as a red cross committee, a cleaning committee, a conflict mediation committee, and others). In these roles, students take an active role in everyday classroom decisions and share the responsibilities for the organization of their schools. By integrating democracy and shared responsibilities into everyday school life, EN aims to increase students' civic competencies and promote a peaceful, democratic society.

The **classroom organization** in EN schools is different from conventional schools in order to facilitate the different learning style. EN schools have multi-grade classrooms, which is not seen as a disadvantage but as an integral part of the model, because it facilitates topic-centered learning and learning from peers. Classrooms are equipped with special flexible furniture that can be moved around to accommodate work in pairs or groups. For each subject matter, there are learning corners that are stocked with didactic materials from the community. Being able to use such familiar materials helps students to better relate with the curriculum contents. There are classroom libraries that contain reference books and literature, and students are encouraged to use the library regularly to learn more about a topic. Last but not least, each classroom contains a series of EN instruments, such as a board of values, a box for suggestions, and a letter box for each student to send and receive friendship mail.

A key element of the EN model are the self-instructional l**earning guides** that serve as pivotal workbooks. The learning guides integrate the core national curriculum with regional or local topics and each child's personal experience. There are guides for teachers and guides for students. Teacher guides support the teacher in the application of the model and in the promotion of active learning. Teachers are guided to adapt the guides to the local context by changing the proposed lesson plans to reflect the priorities and realities of their schools. Student guides are designed to promote active learning: For each lesson, there are basic, practice, and application activities that direct students to work alone, in pairs, in groups, or with their families to acquire skills and knowledge. Even though each student works with his or her own learning guide, all activities are designed to be carried out in note books, so that the learning guides can be reused by the next class.

The new role for teachers as guides and facilitators rather than lecturers requires a new form of **teacher training and support**. When a teacher first starts teaching in an EN school, he or she undergoes a series of pre-service training workshops, each one lasting for several days to a week. The workshops use the EN methodology to train teachers in the correct use of the EN methodology and all of its aspects. Equally important, a special in-service training and support program for teachers helps to foster the exchange of ideas and experiences among teachers. Teachers can visit model schools where the EN model is implemented particularly well. This allows them to observe the model's application in practice and provides them the opportunity for professional interaction. On a monthly basis, they can participate in local "micro centers", which are meetings of teachers of the area that again provide support and an opportunity to discuss problems and to exchange ideas. Additionally, EN teachers receive regular mentoring visits in their own schools, which provides them with feedback and another opportunity to discuss challenges.

Last but not least, EN schools have a focus on **community relations** and a strong appreciation for the local culture. Parents and the wider community are actively involved in the school life, and local culture and knowledge are integrated into the curriculum. This is achieved through school festivities and events, learning activities that require the input of families and community members, and different other instruments such as a "traveling journal" that is passed on from family to family, community maps, community monographs, and more. Furthermore, EN schools have a horizontal and collaborative **administrative structure** that encourages the participation of the entire educational community in school-related decision-making.

1.3 Theoretic support for the Escuela Nueva model

Critical and progressive pedagogy is an important area of theoretical study, and the works of educators such as Maria Montessori (1912) and Rudolf Steiner (1919), with their focus on student-centered learning, have been highly influential. However, the EN model itself was not first conceived on theoretical grounds and then put into practice. Rather, EN emerged as a grassroots solution to the educational challenges faced by practitioners. Likewise, the expansion of the model was not driven by a theoretic paradigm shift in Colombia's primary education sector, but through lobbying efforts of EN practitioners inspired by the perception that EN schools produced tangible results. Even if the theoretical discussion can thus not be based on the model per se, the model's emergence, underlying assumptions, and potential effects on learning outcomes can best be understood against the background of the critical pedagogy movement. This movement emerged in Latin America in the late 1960s and 70s, and its principles are closely associated with the Brazilian educator Paulo Freire.

Paulo Freire was convinced that education is not just about the dissemination of knowledge; it is also a tool of either social oppression or liberation (Freire 1970). Only a type of pedagogy that gives the individuals the central role in their own education can build up an active, informed citizenry. With this in mind, Freire argues against what he calls the *banking model* of education, in which the student is treated as an empty vessel that is to be filled by the teacher through the transfer of knowledge. For him, this traditional model of education—where an all-knowing teacher lectures a group of students who listen quietly—is doomed to fail in educating an active citizenry, as it negates the agency of the individual student by reproducing the oppressive power structure of the society. Freire was convinced that education itself is a political act that should have the aim of liberating all individuals in order for both the "oppressors" and the "oppressed" to regain their humanity. The pedagogy that is employed needs to reflect this true aim of education; it needs to start from each individual's experiences in order to be relevant for the individual's life and to have the potential of political liberation. As knowledge cannot be simply transferred from the teacher to the students, but has to be constructed by the students themselves, only education that is perceived as relevant by the student will result in actual learning. Freire gives an example of his experience in teaching reading and writing skills to a group of peasants. With his method, which focused on first establishing words that had a high practical (and political) significance for the peasants' lives, it took only a few months for peasants to acquire basic literacy.

Even though Freire's writings focus on the education of adults, the EN school model can be understood in the context of his work. Based on Freire's theories, the EN model's student-centered and participatory approach can be expected to lead to increased civic competences through the practice of liberation in everyday school life. It can also be expected to lead to better learning outcomes, as the EN pedagogy is based on integrating the everyday-experience of students into the curriculum in order to make the learning more relevant for students' daily lives.

Freire’s idea of student-centered learning is rooted in constructivist learning theories. Constructivism asserts that new knowledge is gained (constructed) by incorporating new experiences into an existing mental framework. This incorporation can take the form of assimilation (where new knowledge is added to an existing scheme without changing it), or accommodation (where the scheme is altered in order to fit the new experiences) (Piaget 1953; cit. in Ultanir 2012). This theory of learning implies that education has to start with the individual in order to be effective, as every learner has a different existing mental framework that results from their personal biography. Contrasted with traditional teaching methods, where the teacher lectures the same content to the entire class, it is thus expected that the student-centered EN model is more effective in helping students learn.

Student-centered learning theories that follow the traditions of Piaget and Freire have again become very popular among educational theorists starting in the 1990s. A myriad of learning theories have emerged that are all variations of the basic idea, including (but not limited to) problem-based learning, anchored instruction, cognitive apprenticeships, reciprocal teaching, goal-based scenarios, project-based learning, constructivist learning environments, and open learning environments (Land and Hannafin 2000). While these theories vary in their methods, they share the common understanding that learning is best achieved in a student-centered approach. The EN model, while not developed explicitly from the same theoretical basis, joins this list of student-centered approaches.

From the perspective of building citizenship, Freire’s work is strongly influenced by the writings of John Dewey. For Dewey, universal suffrage is not enough to call a society democratic. What is necessary for democracy is to ensure a fully formed public opinion that presupposes effective communication among professional politicians, citizens, and experts. Schools should be the place where children learn and exercise these democratic practices (Dewey 1916). Education as a social process should be designed in a way that reflects the desired (not necessarily the existing) structures of a society, so that it can serve as a model to transform society as a whole. Schools have to foster interest in social relationships and provide children with the necessary tools to take informed action that can provoke social change; these skills are necessary for a society to be fully able to form a public opinion. According to Dewey, the necessary spirit of social cooperation can be developed only if children are taught in a way that reflects that ideal, which requires a pedagogy that encourages students to develop insight, make connections, and participate in the school community (Pitt 2002).

Dewey's idea that the school has to be a live-in role model for democracy in order for students to develop civic competencies has been updated over the course of the century, and is today more actively discussed than ever. Most importantly, student participation in school has come to be seen as the core of citizenship education (Citizenship Advisory Group 1998; Holdsworth 2000; Mager and Nowak 2012). Schools can be a model for a democratic society as a whole. In this setting, students can practice their civic skills in a safe environment by engaging in decision-making processes that affect them, and thus develop the necessary knowledge, skills, and attitudes that will later allow them to take an informed and active role in society (Mager and Nowak 2012). As EN provides such an environment in which students can participate in decision-making both on the classroom- and on the school-level, it is expected that EN students have better civic competences than students taught in conventional schools.

1.4 Problem statement

Two key takeaways emerge from the previous sections. First, Colombia's education sector faces serious problems with regard to quantity, quality, and equity of primary education. Full primary enrollment is still to be achieved – today, about one in ten Colombian children of primary school age do not attend primary school, which means that the quantity of primary education provided needs to increase. Equally important, educational quality needs to improve: Even among those attending school, many children graduate from primary school without having acquired basic numeracy and literacy skills. Finally, educational equity leaves much to be desired: Gender, social background, and area of residency have a large impact on the educational chances of children, leaving especially poor girls from rural areas at a disadvantage.

Second, Colombia has developed a progressive school model that aims to address all of these challenges. The EN model takes a student-centered, active approach to learning and provides flexible answers to the specific challenges that disadvantaged children face. Even though EN itself was not built on a theoretic framework, the theoretic literature on participatory, student-centered learning supports the model's claim and suggests that the school model is suitable to improving both academic outcomes and civic competencies. EN has been in existence for around four decades, and today reaches about half of all Colombian rural schools, as well as a growing number of urban ones.

Taken together, these two takeaways raise a question: Does the EN model work in practice? In other words, if so many Colombian children today attend an EN school, yet schooling outcomes (in terms of overage enrollment, drop-out, and learning achievements) remain so poor, is the school model not living up to its promise? Figure 1 on page 6 suggests that key enrollment indicators have barely improved for the primary sector since the mid-1990s. How can one make sense of this when the EN model has been in wide use for decades?

There are two possible answers to these questions: Either the EN model does not improve learning outcomes, or the EN model works but is not being used in many schools, at least not properly. Clearly, an appropriate education policy response would look very different depending on which of these answers is more in line with empirical evidence. This dissertation will provide some of this evidence and thereby contribute to a better understanding of the challenges that Colombia’s primary education sector currently faces.

2 Literature Review

The aim of this literature review is to summarize the available evidence on Escuela Nueva program implementation and program outcomes. Section 2.1 shows that the knowledge base on program implementation is thin, and that it points towards irregular and incomplete use of the school model's elements. The outcomes of the school model have been studied somewhat more thoroughly, as section 2.2 discusses. The review suggests that the model has a positive effect on learning outcomes, but it also reveals methodological shortcomings in the available literature.

2.1 Program implementation

Previous evaluation studies show that program implementation has been inconsistent. On the one hand, many schools that are officially classified as EN have implemented only parts of the methodology; on the other hand, many conventional schools use some of EN's curriculum and design elements (such as classroom libraries and group work). Consequently, EN and non-EN schools may often differ only on paper.

While not explicitly analyzing the question of program implementation, an early evaluation of the program (Rojas and Castillo 1988, see below) mentions that the student government had been organized in only about half of the EN schools. Similarly, a paper by Loera and McGinn (1992; cited in McGinn 1998, 46) found no significant differences between conventional and EN rural schools in terms of teacher practices. Additionally, there were no significant differences between the two types of schools in terms of the kinds of inputs they receive. The study was not designed as an EN evaluation, yet the national random sample of 180 rural and urban schools included 54 schools classified as EN and 27 classified as conventional rural schools, thus allowing for some comparisons between these school types.

This finding was further substantiated by McEwan (1998). Although the data for this paper was likewise not originally collected to evaluate the EN program, some information is available on program implementation because the surveys collected information on school and classroom inputs as well as on teaching practices in all schools (for instance, on group work and library use). McEwan finds that while EN are more likely than conventional schools to use EN methodologies, there is considerable overlap between the models.

Building on this finding, Forero-Pineda, Escobar-Rodriguéz, and Molina (2006) construct an implementation index for the EN methodology from information that was obtained in a survey in 2001 from students in 25 schools (15 of which were officially EN) in 6 municipalities in the Colombian coffee growing region. Similar to McEwan, the authors conclude that official EN schools tend to have higher index values than conventional schools, but that the difference is not clear-cut.

What is interesting is that variance in the adoption of EN methods can be found both in official EN schools and in conventional schools. On the one hand, this suggests deficiencies in the degree of implementation and irregularities in the use of the model. On the other hand, this may also point to possible spill-over or contamination effects that occur through various channels. For instance, teachers who have been trained in EN pedagogy may have changed their appointment over time and switched to conventional schools, thus taking EN pedagogy components into officially conventional classroom environments. It is also possible that informal contact and exchange of experiences between teachers from EN and non-EN may motivate teachers from conventional schools to adapt some EN elements.

In any case, the official classification does not seem to be a reliable identifier for the use of the EN model. The studies by McEwan and Forero-Pineda, Escobar-Rodriguéz, and Molina are not suited to grasp the extent of program implementation for two reasons. First, the implementation measures used are deficient. The former study uses only proxy data for program implementation (that may additionally be considered outdated today). The latter uses a somewhat more elaborate index consisting of two dimensions (classroom organization[5] and level of teacher training[6]), yet this index is also far from capturing the entire spectrum of EN characteristics (as outlined in section 1.2). Second, the sample size used for both studies is rather small (52 and 25 schools, respectively) and thus not suited for generalizing statements. The present dissertation research project is thus the first one to analyze program implementation in an extensive manner.

There is little discussion in the literature about the reasons for the irregularities in EN implementation among designated EN schools. Anecdotal evidence from both evaluation reports and informal discussions with Fundación Escuela Nueva staff suggest that individual teachers heavily drive the model's implementation. This poses a challenge from a research and implementation science perspective, as many relevant teacher characteristics (motivation, creativity, charisma, etc.) are very hard to measure, let alone control. Further drivers of program implementation are local leaders and authorities, and there even was a push from the federal level for the model's dissemination when it became national policy. Everything put together, it is not clear why schools that officially classify themselves as EN show such large differences in actual

[5] The elements are: desk or table organization (individual, couple, group), the way subjects are presented and developed, the use and number of personal study guides available, the frequency of group activities, the existence and use of classroom libraries and learning corners, and curricular flexibility.

[6] The elements are: number and level of training workshops the teachers have attended, and level of micro-center activities.

implementation. The part of this research project that deals with this topic area is therefore of exploratory, hypotheses-generating nature.

2.2 Program outcomes

Existing evidence provides a strong basis for the belief that the EN model is better in improving learning outcomes and civic competencies than Colombia's conventional school model. Support for this claim can be found on three levels: First, in studies that explicitly evaluate the EN approach; second, in various analyses of the Colombian school system that find elements to be important that are typically associated with EN schools; and third, on a very general level in studies about participatory learning and other program elements in other, international contexts. Evidence from all three levels are presented below for the areas of academic achievements and civic competencies.

2.2.1 Academic achievements

2.2.1.1 Escuela Nueva evaluations

The first evaluation of the EN program was conducted in 1987 and consisted of a simple comparison of average test scores between EN and non-EN students (Rojas and Castillo 1988; cit. in Velez 1991; and Psacharopoulos, Rojas, and Velez 1992). Data was collected for 3,033 students from 168 EN and 60 conventional schools, where EN were randomly selected out of all schools that had, according to official records, completely implemented all EN components for at least five years. Comparison schools were chosen from the same districts where EN schools were part of the sample. As students had no choice in what school to go to but were required to enroll in the closest school, the sample is free of self-selection bias on the student-level. A questionnaire was given to students and teachers to obtain information on student, teacher, and school characteristics. The study makes no attempt to take implementation issues into account but assumes that the model was homogenously implemented across all schools. The purely descriptive data analysis suggests that scores are better in EN schools, but does not control for any other factors that may explain the differences. Velez (1991) recognizes the limitations of the descriptive study design and expands the analysis of the same data set by fitting a standard OLS model to the original data for around 1.677 3rd graders, controlling for some student characteristics. This analysis suggests that EN students do significantly better in mathematics and Spanish.

The same data set is again analyzed by Psacharopoulos, Rojas, and Velez (1992). Standard OLS estimations are run that include student, family, school, and teacher characteristics in order to control for the socioeconomic and environmental factors that might influence school achievements. The study's main finding is that achievements in mathematics and language are generally higher for EN students. Furthermore, drop-out rates are lower for EN.

These results were generally confirmed in the previously mentioned paper published by McEwan six years later (McEwan 1998). His research design is also a posttest-only comparison between EN and conventional schools, but McEwan uses secondary data for a representative random sample of "B calendar" rural primary schools from three Colombian departments (Valle, Cauca, and Nariño). The dataset consists of 52 schools, 24 of which are officially classified as EN. Students' self-selection is again not an issue as the schools are relatively isolated and students have no choice in which school to attend. The data had been collected through surveys that were completed by each school's principal, the teacher of the surveyed classroom, and the students, who also were administered tests in mathematics and Spanish. McEwan uses standard OLS regression analysis in order to control for student, school, teacher, and family characteristics. McEwan finds significant positive effects of the EN model on Spanish and mathematics achievement, even after controlling for specific input differences, which suggests that the EN model is more than just the sum of its measurable input factors.

2.2.1.2 Other studies on determinants of learning outcomes in Colombia

Apart from the studies that explicitly look at EN's academic outcomes, some indications of the model's positive effects can also be found in the more general literature on the determinants of educational achievements in Colombia. Three studies are of particular interest in this context, as their authors explicitly discuss the EN model, despite a broader focus of their studies.

Casassus et al. (2000) analyze the results of an international standardized test, the *Primer Estudio Internacional Comparativo en Lenguaje, Matemática y Factores Asociados*, organized in 1997 by the *Laboratorio Latinoamericano de Evaluación de Calidad de la Educación* (LLECE) among 3rd and 4th graders in countries of Latin America and the Caribbean. They find that about one-third of the variance in LLEVE scores between schools is explained by school-related factors such as the existence of a large library, a good classroom environment, well-trained teachers, classroom-level tutors, and a low student-teacher ratio. Furthermore, parents' involvement in the school life, as well as a higher level of teacher autonomy have positive effects. Given that these factors are all part of the (ideal-typical) EN model, it comes as no surprise that the authors explicitly point out that EN schools seem to achieve better results. Note, however, that this cross-country study design does not allow for well-founded conclusions about the EN model.

Nuñez el al. (2002) analyze differences in student achievements between Colombian public, private, and educación contratada[7] schools. The authors find that, taking into account individual,

[7] Schools that are operated by the Church but co-financed by the state.

family, and school characteristics, private schools generally outperform public ones, even though large heterogeneity exists within each school type. Unobservable factors such as the underlying incentive structure for teachers are hypothesized to drive this result. Furthermore, the analysis finds that in municipalities where EN schools operate, public schools tend to outperform private ones. While the study design does not allow for precise conclusions regarding EN's effect (mainly because it does not explicitly identify them), the paper's results can be seen as an indication of the positive impact of the school model on learning outcomes.

A similar conclusion can be drawn from a study by Uribe (1998) that uses a methodologically very different approach. Uribe estimates production possibility frontiers for Colombian primary schools in an attempt to measure their efficiency. Her goal is to determine what the highest possible output (defined as the results in the standardized test Pruebas SABER) is, given a certain level of inputs (in terms of student, teacher, and school factors). After identifying factors that are related with students' achievements (education of parents, quality of home, education level of teacher, school equipment), she calculates the production possibility frontier and compares each school's actual performance with this benchmark. Uribe finds that EN schools seem to outperform other schools: Between 17% and 20% of EN are efficient, compared to only between 5% and 6% of conventional schools; about 74% of all efficient schools are EN. This again suggests that learning outcomes in EN schools are better than in traditional ones, given the differences in background characteristics.

More indirect evidence pointing to the efficacy of the EN model comes from the body of literature that analyzes the determinants of learning in Colombia without explicit reference to the EN model. In these studies that did not take into account the specific school model, important determinants of learning that also form part of the EN toolbox include: a library (Casassus et al. 2000), a good classroom environment (Casassus et al. 2000), well trained teachers (World Bank

2009; Casassus et al. 2000; Zambrano Jurado 2013; additionally, for secondary education: Jola S. 2011; Cepeda-Cuervo and Núñez-Antón 2013), classroom-level tutors (Casassus et al. 2000), parents' involvement in the school life (Casassus et al. 2000), high levels of teacher autonomy (Casassus et al. 2000; Nuñez et al. 2002; note though that World Bank 2009 finds that this factor is not significant), and targeted and readily available instruction materials (Zambrano Jurado 2013; additionally for secondary education: Manzano Lopez and Ramirez Zambrano 2012). Thus, as these are all elements of the EN model, one would expect schools with high levels of EN program implementation to have better learning outcomes.

2.2.1.3 Evidence from other countries

Finally, when moving up another level to look at international evidence, an enormous pool of evaluation literature opens up that tries to analyze what works in education. Of course, the further removed an evaluation is from the cultural environment and pedagogical model of EN, the more careful one has to be when transferring the findings. Therefore, this discussion will be limited to a few main points.

First, it is worth mentioning that the EN program has also been implemented in other countries, most extensively in Guatemala through the Nueva Escuela Unitaria (NEU) model. Some studies are available that evaluate these international adaptations; as far as international studies are concerned, the findings of these evaluations are arguably most readily transferable to Colombia's EN model. For Guatemala, Juárez and Associates (2003) and Kline (2002) analyze the school model's impact on academic achievements. Both papers agree that Guatemala successfully adapted the EN model to the Guatemalan context. Children in the NEU schools generally outperform their peers in conventional rural schools. The program seems to be of particular benefit to girls. There are no high-quality evaluations of adaptions in other countries, yet two papers (Mulkeen and Higgins 2009; Obwoya et al. 2004) analyze the model's implementation in

Uganda. The two papers agree that the program's adaptation was marked by substantial implementation failures, to the extent where the pilot schools lacked most key features of the original model (for instance, "multigrade teaching" meant, in practice, most of the time "quasi-monograde" instruction where the teachers lectured one age group at a time). There is some limited anecdotal evidence that where the model was implemented more faithfully, results were good.

Second, a look at the general "what works in education" literature suggests that the EN model combines many of the elements that are commonly identified as effective in improving learning outcomes. Meta-analyses of evaluations of educational interventions provide evidence for the effectiveness of the following elements of the EN model: Classroom libraries (Anderson 2005); providing each student with a learning guide, as there is overwhelming evidence for the effectiveness of distributing sufficient textbooks and learning materials (Herz and Sperling 2004; Anderson 2005); targeted and relevant teacher training (Herz and Sperling 2004; Anderson 2005; Glewwe et al. 2011); tutors and mentors, in the case of peer-tutoring for both the mentors and the mentored (Anderson 2005; Glewwe et al. 2011); cooperative learning in pairs or small groups (Marzano, Gaddy, and Dean, n.d.); homework and independent practice (Marzano, Gaddy, and Dean, n.d.); "public" progress charts that provide symbols of recognition for advancements (Marzano, Gaddy, and Dean, n.d.); a constructivist pedagogical approach that helps students to activate prior knowledge and to relate new materials to their daily lives and experiences, for instance through learning corners and community-based curricula (Merrill 2002; Marzano, Gaddy, and Dean, n.d.; Marlowe and Page 2005); participatory classrooms and student participation in decision making at school (Mager and Nowak 2012; to some extent Davies et al. 2006); community and parents' involvement and school decision-making (Herz and Sperling 2004; Glewwe et al.

2011); and individualized flexible schedules and learning programs (Herz and Sperling 2004; Kremer and Holla 2009).

2.2.2 Civic competencies

2.2.2.1 Escuela Nueva evaluations

EN's impact on its students' civic competences and democratic values and behaviors has been almost as interesting to researchers as the academic outcomes. The above-mentioned first round of evaluation studies included a component on EN's impact on civic engagement; the authors of all three papers agree that civic engagement indicators are generally higher for EN than for non-EN students (Rojas and Castillo 1988; Velez 1991; Psacharopoulos, Rojas, and Velez 1992).

Two evaluation studies have looked explicitly at these non-academic outcomes. Following a research design developed by Chesterfield (1994, see below), Pitt (2002) collected data through questionnaires, semi-structured interviews, and observations in three EN and three conventional schools. The EN schools were chosen randomly out of all EN in those parts of Riseralda and Caldas that were considered safe. The comparison schools were chosen by selecting geographically, socially, and economically matched communities where EN did not exist. Pitt conducted comprehensive classroom observations coding the behavior she observed in students, and conducted structured and semi-structured interviews with teachers, students, and alumni. No attempts were made to control for socioeconomic and environmental factors that may be responsible for different outcomes, other than the process of matching comparison schools based on similar characteristics. The study's main finding is that the classroom climate in EN was more positive than in comparison schools, and that the incidence of democratic classroom behaviors was higher in the EN than in conventional schools. No significant difference was found in civic participation of graduates from EN and conventional schools.

To date, the most methodologically rigorous EN evaluation study was a paper by Forero-Pineda, Escobar-Rodriguéz, and Molina (2006). The authors use a quantitative, multilevel regression analysis design to evaluate EN's impact on peaceful social interaction, and probit, ordered-probit, and logit models to study the relationships between schools and families. The study is based on a 2001 survey in 25 schools. Information was collected from 989 students, 49 teachers, 24 school principals, 343 parents, and 179 alumni. Note that 25 schools is a relatively small sample size, especially for a multilevel design. Even so, by using a hierarchical model the authors can control for socio-economic and environmental factors on various levels when estimating the effect of EN. The study's main findings are that EN has a positive impact on peaceful social interaction, on family behavior, and on participation in community life. Additionally, alumni are more inclined towards participatory democracy.

2.2.2.2 Other studies on determinants of civic competencies among students in Colombia

Other authors have analyzed the Colombian education sector without specifically looking at EN to see how peace and democratic behavior are promoted. In a series of articles, a research group from the Universidad de los Andes in Bogotá describes and evaluates the program *Aulas En Paz* that aims at promoting citizenship competencies and peaceful behaviors among Colombian primary school students through classes and extracurricular activities (Chaux 2009; Chaux and Velásquez 2009; Jiménez et al. 2008; Ramos, Nieto, and Chaux 2007). The program is based on seven pedagogical principles, some of which also form part of the EN methodology, such as cooperative learning in small peer groups, learning by doing, and relating the learning activities to the daily lives of the students. Clearly, the evaluation results of one program cannot just be transferred to a pedagogical model. However, the fact that the authors stress the importance of the cooperative aspect of the program that allows students to experience democratic interactions and non-violent conflict resolution can be carefully interpreted as an indication that the cooperation-based learning model of EN may have similar positive effects.

Another informative study about Colombian schools' impact on civic competencies was published by Diazgradanos and Noonan (2015). The authors use multilevel tobit regression analysis to analyze whether safer and more participatory school environments lead to less supportive attitudes toward violence. The data comes from the 2005 round of the Pruebas SABER that also contained a section on citizenship competences; the indices for attitudes toward violence, for a safe and for a participatory school environment, respectively, were constructed from a total of fourteen Likert-statements on the test. The results show clearly that supportive attitudes toward violence can be decreased by a safe and a participatory school environment. Even more important in this context, a participatory learning environment is even more powerful for this purpose than a safe learning environment: While children in safe and participatory schools showed the least support toward violence and those in unsafe, non-participatory schools showed the highest, children in unsafe but participatory environments were less supportive toward violence than those in safe but non-participatory schools. In fact, the effect of participation was found to be twice as large as the effect of safety. The study did not identify specific educational models or school programs, but can be seen as a strong indication that EN as a highly participatory school has a large potential to increase civic competencies and peaceful, democratic behaviors among its students.

2.2.2.3 *Evidence from other countries*

Finally, as was the case for academic outcomes, the broader international literature gives reasons to be confident that EN improves citizenship and democratic behavior indicators. For the Guatemalan adaption of the EN model, the Nueva Escuela Unitaria (NEU), Chesterfield (1994) carried out an extensive qualitative analysis of classroom interactions and found that students in

NEU schools showed improved democratic behaviors, compared to non-project schools. This result was reinforced by another study (de Baessa, Chesterfield, and Ramos 2006) that used multiple methods and confirmed that students in NEU schools engage in significantly more democratic behaviors.

The above-mentioned literature survey by Mager and Nowak (2012) found positive effects of participatory education not only on academic achievements, but also on democratic skills and citizenship. These effects are most likely when participation takes the form of student participation in councils, as is the case in EN's student government and student committees. To a lesser extent the literature survey found the same effect for student participation in class decision-making – likewise a common practice in EN classrooms. A positive effect on democratic skills and attitudes as well as on problem-solving and conflict-resolution skills is also affirmed by the literature survey of Davies et al. (2006).

In summary, there is empirical evidence directly from previous EN evaluations and from studies of education and democratic behavior both within Colombia and in a broader international context that the participatory, student-centered EN model promotes civic competencies.

3 Research Methodology

To solve the puzzle that was laid out in the first chapter (poor learning outcomes in Colombia despite apparent wide use of the Escuela Nueva model), this chapter first formulates research hypotheses and the related research questions (section 3.1). Subsequently, section 3.2 describes the methods that were used to test the hypotheses and answer the research questions. Finally, section 3.3 presents the dataset used for the study.

3.1 Research hypotheses and research questions

This dissertation adds to the empirical literature by collecting evidence against the following two null hypotheses:

H_0^1: There is no difference between observed classroom practices and expected classroom practices based on the official school classification.

H_0^2: There is no *ceteris paribus* difference in learning outcomes between students in conventional schools and in EN schools.

The first null hypothesis deals with the level of implementation of the model, and the second one with the model's effect on learning outcomes. For each of the null hypotheses, the alternative hypotheses of interest can be formulated in different ways:

Program implementation:

$H_A^{1.1}$: There are differences between the ideal-typical EN model and observed program implementation (i.e., implementation is incomplete).

$H_A^{1.2}$: There are differences between official classification and observed program implementation (i.e., program implementation is irregular and the distinction between EN and non-EN schools is not clear-cut).

$H_A^{1.3}$: There are differences in the observed degree of program implementation based on political priorities and teacher preferences.

Program outcomes:

$H_A^{2.1}$: *Ceteris paribus*, students in EN schools achieve higher test scores than students in conventional schools in (a) numeracy; (b) literacy; and (c) civic competences.

$H_A^{2.2}$: *Ceteris paribus*, the effect of socioeconomic status on learning outcomes is smaller in EN schools than in conventional schools.

$H_A^{2.3}$: *Ceteris paribus*, the gender gap in learning outcomes is smaller in EN schools than in conventional schools.

$H_A^{2.4}$: The difference in learning outcomes between EN schools and conventional schools depends on which program elements are being implemented.

Based on the program's goals (section 1.2), the theoretical literature discussing the benefits of progressive education (section 1.3), and existing empirical literature evaluating the EN model (chapter 2), this research was guided by the hypothesis that, other things being equal, students attending an EN school achieve superior results compared to their peers in conventional schools in all three learning outcome dimensions. Given the traditionally strong intergenerational transmission of schooling attainments, EN's positive effect is expected to be particularly strong for children from disadvantaged socioeconomic backgrounds, because these children have less access to support systems at home.[8] Furthermore, because of the student-centered approach of the model, it is expected that gender gaps in learning outcomes are smaller in EN schools than in conventional schools.

The literature gives relatively little indication as to which program elements might be of particular importance. In fact, previous research suggests that trying to separate individual design elements may be futile, as the institutional package may be more important than singular parts (McEwan 1998). Therefore, research hypotheses $H_A^{2.4}$ is meant to explore potential differences in the effect of individual program elements.

To be sure, while these hypotheses state that EN is suitable for improving learning outcomes, the claim is not that the model addresses all shortcomings of the education system or that it can help solving all issues that lead to poor learning outcomes. On the most general level, this research project only considers supply-side constraints – that is, the analysis is based on the assumption that learning outcomes *can* be improved if the right type and amount of school input factors is provided. As shown in the literature review, there are good reasons to believe that this is the case. However, it is possible that demand-side constraints play an equally important role in poor learning outcomes: if children or their caregivers perceive education to be of little practical importance for their lives, no amount of inputs will make them engage in learning. Put in more economic terms, if the returns on investment in education are low or perceived to be low, students will have little incentive to spend time and energy on getting educated. These demand-side issues are not explicitly addressed in this research.

[8] It is possible that the program works better for children from moderately poor than from extremely poor families. The experience from other effective education programs shows that they often cease to produce positive outcomes once socioeconomic deprivation is overwhelming.

The research questions that guided this dissertation reflect the different formulations of the alternative hypotheses. There is one set of research questions related to program implementation, and a second set related to learning outcomes.

1. **Program implementation:**
   - 1.1. To what extent are the EN methodology and its individual components implemented in Colombia?
   - 1.2. How well does the official classification (EN or not) predict the use of the EN methodology?
   - 1.3. What determines whether or not and to what extent a school uses the EN methodology?
2. **Program outcomes:**
   - 2.1. Are EN schools more effective than conventional schools in improving students' (a) numeracy; (b) literacy; (c) civic competences?
   - 2.2. Are EN schools more effective than conventional schools in improving learning outcomes for children from disadvantaged socioeconomic backgrounds?
   - 2.3. Are EN schools more effective than conventional schools in decreasing gender gaps in learning outcomes?
   - 2.4. Which of EN's elements are particularly strongly connected to students' learning success and civic competences?

3.2 Methods

The literature review shows that, to date, a range of different methods have been used to study the outcomes of the EN model. These include studies based on qualitative interviews, focus groups, observations, or on simple descriptive statistics (Misión Social del Departamento Nacional de Planeación de Colombia 1997; Pitt 2002; Sarmiento Gómez 2006; Moore, Florez, and Grajeda 2010; ICFES 2011); quantitative studies based on basic econometric estimation techniques (Velez

1991; Psacharopoulos, Rojas, and Velez 1992; McEwan 1998; Benveniste and McEwan 2000); and a few studies that use more advanced quantitative methods, such as non-parametric data envelopment analysis (Uribe 1998), propensity score matching and logit/probit estimations (Nuñez et al. 2002), and multilevel modeling (Forero-Pineda, Escobar-Rodriguéz, and Molina 2006).

If the goal is to make reliable and generalizable statements about EN's effect on learning outcomes, all of these studies have shortcomings. The first group of qualitative and descriptive studies is by definition not suited for general conclusions and to identify causality. Studies in the second group control for some external factors, but ignore the nested structure of the data and may thereby find statistically significant effects where there actually are none (see below). Furthermore, most of these studies (apart from McEwan 1998 as discussed above) ignore varying levels of program implementation and may as a result not actually measure the effect of the school model. The latter is also true for studies using more advanced estimation techniques, as only Forero-Pineda, Escobar-Rodriguéz, and Monlina (2006) clearly identify EN schools and attempt to measure implementation. However, the sample size used by Forero-Pineda et al. was arguably too small (25 schools) for their chosen approach to generate valid results at the school-level. Furthermore, it was confined to only six Colombian municipalities.

In short, while there are a number of papers that deal in one way or another with the EN school model, the validity of their results is questionable based upon weak methods. In order to add to the body of literature, this dissertation uses a mixed method approach. The design consists of the following three elements:

1. A country-level analysis of all Colombian primary schools based on test results from the standardized test Pruebas SABER 2013, using quantitative multilevel methods

2. A department-level analysis of a representative sample of rural schools from one department (Quindío) based on Pruebas SABER 2013 results and new data on program implementation, using quantitative multilevel methods
3. The triangulation of the results based on a qualitative study of a small number of schools using interviews and observations, combined with key-informant interviews.

A combination of these three research strategies allows for conclusions to be drawn about various levels of program implementation and program outcomes. Table 1 provides an overview on the research design, summarizing which of the research questions are addressed through which element of the research design. Note that the "X" only denotes that the design element addresses the respective research question. The internal and external validity of the corresponding results may differ and will be discussed in more detail in the closing chapter.

Table 1 Research Design Overview

		Country-level	**Department-level**	**Qualitative Part**
1. Implementation	**Extent of Implementation**		X	X
	Drivers for Implementation			X
	Accuracy of official classification		X	
	Distinguishability of school models		X	X
2. Outcomes	**Superiority of EN methodology**	X	X	X
	Role of student's background	X	X	X
	Role of program elements		X	X

3.2.1 Quantitative analysis

The quantitative parts of this study mainly use multilevel modeling techniques, an approach that has rarely been used for studies of national scope in Colombia (Rangel and Lleras 2010). The approach has been applied in the context of EN, but only on a small scale. What follows is a brief introduction to this method and an explanation as to why it is appropriate for evaluating EN’s impact on learning outcomes. At the end of this section, a preliminary multilevel model of the effect of EN on learning outcomes is introduced.

3.2.1.1 Multilevel modeling: theoretic background

As anywhere in the world, academic performance of Colombian students is the result of a wide range of factors. Some of these factors are characteristics of the individual student (such as gender, intelligence, studiousness, etc.); some of them are characteristics of the student’s home and family environment (such as the family’s socioeconomic status, family wealth and income, parents’ education, the value that a family puts on education, etc.); some of them are characteristics of the school (availability and accessibility, number and quality of teachers, pedagogy employed, school infrastructure, school and classroom climate, etc.); some of them are characteristics of the municipality or department (average income and education level, urban or rural areas, education policies and funding, accessibility of the area, ethnic composition, the prevalence of conflicts and crime, etc.); and some are even characteristics of the country as a whole (general characteristics of the education system, education policy and decision making mechanisms, education budgets, etc.). A multilevel framework is a research design that addresses these empirical relationships by explicitly taking into account the multilevel structure.

When evaluating student performance, it is not enough to simply account for the different factors on every level by including them in a standard analysis. It is important to also take into account the very specific cluster structure that underlies the data, as the individual observations are not independent from each other. Figure 4 (adapted from O'Connell and Reed 2012, 6) depicts this fact for a simplified multilevel model. The outcome of interest, individual student performance, resides at the lowest level of the hierarchy, together with other student characteristics. Student characteristics on this first level vary among all students, both within and between different schools (and regions). Students are, however, nested within different schools that vary in important characteristics, both observed ones (such as the number of teachers) and unobserved ones (such as school climate). These level-two characteristics vary among students from different schools, but not among students within schools. Schools on their part are nested within regions that share common characteristics, again both observed and unobserved ones. These level-three characteristics are the same for all schools, and thus for students within the region, and hence do not vary between them, but they do vary between students and schools from different regions. Extending this basic model further is straightforward. What is important is that variability exists on each level of the multilevel analysis; for instance, if an evaluation is only carried out within region 1 in the universe of Figure 4, level-three characteristics would not be included in the analysis as they would not vary across the sample.

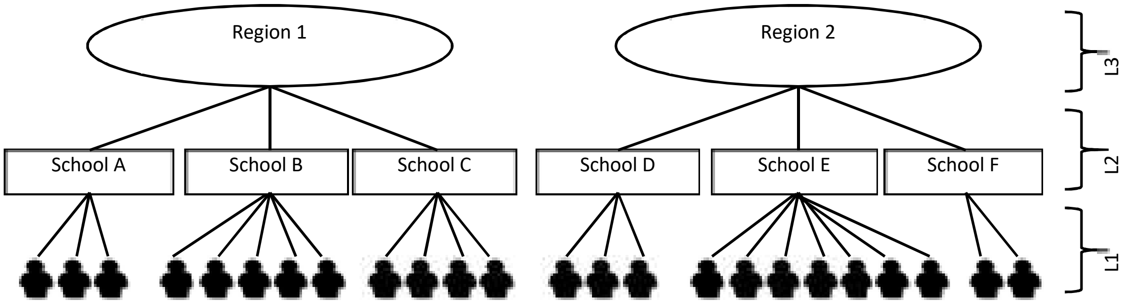


Figure 4 Simplified model of cluster data (adapted from O'Connell and Reed, 2012)

While standard econometric modeling techniques (that is, Ordinary Least Squares, OLS) have been used to evaluate student performance, the clustered data structure clearly violates the Gauss-Markov Assumptions underlying OLS modeling. With clustered data, one cannot assume that the observations are statistically independent, and in fact it is easy to show empirically for a clustered data set that observations obtained from individuals (students) within the same cluster (school or region) are more similar to each other than to observations from different clusters (O'Connell and Reed 2012, 7). This point will be explained formally below by analyzing the error term in the econometric model. Not taking into account the similarities in clustered data sets leads to an underestimation of standard errors and thus to a higher-than-normal Type I error rate (McCoach and Black 2012, 27), i.e., the incorrect rejection of the null hypothesis: for instance, a pedagogical intervention is found to have a positive effect when in fact it does not. This is the reason why student performance should better be evaluated within a multilevel framework that uses multilevel models.

3.2.1.2 Formal example

The following example demonstrates this argument formally and shows where a simple OLS model goes wrong. For the sake of the demonstration a simple econometric multilevel model of educational outcomes is developed which only includes two independent variables (one for level-one and one for level-two); yet, the model can be easily expanded and more variables on each level of analysis can be included. The discussion is based on Raudenbush and Bryk (2002).

Let us assume that a student's academic performance (*SCORE*, their score on a nationwide standardized language exam) depends on her family's socioeconomic status and on the school model she attends (EN or not). Both a higher socioeconomic status and attending an EN school are hypothesized to improve the test score. Socioeconomic status (*SES*) is treated as a student-level characteristic and measured by parents' educational attainment (centered around the sample mean). Additionally, the effect of the school model is assumed to differ by socioeconomic status, which requires the inclusion of the interaction term EN**SES.* This simple model only includes one level-two variable, namely, the school model. The school type is modeled as a dichotomous variable (*EN*) indicating whether (EN=1) or not (EN=0) the school is an EN. First, consider the following simple standard model that does not account for the multilevel data structure:

Eq. 1: $$SCORE_i = \beta_0 + \beta_1 SES_i + \beta_2 EN_i + \beta_3 (EN * SES)_i + r_i$$

where $SCORE_i$ is the educational outcome of interest for student i, namely, the standardized test score on the language exam; β_0 is the intercept (in this example, β_0 is the expected test score of a student who attends a traditional school and whose socioeconomic status is the sample average); β_1 is the expected change in the exam score, for a student in a conventional school, when socioeconomic status increases by one unit; SES_i is the measure for socioeconomic status, measured by parents' average years of education; β_2 is the expected difference in the exam score between a student in an EN and a student in a conventional school, for a student of average socioeconomic level ($SES = 0$); EN_i is the dummy indicating whether a student attends an EN ($EN = 1$) or not ($EN = 0$) [Note that in this model, EN is defined at the level of the individual, i] ; β_3 is the expected additional difference in the exam score between a student in an EN and a student in a conventional school for each unit increase in the socioeconomic level; $(EN * SES)_i$ is the interaction of school model and socioeconomic status; and r_i is the error term and represents a unique effect associated with each student.

This simple standard model implies that both the intercept (β_0) and the effect of attending an EN (β_2) are the same for all students. It assumes that the error term r_i has an expected value of 0 for any given student once SES_i and EN_i are controlled for; it is independent of the explanatory variables ($E(r_i)|X = 0$). r_i is normally distributed with a mean of zero and a constant variance σ^2 (that is, $r_i \sim N(0, \sigma^2)$). However, given the multilevel structure of the model, there is good reason to believe that this model does not capture reality very well. For instance, it is likely that the mean test score varies between schools and that this variance is not entirely random but can be explained by certain school characteristics; the same is true for the slope estimates. It is also unlikely that the error term can indeed be expected to be independent and of constant variance; for instance, the error term is likely to be bigger for students in schools with especially motivated teachers, and the error variance is likely to be higher in schools with a very heterogeneous group of students than in schools where students are more similar to the sample's average student. Therefore, another model is necessary.

Consider the following model for *student-level determinants* of educational outcomes:

Eq. 2: $$SCORE_{ij} = \beta_{0j} + \beta_{1j} SES_{ij} + r_{ij}$$

where the interpretation of all parameters is similar to Eq. 1, with the difference that the subscript j was added in order to point out differences across schools: $SCORE_{ij}$ is the educational outcome of interest for student i in school j; β_{0j} is the intercept for school j (in this example, β_{0j} is the expected test score of a student attending school j whose socioeconomic status is the sample average); β_{1j} is the expected change in the exam score in school j when socioeconomic status increases by one unit; SES_{ij} is the measure for socioeconomic status for student i in school j; and r_{ij} is the error term and represents a unique effect associated with student i in school j.

Note that Eq. 2 does not contain the dummy EN, as EN is not a level-one but a level-two variable and thus only varies between students from different schools, not among students from the same school. Because of the subscripts j for each parameter, Eq. 2 is actually a set of equations with different parameters for every school; hence, it does not include any school-level predictors. The interaction term is missing for the same reason.

School-level characteristics can impact the student-level relationship between socioeconomic status and test scores in two major ways. First, it is possible (and even likely) that average student performance varies across schools. For instance, this study's research hypothesis expects average test scores to be higher in EN schools than in conventional schools. In other words, the intercept β_{0j} is expected to vary by school type. Second, it is equally possible that the effect of socioeconomic status on a student's performance is different for different school models. For instance, children from disadvantaged backgrounds may be less likely than children with a higher socioeconomic status to perform well in conventional schools, as the latter might find it easier to find other help in case they need it. Progressive pedagogy might be especially beneficial for disadvantaged students as a student-centered approach is more likely to take into account the weaknesses and strengths of each individual and to help disadvantaged students to catch up. Hence, the slope estimate β_{1j} is also expected to vary by school type. Formally, this second level is modeled as follows:

Eq. 3a: $\beta_{0j} = \gamma_{00} + \gamma_{01}EN_j + u_{0j}$

Eq. 3b: $\beta_{1j} = \gamma_{10} + \gamma_{11}EN_j + u_{1j}$

where β_{0j} and β_{1j} are the intercept and slope parameter from Eq. 2; EN_j is the dummy variable indicating whether a student attends an EN ($EN = 1$) or a conventional school ($EN = 0$) [in this model, EN is defined at the level of the school, j]; γ_{00}, the intercept, is the mean test score for conventional schools; γ_{01} is the mean achievement difference between EN and conventional schools; γ_{10}, the intercept, is the average SES-achievement slope in conventional schools; γ_{11} is the mean difference in SES-achievement slopes between EN and non-EN; u_{0j}, the error term, is the unique effect of school j on mean achievement given EN_j; and u_{1j} , the error term, is the unique effect of school j on the SES-achievement slope given EN_j.

When substituting Eq. 3a and 3b into Eq. 2, the model becomes:

Eq. 4: $$SCORE_{ij} = \gamma_{00} + \gamma_{01}EN_j + \gamma_{10}SES_{ij} + \gamma_{11}EN_jSES_{ij} + u_{0j} + u_{1j}SES_{ij} + r_{ij}$$

Even though Eq. 1 and Eq. 4 include the same dependent and independent variables, the two models vary considerably. The most important difference concerns the nature of the error term. While Eq. 1 had a simple random error, r_i, the random error of Eq. 4 is more complex and consists, in fact, of three terms: $u_{0j} + u_{1j}SES_{ij} + r_{ij}$. This implies that:

- The errors are not independent: u_{0j} and u_{1j} are the same for each student in school j.
- The error variance is not constant: u_{0j} and u_{1j} vary across schools, and SES_{ij} varies across students.

Thus, the Gauss-Markov Assumptions necessary for accurate hypothesis testing based on OLS do not hold, so that standard regression analysis is not appropriate. However, iterative maximum likelihood procedures can be used in order to estimate the random parameters for a multilevel model, provided that the data sample is large enough. This is the approach that the quantitative section of this dissertation takes in order to estimate EN's effect.

3.2.1.3 Alternative strategies for dealing with clustered data

Multilevel modeling is not the only way to deal with clustered data. The problem of errors that are not independent and not constant can also be resolved by calculating clustered standard errors while still using OLS estimations. This is the approach taken by standard survey estimation techniques. Correcting standard errors for clustering leads to less efficient but unbiased estimates. The big advantage of this method is the less complicated estimation process. However, this strategy does not allow for an analysis of the nature of the variance between schools (or between clusters more generally). It also does not produce separate estimates for the individual clusters (in this context, schools, municipalities, or departments). In this sense, survey estimation techniques see the clustered structure of the data as a nuisance to be corrected, not as a source for additional insights. For these reasons, multilevel modeling is the preferred approach for this dissertation.

A major practical disadvantage of multilevel estimation is that Stata13 does not allow for the inclusion of a finite sample correction factor in its multilevel estimation command (see section 3.2.1.8). This is not a problem for the country-level study, which is based on a large sample—but it means that standard errors are overestimated in the department-level study, which is based on a sample that consists of observations from over 50% of the study population. Therefore, the department-level multilevel estimates in chapter 6 are complemented by survey estimates as an alternative way to deal with school-level clusters while at the same time accounting for the large relative sample size.

3.2.1.4 A remark on notation

The “two-stage-formulation” (departing from a level-one model, and explaining its components as a function of level-two equations) is the way that Raudenbush and Bryk (2002) and Snijders and Bosker (2011) approach multilevel modeling – but it is not the only approach. Rabe-Hesketh and Skrondal (2012) propose a “one-stage formulation”, where regressors from all levels are modeled at the same time, and random effects are added progressively without first separating the model into its different levels. While the two-stage approach provides an intuitive way of understanding multilevel models, the one-stage approach is more intuitive for actually estimating the model. It also seems to limit the risk of over-identification; it appears that papers which use the two-stage approach tend to include more cross-level interaction terms and more random coefficients in the models, "because the level-2 models look odd without residuals" (Rabe-Hesketh and Skrondal 2012, 213).

The rest of this dissertation will use the one-stage formulation approach, and the notation used by Rabe-Hesketh and Skrondal. To show how the notations compare, Eq. 4 from above is contrasted below with Eq. 5, which is the same model using notation as proposed by Rabe-Hesketh and Skrondal:

Eq. 4: $$SCORE_{ij} = \gamma_{00} + \gamma_{01}EN_j + \gamma_{10}SES_{ij} + \gamma_{11}EN_jSES_{ij} + u_{0j} + u_{1j}SES_{ij} + r_{ij}$$

Eq. 5: $$SCORE_{ij} = \beta_0 + \beta_1 EN_j + \beta_2 SES_{ij} + \beta_3 EN_jSES_{ij} + \zeta_{1j} + \zeta_{2j}SES_{ij} + \varepsilon_{ij}$$

The main difference is that the all fixed parameters in the model are denoted as β_n (for a model with N fixed parameters, independent of the variables' level), all random parameters (except the level-one residual) are denoted as ζ_{mj} (for a model with M random parameters, independent of whether the parameter refers to a random intercept or a random coefficient), and the level-one error term is denoted as ε_{ij}. There is also a term for the total error: ξ_{ij}, which is defined as the sum of all random parts of the model (in the case of Eq. 5, the total error is $\xi_{ij} = \zeta_{1j} + \zeta_{2j}SES_{ij} + \varepsilon_{ij}$).

3.2.1.5 Conceptual framework: Evaluating Escuela Nueva in a multilevel framework

When evaluating Colombia's EN program, the data structure discussed above needs to be taken into account just as in any other educational outcome evaluation. It is important to identify all relevant factors that may have an effect on schooling outcomes, to establish the level that they belong to, and to identify potential interactions. In this section, first, four important levels of analysis are established; second, possible explanatory variables on each of these levels are identified; and lastly, a model of the structural effects (intercepts and slopes) on each level is outlined.

Level one always contains the dependent variable and may contain a varying number of independent variables. Typically, in the context of educational evaluations—and this is also the case for this analysis—the outcome of interest will be a student-level variable (academic achievement, as measured by standardized test scores; or civic behavior, as measured by a score on a nationally standardized survey instrument). Hence, the student-level represents the lowest level of data.[9]

A review of the literature on determinants of schooling outcomes in Colombia shows that there are many student-level variables that influence learning outcomes. These include:

- Gender (evidence for primary schools: World Bank 2009; Zambrano Jurado 2013; evidence for secondary schools: Carcamo Vergara and Mola Avila 2012; Baron 2012);
- Interest and joy in learning or going to school (Zambrano Jurado 2013; Jola S. 2011 [for secondary education]);
- Socioeconomic background and parental education (evidence for primary schools: World Bank 2009; Rangel and Lleras 2010; evidence for secondary schools: Jola S. 2011; Cepeda-Cuervo and Núñez-Antón 2013; Gaviria and Barrientos Marín 2001; Manzano Lopez and

[9] However, level one does not need to be the student-level. For instance, if the evaluation was to focus on drop-out rates as the outcome variable, schools would represent the lowest level of analysis, and the main explanatory variable (implementation of Escuela Nueva) would be located at the same level as the outcome variable.

Ramirez Zambrano 2012; Bonilla Mejia and Galvis 2012; Ayala Garcia, Marrugo Llorente, and Saray Ricardo 2011);

- Having books, computers, or educational materials at home (evidence for primary schools: World Bank 2009; Zambrano Jurado 2013; evidence for secondary schools: Jola S. 2011; Ayala Garcia, Marrugo Llorente, and Saray Ricardo 2011);
- Working (evidence for secondary schools: Bonilla Mejia and Galvis 2012; Manzano Lopez and Ramirez Zambrano 2012 [leads to higher drop-out]); and
- Ethnicity (Bonilla Mejia and Galvis 2012 [for secondary schools]).

While there may be good reasons to include a separate level for family-level variables, this is practically not feasible for the context of Colombia's EN because of data limitations: The data does not allow for an identification of family ties between individual test takers. This is not problematic if the portion of siblings in the sample is not high, and is in fact common practice in multilevel models of educational outcomes (see, e.g., Somers, McEwan, and Willms 2004; Huang and Moon 2009; McArdle, Paskus, and Boker 2013).

Going up from the student-level, possible next levels are classrooms or schools. There are both theoretical and practical reasons for not including the classroom as a separate level. One feature of the EN model is that it is a multigrade school. Students of different age groups are being taught in the same class, both as a matter of pedagogical belief and in response to resource constraints, as rural schools are oftentimes too small to operate efficiently as monograde schools. Therefore, the distinction between the classroom and the school-level is not clear. Additionally, even if a classroom-level would be justified on a theoretical level, the available data again makes it impossible to include it, as the database of the standardized test SABER does not assign students to classrooms (or include classroom characteristics).

Instead, the second level of analysis will be the school-level.[10] This level contains the main explanatory variable, namely, an identifier for EN schools. As discussed, the operationalization of this identifier is one of the key questions of this research project, because the level of program implementation is uncertain. For the first part of the quantitative analysis (the country-level study), the official EN classifier is used to identify the school model. For the second part of the quantitative analysis (the department-level study), the EN implementation index is used to identify the school model.

In addition, level two contains a range of other school-level factors. The literature suggests that important school-level determinants of learning outcomes include:

- Geographic area, i.e. a rural or urban environment (evidence for primary schools: World Bank 2009; Zambrano Jurado 2013; evidence for secondary schools: Jola S. 2011; Bonilla Mejia and Galvis 2012; Carcamo Vergara and Mola Avila 2012);
- School type, i.e. public or private (Gaviria and Barrientos Marín 2001; Nuñez et al. 2002; World Bank 2009; Zambrano Jurado 2013; Bonilla Mejia and Galvis 2012 [for secondary education]);

[10] Schools in Colombia are administratively organized in three different levels: instituciones educativas, sedes, and jornadas escolares. An institucion educativa ("educational institution") may have one or more sedes ("branches"), which in turn may offer one or more jornadas escolares ("sessions": classes either meet for morning, afternoon, evening, full day, or alternative [e.g. weekend] sessions). While it may be desirable to treat the session-branch-institution hierarchy as three separate levels of analysis, this is not possible due to data constraints (most clusters would only include one or very few sub-groups or observations). Therefore, it is necessary to choose one institutional level as the "school" cluster. While branches are administratively dependent on their mother institutions, they may have very different characteristics: the mother institution may be in an urban setting, while the branch is in a rural setting; the former may offer a conventional curriculum, while the latter may implement the Escuela Nueva model; they may have different socioeconomic classifications; and so on. However, the different sessions offered by the same branch show much less variation. Most importantly, they typically do not differ in terms of the educational model offered. For these reasons, branches are treated as the main unit of analysis for this study, and the term "school" usually refers to branches, unless otherwise stated.

- School schedule, i.e. full day of half-day session (Bonilla Mejia and Galvis 2012 [for secondary schools]);
- School resources and infrastructure (evidence for primary schools: Casassus et al. 2000; Rangel and Lleras 2010; Zambrano Jurado 2013; evidence for secondary schools: Jola S. 2011; Manzano Lopez and Ramirez Zambrano 2012);
- Good classroom environment (Casassus et al. 2000);
- Socioeconomic composition of the school (Rangel and Lleras 2010);
- Low student-teacher ratio (Casassus et al. 2000; Gaviria and Barrientos Marín 2001 [for private secondary schools]);
- Offering science and vocational activities (Jola S. 2011 [for secondary schools]);
- Tutoring (Manzano Lopez and Ramirez Zambrano 2012 [for secondary education]); and
- Parental involvement (Casassus et al. 2000).

Furthermore, with regard to teachers:

- Teacher training level (evidence for primary schools: Casassus et al. 2000; World Bank 2009; Zambrano Jurado 2013; evidence for secondary schools: Gaviria and Barrientos Marín 2001 [only in private schools]; Jola S. 2011; Cepeda-Cuervo and Núñez-Antón 2013; Bonilla Mejia and Galvis 2012); and
- Good incentives and autonomy for teachers (Casassus et al. 2000; Nuñez et al. 2002; World Bank 2009; Gaviria and Barrientos Marín 2001 [for secondary schools]).

Latent teacher characteristics such as motivation, charisma, and other personality traits also enter this level of analysis through the school-level error term. It is likely that these unobserved teacher characteristics account for a substantial part of unexplained between-school variability, both with regard to EN implementation and with regard to school outcomes.

Schools for their part are nested within communities or regions, and there are good reasons to believe that, say, students from the capital city of Bogotá are more similar to each other than to students from isolated areas of Santander or Bolivar. In fact, studies by Mina (2004) and Bonilla Mejia and Galvis (2012) found that municipality-level variables such as expenditure per student, poverty, income inequality level, geographic location, and homicide rates have an impact on learning outcomes. Additionally, the political autonomy that municipalities enjoy, especially with regard to education policy, means that municipality-level variables may well have an effect on learning outcomes. Thus, municipalities represent the third level of the analysis.

The inclusion of a municipality-level alone does not do justice to the political and geographical structure of Colombia. Municipalities belong to departments, which are, together with the national government, the main level of policy making, and which often have distinct cultures. For instance, Secretaries of Education in the different departments have sovereignty over educational budgets. Therefore, departments should be considered as a fourth level of analysis.

Put together, four levels seem relevant for analyzing the effect of EN: the student-level, the school-level, the municipality-level, and the department-level. Of course, for the department-level analysis (chapter 6), only the first three levels are relevant. This structure is depicted in the conceptual model in Figure 5. The exemplary multilevel model presented in the previous section only included two levels, yet the inclusion of a third and fourth level follows a very similar logic (Raudenbush and Bryk 2002, 228–51). To be sure, an additional level adds complexity and a wide range of modeling possibilities, as the relationships of each higher-level variable to each of the lower-level variables have to be modeled. Each of the regression coefficients in the level-one model can be viewed as either fixed, non-randomly varying, or random; the same is true for each higher-level coefficient. Hence, in the context of EN evaluation using a three- or four-level model, one needs to decide for each level which predictors are being introduced (see below), and whether the structural effects related to these predictors (that is, the intercepts and slopes) are considered fixed, non-randomly varying, or random.

A final decision regarding the inclusion or exclusion of specific levels of analysis can only be taken after a first round of data analysis. It is good practice in multilevel modeling to start with an unconditional model, that is, with a model that includes only random effects at the different levels. With this so-called null model, it is possible to analyze the error term by decomposing the variance into its between-students/within-school, between-schools/within-municipalities, between-municipalities/within-departments, and between-departments components. Only if there is variance at each of these levels does it make sense to subsequently include varying or random effects at these levels. Hence, while the following discussion about predictors and the modeling of structural effects provides a model to be tested, the decision about the final model has to be taken based on the data.

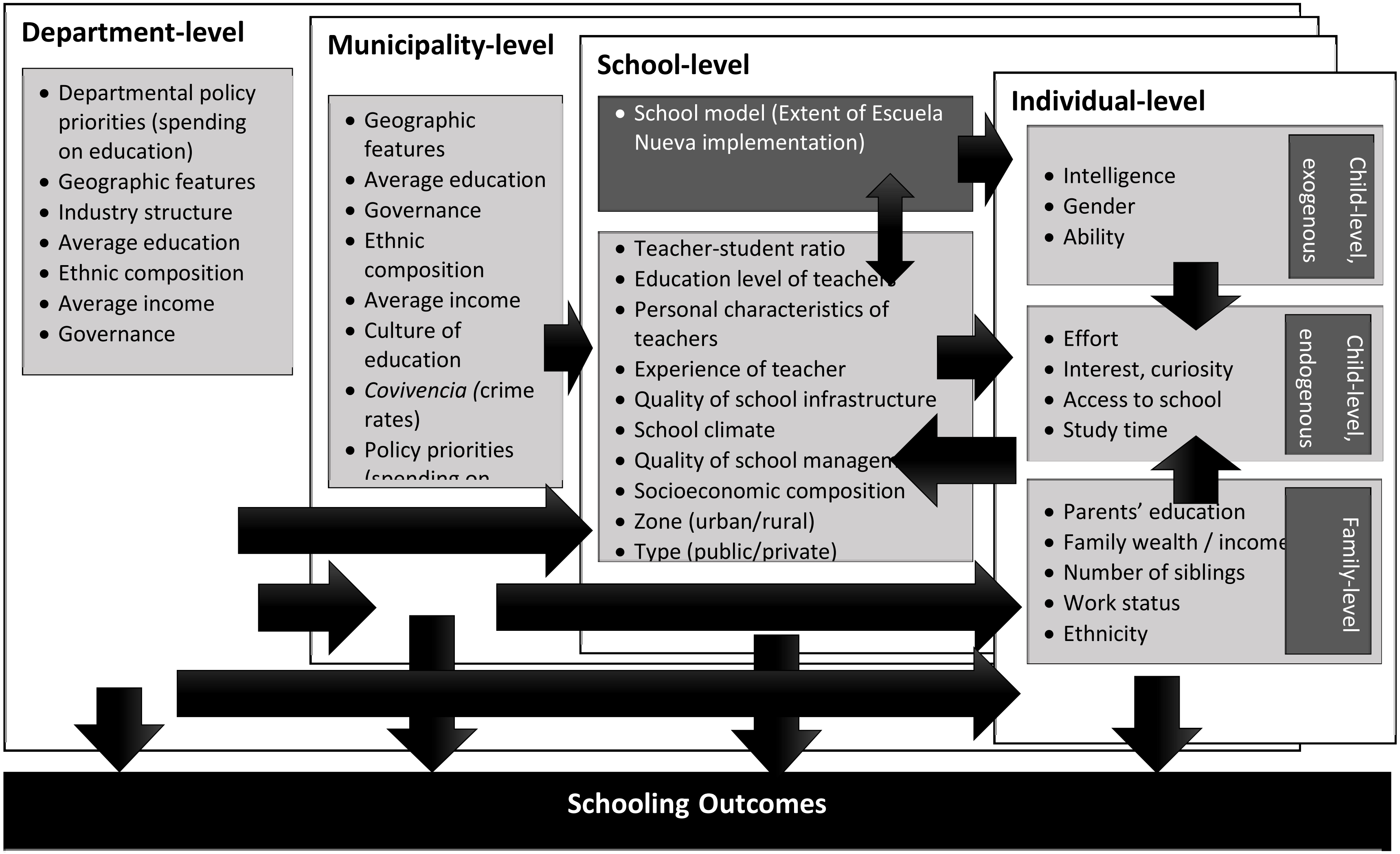


Figure 5 Conceptual Framework

The conceptual model depicted in Figure 5 summarizes the factors that influence schooling outcomes on the four levels as defined in the previous paragraphs. As becomes clear when looking at the model, many of the factors are interrelated. For instance, variables on the individual-level do not only influence learning outcomes, but may also impact one another: a student's curiosity will impact learning outcomes, but will also drive other factors such as study time; in turn, it will be influenced by other individual-level characteristics such as parents' education, and so on. Furthermore, variables can also have an impact on variables at another level. For instance, higher levels of parents' education and higher family wealth are likely to be associated with better school-level characteristics, while school-level characteristics such as the teacher-student ratio will impact individual-level characteristics such as effort, and so on. Unfortunately, for many of the variables outlined in the conceptual model no data is available. For instance, the available dataset does not contain a measure for students' intelligence, school climate, the cultural value of education, or departmental politics (on levels one, two, three, and four respectively). Hence, these variables will be omitted and become part of the error term, which demonstrate why it is so important to consider the structure of the composite error term.

While including all of the variables in Figure 5 would certainly be desirable, data could be obtained and merged into one dataset for the variables listed below. As described in the next section, the final decision about the inclusion of each of the variables into the model will depend both on the theoretical considerations presented above and on evidence for their relevance found in the data. The variables and their sources will be described in more detail in section 3.3 (Data) and in the context of the analysis (chapter 4 for the country-level analysis and chapter 6 for the department-level analysis).

Level-one (student-level) variables:

Dependent variable: Academic outcomes

- Test score (in Mathematics, Language, and Civic Competencies, respectively) ($score_{ijmd}$)

Independent variables: Student characteristics[11]

- Grade ($grade_{ijmd}$)
- Gender ($male_{ijmd}$)

Level-tow (school-level) variables:

- Escuela Nueva (EN_{jmd})
- Socioeconomic level of school (NSE_{jmd})
- Zone (urban or rural) ($rural_{jmd}$)
- School type (public or private) ($private_{jmd}$)
- Session type (full day, morning, afternoon ($morning_{jmd}$ and $afternoon_{jmd}$)
- Presence of students of ethnic background ($ethnic_{jmd}$)
- Presence of students who are conflict victims ($conflict_{jmd}$)

Level-three (municipality-level) variables:

- Governance ($governance_{md}$)
- Crime rate ($homicides_{md}$)

[11] The dataset for earlier rounds of the Pruebas SABER contained more student-level characteristics, such as ethnicity, parents' education, and a proxy for socioeconomic background. This data is not available anymore starting with the 2013 round. However, the recent rounds provide better quality data in other aspects.

- Education expenditure per student ($expenditures\_m_{md}$)

Level-four (department-level) variables:

- GDP per capita (GDP_d)
- Education expenditure per student ($expenditures\_d_d$)

3.2.1.6 Development of the multilevel model

Much more than for other econometric techniques, the formulation of a multilevel model is a highly iterative process. The correct specification of the structural effects, including the decision about the nature of these effects (fixed, non-randomly varying, random), needs to be based both on theory and on available empirical evidence. *Overfitting*, which means including more explanatory variables than necessary, especially at higher levels, is highly inefficient, as a multilevel model can quickly become very complex and difficult to compute (Snijders and Bosker 2011). For this reason, at this point only a preliminary model is presented that is based on the hypotheses developed in chapter 2 and on the conceptual framework presented in Figure 5. The actual model will be developed throughout chapter 4 (for the country-level analysis) and chapter 6 (for the department-level analysis).

Model A1, the preliminary model for the country-level analysis, looks as follows:

Model A1: $$score_{ijmd} = \beta_0 + \beta_1 EN_{jmd} + \beta_2 male_{ijmd} + \beta_3 (male * EN)_{ijmd} + \beta_4 rural_{jmd} + \beta_5 private_{jmd} + \beta_6 NSE_{jmd} + \beta_7 (NSE * EN)_{jmd} + \beta_8 ethnic_{jmd} + \beta_9 conflict_{jmd} + \beta_{10} morning_{jmd} + \beta_{11} afternoon_{jmd} + \beta_{12} governance_{md} + \beta_{13} homicides_{md} + \beta_{14} expenditures\_m_{md} + \beta_{15} GDP_d + \beta_{16} expenditures\_d_d + \zeta_{1d} EN_{jmd} + \zeta_{2md} EN_{jmd} + \zeta_{3d} + \zeta_{4md} + \zeta_{5jmd} + \varepsilon_{ijmd}$$

The subscripts indicate the level to which the variables and parameters belong: d for departments, m for municipalities, j for schools, and i for individual students. The model contains three random intercepts (ζ_{3d}, ζ_{4m} , and ζ_{5jmd} for the department-, municipality-, and school-level, respectively), as well as predictor variables for all four levels (with the fixed coefficients β_1 to β_{16}). There are two random coefficients in the model: a department-level (ζ_{1d}) and a municipality-level (ζ_{2md}) random coefficient of EN, reflecting the assumption that the effect of the EN model might be different between municipalities and departments, depending on the local context. This idea is further discussed in chapter 4. There are also two interaction terms: $\beta_3(male * EN)_{ijmd}$ and $\beta_7(NSE * EN)_{jmd}$. These interaction terms test the hypotheses that the effect of EN differs by gender, and that the model is particularly beneficial for children from lower socioeconomic levels.

Model A2, the preliminary model for the department-level analysis, is defined as:

Model A2: $$score_{ijm} = \beta_0 + \beta_1 EN_{jm} + \beta_2 male_{ijm} + \beta_3(male * EN)_{ijm} + \beta_4 private_{jm} + \beta_5 NSE_{jm} + \beta_6(NSE * EN)_{jm} + \beta_7 ethnic_{jm} + \beta_8 conflict_{jm} + \beta_9 morning_{jm} + \beta_{10} afternoon_{jm} + \beta_{13} governance_m + \beta_{14} homicides_m + \beta_{15} expenditures_{m_m} + \zeta_{1m} EN_{jm} + \zeta_{2m} + \zeta_{3jm} + \varepsilon_{ijm}$$

This model only contains two random intercepts ($\zeta_{2m}$and ζ_{3jm} for the municipality- and school-level, respectively), and one random coefficient (ζ_{1m}, which tests for municipality-specific effects of the EN implementation index). As all observations are from the same department, no department-level variables are included, and the subscript d is dropped.

The observant reader may have noticed that the student-level variable $grade_{ijmd}$ is not included in either of these models. The reason is that the analysis is done separately for the different grade areas, in order to allow for all coefficients (fixed and random) to differ by grade. A comparison of the estimation results for grades 3 and 5 will nevertheless allow one to draw conclusions about a “time of exposure” effect of the EN model: If EN indeed improves learning outcomes, one can expect the effect for 5th graders to be larger than for 3rd graders, as these students have been exposed to the program for more time. The EN model only starts at the 2nd grade level, that is, after children have obtained basic reading skills. This implies that at the time of taking the first Pruebas SABER test, 3rd graders have only been exposed to the EN method for about a year and a half, potentially too little time for robust effects. Thus, the comparison between the two age cohorts can be used as an additional estimator for the program effect size. In order to better understand how the influence of other factors on learning outcomes changes over the course of a primary school career, the two grade models are estimated separately (an inclusion of an interaction term for every random effect might quickly overload the model).

The development of the final model is done in five distinct stages. First, an extensive exploratory data analysis (EDA) is carried out (section 4.1 for the country-level analysis, and section 6.1 for the department-level analysis). The purpose of the EDA is to explore the variance of all variables and to check for correlations between the control variables and learning outcomes. This will help guide the decision about which variables will enter into the model.

The second step is a series of analysis of variance (ANOVA) models of the outcome variables, also called *null models* or *unconditional models* (sections 4.2 and 6.2). As laid out by Raudenbush and Bryk (2002), this analysis decomposes the overall variance in test scores into level-specific components and thus helps to understand at which levels (students, schools, municipalities, or departments) the factors lie that are responsible for differences in learning outcomes. At the same time, through a comparison of different specifications of the null model it becomes possible to decide which levels should enter as random intercepts into the model. As part of that analysis, intraclass correlation coefficients (ICC) are calculated to measure cluster homogeneity. The ICC represents the proportion of total variance in the outcome that is due to between-group variance at a given level of analysis. In the absence of between-group variability, the value of the ICC is zero, and the use of multilevel estimation methods is not necessary. A positive ICC, however, indicates a lack of independence of the lower-level observations, which justifies the use of a hierarchical model (O'Connell and Reed 2012).

Third, a series of random intercept models are developed (sections 4.3.1 and 6.3). These models contain a separate error term for each level of the model, the number of the levels being dependent on the theoretic considerations presented in this chapter and on the results of the ANOVA. The random intercept model also contains the explanatory variables of all levels. The inclusion of the explanatory variables is based on the discussion in this chapter, on the results of the EDA, and on analyses of model fit.

Fourth, and based on the results of the random intercept model formulation, the model is expanded into a random coefficient model, i.e. random slopes are added where appropriate (section 4.3.2). Fifth and finally, model diagnostics are being done to evaluate the overall fit of the model and the agreement with modeling assumptions. Changes to the model are possible up until this very last stage.

3.2.1.7 A note on the comparison group

In its core, the quantitative approach tries to isolate the EN effect from all other influences on learning outcomes. This necessitates the definition of a comparison group, a school model against which the EN model is contrasted. Several challenges arise from this simple fact. The first set of challenges is the clear definition of a distinct school type that the EN can be contrasted with. The second set of challenges relates to the possible endogeneity in the model: Whether or not a school is an EN is most likely not independent from a range of context factors that also influence learning outcomes – which begs the question whether a clear comparison group can ever be defined.

3.2.1.7.1 Potential comparison groups

Two objections may be raised about comparing EN schools to “conventional schools”. First, as was pointed out in section 1.2, Colombian schools and local authorities can choose from a range of educational models, the EN approach being one of them. Hence, not every non-EN school is automatically a conventional school; for example, a public primary school may also be part of the Etnoeducación or the Aceleración del Aprendizaje program. Or, even more problematic for a large-scale quantitative analysis, schools might make use to a varying degree of their autonomy to adapt the curriculum and teaching methods in their own ways. Short of collecting data on the actual classroom practices in every school, there is little that can be done to eliminate the risk that allegedly “conventional” schools are actually using very innovative methods.

A second possible objection is that conventional *multigrade* schools, rather than *all* conventional schools, may be the better comparison group. As outlined, EN was created in the context of UNESCO’s strategy to promote multigrade classrooms in the 1960s. This strategy aimed at increasing access to education in rural areas where monograde classes are not viable economically. The EN model was just one response; there are multigrade schools in Colombia that have not adopted the EN model. Thus, this universe of all multigrade schools might be considered more relevant for studying the benefits of the EN approach. This objection is based on the proposition that monograde schools differ from multigrade schools in ways that cannot be easily measured (that is, in terms of quality, community attributes, or culture), which may bias the results. More specifically, if it is mostly the poorest, lowest-quality schools that are multigrade schools, then comparing EN to monograde schools—without controlling for it—would underestimate the effect of the school model. Unfortunately, there is no data available on whether Colombian schools are multi- or monograde.

The literature is not conclusive about the effect of multigrade teaching on learning outcomes, but points in a general direction: literature reviews find that, all other things being equal, learning outcomes in multigrade classrooms are comparable to monograde schools (Brown and Martin A.B. 1989; Veenman 1997; Mason and Burns 1997). Mason and Burns (1997) qualify their result by concluding that the assignment of better students or teachers to multigrade classrooms may have driven the overall result; after controlling for that bias, multigrade classes may have a slightly negative effect. However, especially in the context of developing countries this bias may well go in the opposite direction (worse or less qualified teachers may be more likely to work in rural multigrade schools). Second, as Mason and Burns themselves acknowledge in their review of the literature, any potential negative effect may be the result of a methodological flaw in study designs, namely, of mixing two types of multigrade schools: those with purposeful and those with non-purposeful assignment. In schools where multigrade teaching is used solely as a response to budget pressure, and without changing pedagogic concepts, learning outcomes tend to be worse. In schools were the decision to use multigrade teaching is taken deliberately as part of a larger pedagogical strategy, the method may actually be superior, *if* teachers are provided with adequate support and resources (Mason and Burns 1997; Veenman 1997).

Everything put together, there are some good reasons to compare the EN model to all non-EN schools. First and foremost, the model set out to redesign rural education, not to redesign multigrade education. The vision of EN is to improve learning outcomes by addressing some of the major obstacles faced by rural communities; teaching various grade levels in one class is just one aspect of the model. It is, for that matter, a purposeful decision that is part of a larger pedagogical package.

Second, a comparison to other multigrade schools requires knowledge about where these schools are. However, there is no official data on Colombian multigrade schools; they are not differentiated from other—monograde—rural schools in the educational databases of DANE. A sample frame could only be created by either making certain assumptions (e.g. by defining schools as muligrade if there is not at least one teacher per grade level), or by adding an additional survey round. The latter option is of course not feasible, and the former may be problematic if, for instance, only some classes are combined.

3.2.1.7.2 Becoming an Escuela Nueva: Endogeneity

Even if a clear comparison group could be established against which official EN schools can be contrasted, another problem would remain, and that is the question of how and why a school becomes an EN school. As discussed in section 1.2 not much is known about the process, and no research has been done to explore which schools adopt the EN model faithfully, and why. It is very likely that the factors influencing implementation *also* influence learning outcomes: particularly motivated teachers may be more likely to adopt the model, highly committed parents may be more likely to support the model, and engaged politicians may be more likely to provide the necessary funding. All of these factors would probably also improve learning outcomes in the absence of the EN model. Unfortunately, the ways of dealing with this problem are limited. The fact that the data are analyzed with multilevel models helps to account for some of the noise surrounding individual schools (through a school-specific error term), but in the absence of a true experiment (where some schools, teachers, and students are randomly assigned to the EN model and others are not), it is not possible to compare the outcomes of one school that adopts the EN model with the outcomes of another non-EN school that is exactly the same in all unobservable factors. For the estimation results this means that the effect of EN might be confounded with the effect of more favorable background characteristics, resulting in an upward bias.

3.2.1.8 Estimation procedure

3.2.1.8.1 Software

The models are estimated using Stata (versions 12 and 13). The main command used for estimation is `-xtmixed-` (`-mixed-` in Stata13). While there are other ways of estimating multilevel models, `-xtmixed-` turned out to be most flexible for the purposes of this study, allowing for several levels of hierarchy. Generally, maximum likelihood estimation is used to fit the models (by specifying the `-mle-` option). However, if models failed to converge, the options `-reml-` (restricted maximum likelihood) or `-matlog-` (use of matrix logarithms for the estimation of variance components) were also attempted. The Stata estimation approach is based on the reference book on multilevel modeling using Stata by Rabe-Hesketh and Skrondal (2012), and on the University of Bristol Centre for Multilevel Modelling's online course on the subject (University of Bristol 2013).

3.2.1.8.2 Plausible values

ICFES provides he test results for the Pruebas SABER as plausible values (see the extensive description of this database below and in the annex), which has become common practice in large-scale educational assessments. As van Davier, Gonzales, and Mislevy (2009) show, plausible values are more reliable for group-level analyses of the latent variable student ability than the simple score from an exam, as a student's long-term proficiency is not necessarily reflected by her performance on a single test. Thus, measuring individual proficiency in a large-scale assessment is typically only achieved with a substantial amount of measurement error. In order to obtain more accurate estimates of students' true ability as well as of the variance therein—and thus more accurate estimates of the effect of specific educational interventions—it is better to work with a range of plausible values instead. Plausible values are a set of imputed values that reflect the likely distribution of a student's "true" ability. They are randomly drawn from a Bayesian posterior distribution, which is derived from the empirically observed exam scores and background characteristics. In this set-up, the "true" exam scores are missing by design, which makes plausible values a special case of an imputed dataset (Snijders and Bosker 2011, 136). The Pruebas SABER dataset contains five plausible values for each student and test area.

The correct way of using plausible values for estimations is to first estimate the models separately using each of the five plausible values, and then average the results using a set of imputation rules. It is *not* correct to take an average of the plausible values and run the estimates based on that average (von Davier, Gonzales, and Mislevy 2009; Snijders and Bosker 2011, 135–44). If one takes first the average of the plausible values, or picks only one of the five values to run the estimates, the estimation results are biased, particularly with regard to estimated variances (Carstens and Hastedt 2010; Marchant 2015).

STATA has a class of multiple imputation commands that correctly handle this two-step imputation process. The average estimates from the separately run models are obtained by first using `-mi import-` to declare that the specified variables are imputed, and then `-mi estimate-` as a prefix for the estimation commands. However, there are some shortcomings in using the multiple imputation commands. They are designed to provide point estimates and cannot fulfill all the requirements of this analysis – for instance, standard deviations for the exploratory data analysis, or many multilevel post-estimation tests, cannot be performed through multiple imputation. In these cases, five models (one for each plausible value) are run separately, and the results are presented jointly or as averages.

3.2.1.8.3 Weights

The analysis of complex survey data in a multilevel design is an emerging field, because the use of weights in maximum likelihoods estimation is more complex than the use of weights in OLS regression. This is particularly true for multilevel designs where sampling probabilities may differ from level to level. Stata added survey support for multilevel models only in its most recent release, Stata14. As the data for this dissertation were analyzed in Stata12 and Stata13, the respective commands were not available for the analysis. In general, not including sampling weights properly may lead to biased parameter estimates, while estimates are unbiased but less efficient when weights are included. In the context of maximum likelihood estimation, the increased complexity of the model due to sampling weights increases the chances that the model does not converge (a problem that did in fact arise in the process of analyzing the data). It is therefore important to carefully consider the use of weights.

The Pruebas SABER database contains student-level weights for each testing area. These weights result from the implicit sampling probabilities for each student and testing area within a class, as each student is tested only in a part of the overall exam. Each 3rd grade student participates only in either the language or the mathematics exam, while each 5th grade student participates only in two out of the three testing areas (mathematics, language, and civic competencies). The weights are therefore calculated as follows (Atención al Ciudadano del ICFES 2016):

$$w_{igja} = \frac{N_{gj}}{n_{gja}}$$

where w_{igja} is the weight of student i of grade g in school-session j in testing area a; N_{gj} is the total number of students in grade g in school-session j; and n_{gja} is the number of students in grade g in school-session j who participate in the test for area a. For instance, if a school has 20 students in grade 3 and 10 of them participate in the mathematics exam, the associated implicit sampling weight of each of these students is 2. The weights reported in the database vary between 1 and 6, indicating that each student represents between one and six students of his or her school-session and grade in the given testing area. This, in turn, implies that the weights should not vary for students within the same school-session and grade for a given testing area. Because all Colombian schools participate in the Pruebas SABER, the sampling probability for all higher levels is one by design.

Kim, Anderson, and Keller (2013, 414) explain that for multilevel analysis, the decision of whether or not to include sampling weights should be based on whether the weights are likely to have an impact on the results (whether they are, in fact, informative), and whether the probability of selection is related to the probability model for the data. There are no reasons to believe that the latter would be the case: all students in a given grade and school-session are randomly assigned to the different parts of the exam; this process is not based on any type of stratification or a similar non-fully-random processes. Therefore, there should be no systematic relationship between inclusion probability for a given testing area and the resulting test score.

The former consideration (whether weights are informative) requires some more analysis. Theoretically, the Pruebas SABER weights should be completely uninformative for a multilevel analysis, given that all students in a cluster (any given school-session/grade/area combination) should have the same sampling probability. In the actual dataset, there is a small number of schools with variations in the reported sampling weight for students in the same cluster (between 0.4% and 0.5% of all schools, depending on the testing area). No explanation for this variation could be obtained from ICFES. Most likely, the variation is due to errors in the database. However, even if that is not the case, the low within-school variance suggests that the student-level weights are not informative and will likely have no impact on the results. Therefore, weights will not be included in the multilevel analysis, following the recommendation of Kim, Anderson, and Keller (2013) and Carle (2009). They will, of course, be included in the survey data analysis.

3.2.2 Qualitative analysis

There is very little discussion in the EN literature about the process of program implementation, especially about the motives for adapting the school model or its components in a given school or municipality, or the reasons why implementation is so irregular across schools. The last part of the research project is a small qualitative component that aims to explore these questions by collecting in-depth information on program implementation and on perceived benefits of the program in a few selected schools and from some key stakeholders and experts. Specifically, the aims of this component are as follows:

- Gain a better understanding of the EN methodology by observing typical school days in EN and non-EN schools
- Obtain a richer, in-depth assessment of the program's perceived benefits and shortcomings from its primary stakeholders, that is, children and teachers
- Explore how the different program components are being implemented, used, and potentially altered in practice
- Explore reasons for different degrees of program implementation and for decision in favor or against the school model
- Gain a better understanding on the decision-making processes related to EN program implementation
- Explore perceptions held about the program by key stakeholders, including beliefs related to the functioning and success of the approach for students from different backgrounds

3.2.2.1 Data collection

The qualitative exploratory study is based on visits to seven rural schools in Quindío. The schools were chosen purposefully after a preliminary analysis of the quantitative data: in order to be able to observe different classroom practices, three schools with particularly high implementation scores and four schools with particularly low implementation scores were selected and revisited for additional data collection. The field work team who gathered the quantitative data provided consulting for the selection, with the aim of obtaining a diverse sample of schools in many regards (size, remoteness, teaching style, student population, etc.).

Data was collecting in the following ways:

- Semi-structured interviews with 5th grade teachers. These interviews focused on the teacher's role in the classroom, and his or her perceptions on the school model.
- Focus groups with 5th grade students to obtain an idea on students' perceptions of their respective schools. If the school size permitted it, two groups were organized separated by gender in order to allow boys and girls to express themselves more freely. However, most schools were very small, often with less than fifteen children in all primary grades combined.
- Non-participatory observations to document everyday life at school. During the school visits, the set-up of the school and the classroom, the way students and teachers interact, the use of specific pedagogical elements, etc. were observed.

Additionally, interviews were conducted with key staff of Fundación Escuela Nueva, and with the person responsible for the EN model within the Secretary of Education in Quindío.

3.2.2.2 Data analysis

The qualitative data analysis was guided by the attempt to identify which factors explain the process of EN implementation, to gain a better insight into everyday life at school, and to uncover benefits and shortcomings of the model that may have been overlooked so far. This was done through qualitative content analysis, a commonly used strategy in social sciences research as a way to discover general patterns and meanings from responses (Rubin and Babbie 2007). The first step was the transcription of the interviews. In a next step, the transcriptions and observation protocols were coded using descriptive, topic, and analytical coding (Richards 2005). The codes for this process were derived from the conceptual framework, research questions, and the theoretical and empirical literature upon which this project is based; in further iterations, codes were induced from the thematic patterns emerging in the data. Based on the topics that recurred in this analysis, hypotheses about the research questions were formulated. Finally, the interview transcriptions and observation protocols were searched for further instances that either support and refine or contradict these hypotheses (Rubin and Babbie 2007). This coding process, and the related analysis, was done using UCLA's mixed methods web application *dedoose*.

It is important to reiterate that this design element does not allow drawing conclusions about causalities. It is employed in this study as a way to provide a richer context for the quantitative analysis, and to generate hypotheses about the parts of the research question that the quantitative section cannot address. In that sense, the qualitative information collected through the research project helps strengthening the conclusions of the project.

3.3 Data

Several datasets are merged for this study. The three main datasets are:

1. The results of the Pruebas SABER standardized test, provided by ICFES;
2. Administrative school data, including the official school model classification, provided by DANE in the EDUC dataset;
3. Data on the implementation of the EN school model, collected in Quindío for the purpose of this study.

Furthermore, some additional municipality or department-level data were added from other sources. This section gives a brief overview of these datasets and describes the final (joint) dataset. More information on the individual datasets can be found in Annex A.

3.3.1 The Pruebas SABER dataset

The main dataset for this study is the dataset containing the results of the 2013 round of the Pruebas SABER. The Pruebas SABER 3º, 5º Y 9º (henceforward, Pruebas SABER) is a standardized evaluation of learning outcomes that has been carried out in Colombia since 1991. Participation is mandatory since 2001 for both private and public schools. In 2012, 3rd graders were added to the test. Since the same year, the tests have been carried out yearly, and every round of evaluation contains a mathematics and a language exam; additionally, alternating each year there is also an evaluation in natural sciences or civic competencies for students in grades 5 and 9. Each 3rd grade student is randomly assigned to either the language and mathematics test, while each 5th and 9th grader is assigned randomly to two of the three test areas of the given year (ICFES 2015b).

For the 2013 round, results are available for 704,697 3rd graders and for 706,204 5th graders (1,410,901 in total), as well as for 17,073 educational institutions and 31,050 uniquely identifiable branches. When only looking at students in uniquely identifiable branches, data is only available for 567,939 3rd graders and for 574,948 5th graders (1,142,887 in total) in 14,729 educational institutions with 31,050 branches. This discrepancy in the numbers is due to reporting errors: for schools that did not report results correctly or reported significant cheating, ICFES provides test scores only at the institution level, and omits all branch-level information (including the branch-level identifiers) from the database. Because the school model is defined at the branch-level, observations without a unique school branch identifier have to be removed from the dataset, which may lead to a sample selection bias in the results if the likelihood of exclusion is correlated with learning outcomes and with the likelihood of being an EN school. This possible bias is discussed in detail in section 1.2 of Annex A. Unfortunately, it is not possible to fully determine whether a bias exists, and how it would influence the results. However, a careful judgement can be made based on the available evidence: Most likely, reporting results correctly is an indicator for overall school quality or administrative capacity, which likely is correlated positively with higher test scores. Furthermore, it is likely that schools that implement more EN elements tend to be higher-quality schools. These correlations suggest a positive bias in the estimated effect of the EN model. Furthermore, given a likely positive correlation between average socioeconomic level and school quality, the main effect of socioeconomic status as well as its interactions with the EN model are probably overestimated as well.

The following variables come from the Pruebas SABER database:

- Test results
- Gender
- Grade
- Socioeconomic level of school (average socioeconomic level of students)
- Zone (urban or rural)
- Sector (public or private)
- Session type (full day, morning, afternoon)

3.3.2 The administrative dataset EDUC (C-600)

Official administrative data, including data on the educational model used in each school and school branch, is provided by DANE in the EDUC dataset. All Colombian public and private school are required by law to respond to the “C-600 survey” every year, which is why the dataset covers most schools. Educational institutions and school branches are identified by the same unique codes that are also used in the Pruebas SABER dataset (and for other education-related purposes). This study uses data from 2013, which is the year of the Pruebas SABER results. According to this database, there are 49,932 school branches in Colombia which offer primary education.

The following variables come from the C-600 dataset (see Annex A for more information):

- Geographic information: department, municipality, urban or rural area
- Sector: private or public (for robustness analysis)
- Zone: rural or urban (for robustness analysis)
- Educational offers: education level and school model
- Student population: Number of students in current and previous year, presence of students from ethnic minorities or students who are victims of the armed conflict

3.3.3 Primary data: Escuela Nueva implementation

The third large component of the study dataset is primary data collected in order measure the extent of EN program implementation in Quindío, a department of Colombia. Quindío is well suited for an implementation study because the model has been implemented in this department for three decades, and continues to be (officially) supported by the Secretary of Education of Quindío. It therefore was expected that a sample of primary schools from this department would reveal a range of implementation states, from overly conventional schools to schools with considerable experience and diligence in EN implementation. Additionally, Quindío is a small department, which means that it was possible to obtain a representative sample of schools with high statistical power with limited project funds. Even though the department is mountainous and some of the schools are located in remote areas of the Andes, average distances to the schools are moderate. The department is also considered one of Colombia's safest areas. All these characteristics mean that Quindío is not necessarily representative for other regions of Colombia, hence the results from the department-level analysis might not be directly transferable to other parts of the country.

3.3.3.1 Sample

Program implementation is measured at the branch-level, hence the sampling units for this project were school branches. The 2013 application of the Pruebas SABER was used as the sampling frame. From this list, urban schools and schools that only offer (or only report results for) one of the two grades of interest were eliminated. Of the remaining 149 schools, 80 schools[12] where randomly drawn using the `-sample-` command in Stata. The numbers are summarized in Table 2.

Within the sampled schools, all 5th grade students and their teachers were included in the sample. In case that a school had more than one class of grade 5, one class was to be randomly picked. In case that a class had more than fifteen 5th grade students, fifteen students were to be randomly picked. In order to be included in the sample, students and their parents had to sign consent/assent forms, and the students had to be present in school on the day of the data collection. No attempt was made to follow up on 5th grade students who normally attend a sampled school but were not present on the day of the data collection.

[12] According to DANE, 89.5% of rural schools in Quindío are implementing the Escuela Nueva model. The standard formula to calculate the required sample size n for obtaining a sample in which a percentage p has a characteristic of interest is as follows: $n = m/(1 + n/N); m = 1.96\text{^}2 * (p * (1 - p))/CI\text{^}2$, where N is the Population size and CI the confidence interval expressed as a decimal (for ±5, CI = 0.05). If one wants to get a sample of rural schools of Quindío in which, with a confidence interval of 0.05, 90% of schools implement the model, m is 138.297 and n is 71.73. This number was rounded up to 80 schools.

Table 2: Overview of sampling process

	Number of branches	Of which eliminated
Total	31,050	30,760
Of which in Quindío	290	109
Of which in rural areas	181	32
Of which with results for grades 3 and 5	149	69
Of which in sample	80	

3.3.3.2 Survey instrument and data collection

The survey instruments are designed to measure the degree to which EN principles are implemented in schools. They were initially developed and field-tested by Fundación Escuela Nueva, but were slightly adapted for this study. The instruments can found in Annex D.

There are two different survey forms: One for teachers and one for students. Each version of the survey contains some background information about the school (such as size and location) and about the student or teacher, respectively. The main part of the survey collects data on school life: Classroom and course organization, implementation of different pedagogic techniques, use of text books and supportive materials, student government, community relations, and EN teacher training and class preparation. Administering surveys to both students and teachers allows for capturing different aspects of program implementation and serves to triangulate the data, given that the reported classroom experience of students and teachers may well differ.

The surveys were administered in the classrooms during normal school hours. A team of four field workers visited the sampled schools between May and November 2016. The field workers were recruited with the help of an associate professor from the Universidad del Quindío. They are graduate students in Education, and all of them have experience working as teachers in rural Quindío. After obtaining permission from the director of each educational institution in the sample, each school (branch) was visited at least twice: one time in order to invite students and teachers to participate and to drop off assent/consent forms, and a second time to administer the survey instruments. Field workers also documented some meta-data about the school. Students and teachers received a compensation for their participation.

As is to be expected in data collection field work, some challenges arose – some of which provided valuable qualitative background information on the workings of Colombia's primary school sector. For instance, a significant number of schools had to be visited more than the scheduled two times, because teachers were absent, school was cancelled due to bad weather or due to the UEFA soccer championship, or the official school schedule had changed in the short time between first and second visit. The field work team was instructed to follow up with every school up to five times. There was also one case in which the school could not be located. In ten cases a sampled school had to be discarded because no data collection was possible (refusal of the teacher or the parents, or repeated failure to successfully distribute the survey instruments). In eight of these cases, the schools were replaced by a randomly selected alternative school from the same municipality. Some of the challenges encountered during the data collection process will also be discussed in the context of research limitations (section 7.2).

3.3.3.3 The implementation data base

Program implementation data is available from 81 teachers in 76 schools and from 260 students in 68 schools. There are 2 schools with data only from students, and 10 schools with data only from teachers. The student database contains around 70 variables measuring the different parts of program implementation, plus some information about the students. The teacher database contains over 140 variables measuring program implementation, plus some information about the teacher. Most of the variables are yes/no dummy variables or Likert-type scales. To protect the privacy of the respondents, the schools and respondents are identified by a project-specific ID, which connects to a meta-database that includes the official DANE ID of the school.

3.3.4 Other data

Some additional variables from other data sources complement the analysis. Unfortunately, the official statistics system in Colombia is relatively scattered, and municipality-level statistics are rarely available. Data are thus chosen based on availability, even if this means that the data may not be perfectly comparable. The following variables were added (for details, consult the annex):

- Homicide rates by municipality, provided by Federación Colombiana de Municipios (2016)
- Departmental gross domestic product, provided by DANE (2016)
- Public education expenditure per student at municipality- and department-level, provided by Iregui, Melo, and Ramos (2006)
- Municipal governance index, provided by the National Planning Department (DNP 2014)

3.3.5 Description of the dataset

After merging the datasets and removing cases where the EDUC and the SABER dataset could not be matched, the final dataset (henceforward: study dataset, final dataset, or simply dataset) is structured as follows (see Table 3). The dataset contains observations from 1,100,397 students. About half of these students (49.7%) are in grade 3, the rest (50.3%) are in grade 5. The data comes from students from 30,099 branches and 14,371 institutions. For around 80% of branches and 89% of institutions, data is available for both grade levels. About 24% of student-level observations and 69% of branch-level observations come from rural areas, which shows that rural branches are typically much smaller than urban branches. Finally, 53% of institutions have at least one dependent branch that is located in a rural area. Table 4 summarizes the study dataset.

Table 3 Overview of study dataset: Observations by data level

	Total	In grade 3 / results for grade 3		In grade 5 / results for grade 5		Results available for both grades		Located in rural zone	
		N	%	N	%	N	%	N	%
Institutions	14,371	13,849	96.4%	13,248	92.2%	12,726	88.6%	7,579	52.7%
Branches	30,099	27,772	92.3%	26,264	87.3%	23,937	79.5%	20,732	68.9%
Students	1,100,397	546,450	49.7%	553,947	50.3%	--	--	267,273	24.3%

Table 4 Overview of study dataset: missing observations

Variable	Missing observations	Non-missing observations	Unique values
ID Student	0	1,100,397	1,100,397
ID Institution	0	1,100,397	14,371
ID Branch	0	1,100,397	30,099
ID Session	0	1,100,397	31,908
ID Department	0	1,100,397	33
ID Municipality	0	1,100,397	1,079
Area of branch (rural vs urban)	0	1,100,397	2
Area of institution (rural vs urban)	0	1,100,397	2
Sector (public vs private)	0	1,100,397	2
Session type	234,439	865,958	3
Socioeconomic level of institution/branch	49,542	1,050,855	4
School calendar	0	1,100,397	2
Grade	0	1,100,397	2
Sex	30,184	1,070,213	2
Plausible values Language	459,132	641,265	>500
Plausible values Mathematics	463,901	636,496	>500
Plausible values Civic competencies	733,449	366,948	>500
Model: Traditional	18,145	1,082,252	2
Model: Escuela Nueva	18,145	1,082,252	2
Students of ethnic minorities	18,145	1,082,252	2
Students who are conflict victims	18,145	1,082,252	2
Students primary, previous year	38,896	1,063,257	>500
Students primary, current year	27,752	1,072,645	>500
Homicides rate per 100,000 inhabitants	95,702	1,004,695	137
Governance Index	1,897	1,098,500	>500
Expenditures by student (department)	0	1,100,397	33
Expenditures by student (municipality)	635,982	464,415	223
GDP per capita	0	1,100,397	33
EN Index*	1,099,296	1,101	74
EN teacher index*	1,099,308	1,089	66
EN student index*	1,099,390	1,007	66

* The construction of the index is described in chapter 5.

4 Country-Level Analysis of Learning Outcomes

The Escuela Nueva school model has been officially implemented throughout Colombia for many years. Today, about half of Colombian primary schools and around three quarters of Colombian rural primary schools are officially EN. However, to the best of the author's knowledge, no country-level analysis of the effects of the model has been carried out, with the exception of a recent dissertation (Hincapié 2014). This chapter aims to fill that gap based on an analysis of the results of the 2013 round of the Pruebas SABER and the official EN classifier provided by DANE. Data for this exercise is available for a total of 810,324 students (393,212 in grade 3, 417,112 in grade 5) in 21,235 schools.

After exploring the data (section 4.1) and decomposing the variance in exam scores into the levels of its source (section 4.2), a department-level random coefficient model is developed that describes how the EN model influences learning outcomes, taking into account student-, school-, municipality-, and department-level effects (section 4.3). The results show a statistically and practically significant effect of the model, with some important variations across municipalities and departments (section 4.4).

4.1 Exploratory data analysis

Exploratory data analysis of the predictors helps to understand the data and the relationship between test scores and possible explanatory variables. This information is even more useful for multilevel models than for standard OLS analysis, given the higher complexity of the model.

4.1.1 Plausible value outcomes

Table 5 summarizes the sample mean and standard deviations for each grade-area for all five plausible values provided in the dataset, together with the imputed mean across the five plausible values and the imputed overall standard deviation.[13]

The table confirms that the five plausible values are congruent in the aggregate (despite the large variation of plausible values within persons: for example, the mean within-person standard deviation in the language score plausible values is 22.96 for grade 3 and 28.81 for grade 5; the magnitude is similar for the other areas). The table also shows that the imputed estimations generated by Stata mirror the descriptive statistics of the individual plausible values. Furthermore, Figure 6 and Figure 7 show for language exam scores in grade 3 and 5, respectively, that the distribution across persons is very similar between the five plausible values, and that it appears reasonable to assume an approximate normal distribution. This assumption is substantiated by the statistics on skewness and kurtosis presented in Table 5: even though there is a slight positive skew, the values are close to a normal distribution. The histograms for the other testing areas look similar and are not presented.

[13] The standard deviations are calculated “manually” as $\sqrt{E[pv^2] - (E[pv])^2}$, where $E[.]$ denotes the estimated imputed values (following Kolenikov 2010).

Table 5 Exploratory Data Analysis: Plausible Values in country-level study

	Language				**Mathematics**				**Civic Competencies**	
	Grade 3		**Grade 5**		**Grade 3**		**Grade5**		**Grade 5**	
	mean	**sd**	**mean**	**sd**	**mean**	**sd**	**mean**	**sd**	**mean**	**sd**
PV1	298.02	75.22	298.02	79.68	297.74	77.59	295.78	76.79	293.88	75.17
PV2	298.01	75.19	298.01	79.73	297.79	77.52	295.97	76.85	293.98	75.24
PV3	298.16	75.22	298.16	80.01	297.85	77.54	295.89	76.65	294.05	75.26
PV4	298.08	75.17	298.08	79.68	297.79	77.53	295.93	76.80	293.97	75.27
PV5	298.02	75.03	298.02	79.58	297.74	77.48	295.97	76.68	293.97	75.17
Imputed	298.06	75.16	299.48	79.74	297.78	77.53	295.91	76.75	293.97	75.22
Skewness*	0.413		0.319		0.460		0.406		0.493	
Kurtosis*	3.246		2.942		3.385		3.003		2.993	
N	197,234		277,179		195,978		274,404		276,169	

* These statistics are provided exemplarily for plausible value 1.

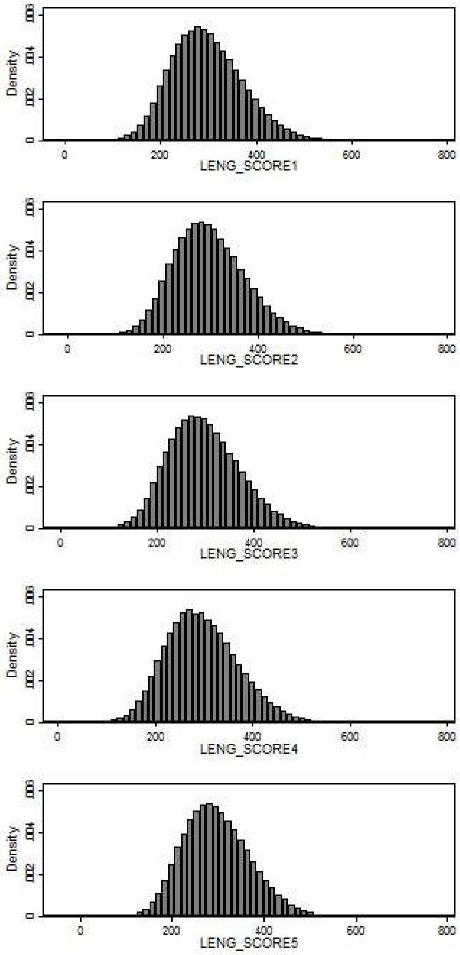


Figure 6 Histograms of Plausible Values for Exam scores, Language Grade 3 (country-level study)

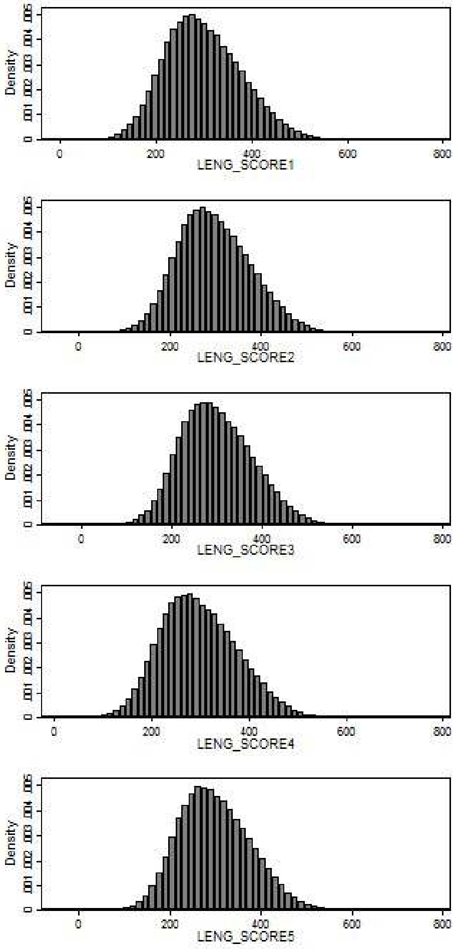


Figure 7 Histograms of Plausible Values for Exam scores, Language Grade 5 (country-level study)

4.1.2 Predictors

4.1.2.1 Student-level

Table 6 through Table 10 summarize the results of the exploratory data analysis: For each level of all the predictor variables, the tables contain the percentage of student-level observations in the given level, the average test score, and the standard deviation. The tables do not report the standard errors of the mean, though these were typically estimated to be well below 1, indicating a high level of confidence about the respective means. The information provided in the tables is purely descriptive. Unless otherwise stated, all estimates henceforward are based on the same group of observations, which are all observations for which there is non-missing information on the respective test scores, EN classification, gender, area, school type, socioeconomic level, ethnic students, conflict victims, session type, municipal governance, per capita departmental GDP, and education expenditure per student (department-level data).

Some clear patterns emerge. For most predictor variables and grade-areas, there are considerable differences in the mean scores in line with the expectations (except in the case of the homicide rate and municipality-level data on education expenditure per student, where no trend is apparent). For instance, students in urban schools score better, on average, than students in rural schools in all areas and grades; the higher the average socioeconomic level of students in a school, the better the average test results; and students do better the higher the per capita GDP of their department. Girls do better than boys in all areas except in mathematics (where there is no clear gender difference in grade 3; boys do better in grade 5). The standard deviations are generally estimated to be between 70 and 80, with only a few values below or above that range. This is a relatively high value in relation to the mean estimate and indicates a large variation in test scores across students within each of the groups.

Table 6 Exploratory Data Analysis, Language Grade 3 (country-level study)

Statistics presented at student-level		**%**	**Mean**	**Std. Dev.**
Level-one Variables (Student-level, n = 197,234 students)				
Male	Male	50.41	292.24	73.80
	Female	49.59	303.98	76.07
Level-two Variables (School-level, j = 17,652 schools)				
EN school	EN school	12.60	286.26	78.94
	Non-EN school	87.40	299.76	74.45
Session type	Complete day	10.75	309.10	80.48
	Morning	57.80	295.64	75.85
	Afternoon	31.45	298.74	71.59
Rural	Rural	24.74	279.40	77.09
	Urban	75.26	304.19	73.49
Private	Private	9.49	349.67	79.98
	Public	90.51	292.65	72.55
Socioeconomic level	NSE 1	21.69	269.37	76.01
	NSE 2	25.67	282.53	70.60
	NSE 3	31.56	303.10	67.67
	NSE 4	21.08	338.96	71.28
Ethnic population	With ethnic students	34.01	292.40	72.99
	Without ethnic students	65.99	300.98	76.09
Victims of conflict	Has conflict victim students	53.88	295.51	70.19
	Has no conflict victim students	46.12	301.04	80.49
Level-three variables (municipality-level, m = 1,007 municipalities)				
Homicide rate #	Lowest quartile (Q1)	37.02	301.51	74.45
	Q2	10.61	298.16	76.83
	Q3	24.74	292.06	74.02
	Highest quartile (Q4)	27.63	305.29	74.82
Governance index	Lowest quartile (Q1)	30.35	291.23	76.07
	Q2	19.50	287.47	74.00
	Q3	27.77	301.90	75.71
	Highest quartile (Q4)	22.38	311.79	71.72
Educ. expenditure per student, municipal-level data*	Lowest quartile (Q1)	26.93	300.20	75.02
	Q2	27.98	289.25	74.75
	Q3	37.29	314.64	72.15
	Highest quartile (Q4)	7.81	301.21	70.30
Level-four variables (department-level, d = 33 departments)				
GDP per capita	Lowest quartile (Q1)	26.75	284.32	76.12
	Q2	24.14	296.42	72.88
	Q3	25.18	300.83	76.67
	Highest quartile (Q4)	23.93	312.16	71.89
Educ. expenditure per student, department-level data	Lowest quartile (Q1)	26.41	283.40	76.58
	Q2	27.34	296.84	75.76
	Q3	25.75	304.64	73.59
	Highest quartile (Q4)	20.50	310.31	71.17

* Data for this variable is only available for n= 74,746 in j= 3,656, m= 193 and d= 10.

# Data for this variable is only available for n= 178,634 in j=14,176, m=686 and d=32.

Table 7 Exploratory Data Analysis, Language Grade 5 (country-level study)

Statistics presented at student-level		**%**	**Mean**	**Std. Dev.**
Level-one Variables (Student-level, n = 277,179 students)				
Male	Male	50.42	291.69	79.34
	Female	49.58	307.40	79.36
Level-two Variables (School-level, j = 17,586 schools)				
EN school	EN school	11.74	289.24	78.61
	Non-EN school	88.26	300.84	79.79
Session type	Complete day	10.78	314.72	83.26
	Morning	57.26	297.19	79.52
	Afternoon	31.96	298.44	78.33
Rural	Rural	23.81	281.89	77.65
	Urban	76.19	304.98	79.58
Private	Private	8.84	341.75	86.77
	Public	91.16	295.38	77.81
Socioeconomic level	NSE 1	21.25	270.38	75.48
	NSE 2	25.46	283.02	73.92
	NSE 3	31.75	302.95	74.13
	NSE 4	21.54	342.51	79.44
Ethnic population	With ethnic students	34.54	293.43	77.30
	Without ethnic students	65.46	302.67	80.81
Victims of conflict	Has conflict victim students	54.67	298.16	76.54
	Has no conflict victim students	45.33	301.07	83.40
Level-three variables (municipality-level, m = 1,011 municipalities)				
Homicide rate[#]	Lowest quartile (Q1)	37.85	304.93	79.12
	Q2	10.92	298.05	80.18
	Q3	24.55	294.37	79.23
	Highest quartile (Q4)	26.68	303.11	80.55
Governance index	Lowest quartile (Q1)	30.76	293.45	79.20
	Q2	19.35	287.85	77.42
	Q3	27.46	302.35	79.71
	Highest quartile (Q4)	22.44	314.26	79.98
Educ. expenditure per student, municipal-level data*	Lowest quartile (Q1)	25.31	298.64	81.55
	Q2	27.17	290.16	78.66
	Q3	39.89	314.04	78.38
	Highest quartile (Q4)	7.63	302.96	78.49
Level-four variables (department-level, d = 33 departments)				
GDP per capita	Lowest quartile (Q1)	26.39	283.96	77.78
	Q2	24.39	299.95	78.22
	Q3	24.05	297.99	81.38
	Highest quartile (Q4)	25.17	316.72	78.14
Educ. expenditure per student, department-level data	Lowest quartile (Q1)	26.05	281.45	79.16
	Q2	26.64	295.97	79.02
	Q3	25.67	308.88	79.27
	Highest quartile (Q4)	21.65	314.34	77.22

* Data for this variable is only available for n= 105,396 in j= 3,616, m= 191 and d= 10.

# Data for this variable is only available for n=250,813 in j=14,094, m=689 and d=32.

Table 8 Exploratory Data Analysis, Mathematics Grade 3 (country-level study)

Statistics presented at student-level		**%**	**Mean**	**Std. Dev.**
Level-one Variables (Student-level, n = 195,978 students)				
Male	Male	51.21	297.42	78.84
	Female	48.79	298.16	76.13
Level-two Variables (School-level, j = 17,475 schools)				
EN school	EN school	12.46	298.93	87.33
	Non-EN school	87.54	297.62	76.03
Session type	Complete day	10.76	313.97	84.15
	Morning	57.58	296.17	78.17
	Afternoon	31.66	295.22	73.26
Rural	Rural	24.51	287.28	83.87
	Urban	75.49	301.19	75.04
Private	Private	9.53	344.66	82.59
	Public	90.47	292.85	75.30
Socioeconomic level	NSE 1	21.57	276.49	83.26
	NSE 2	25.67	284.73	74.28
	NSE 3	31.53	299.60	70.02
	NSE 4	21.23	332.50	73.76
Ethnic population	With ethnic students	34.02	290.12	74.28
	Without ethnic students	65.98	301.74	78.86
Victims of conflict	Has conflict victim students	53.94	293.41	72.05
	Has no conflict victim students	46.06	302.90	83.20
Level-three variables (municipality-level, m = 1,009 municipalities)				
Homicide rate#	Lowest quartile (Q1)	37.11	301.49	76.85
	Q2	10.60	298.20	79.62
	Q3	24.64	291.88	76.40
	Highest quartile (Q4)	27.65	301.80	76.60
Governance index	Lowest quartile (Q1)	30.43	290.85	78.33
	Q2	19.45	290.27	77.93
	Q3	27.75	302.38	78.13
	Highest quartile (Q4)	22.37	308.04	73.66
Educ. expenditure per student, municipal-level data*	Lowest quartile (Q1)	26.95	292.74	74.59
	Q2	27.81	287.01	75.58
	Q3	37.41	309.94	74.29
	Highest quartile (Q4)	7.83	305.56	74.53
Level-four variables (department-level, d = 33 departments)				
GDP per capita	Lowest quartile (Q1)	26.58	285.41	78.98
	Q2	24.15	299.45	75.98
	Q3	25.27	294.78	76.80
	Highest quartile (Q4)	23.99	312.99	75.53
Educ. expenditure per student, department-level data	Lowest quartile (Q1)	26.42	279.24	75.90
	Q2	27.34	298.58	78.37
	Q3	25.66	307.51	77.13
	Highest quartile (Q4)	20.58	308.39	74.59

* Data for this variable is only available for n= 74,251 in j= 3,610, m= 193 and d= 33.

# Data for this variable is only available for n=177,602 in j=14,059, m=687 and d=32.

Table 9 Exploratory Data Analysis, Mathematics Grade 5 (country-level study)

Statistics presented at student-level		**%**	**Mean**	**Std. Dev.**
Level-one Variables (Student-level, n = 274,404 students)				
Male	Male	50.38	298.57	78.62
	Female	49.62	293.20	74.71
Level-two Variables (School-level, j = 17,200 schools)				
EN school	EN school	11.55	292.02	80.31
	Non-EN school	88.45	296.41	76.26
Session type	Complete day	10.75	313.18	81.61
	Morning	57.09	294.55	77.06
	Afternoon	32.15	292.55	73.74
Rural	Rural	23.51	282.25	78.17
	Urban	76.49	300.10	75.82
Private	Private	8.78	334.18	82.66
	Public	91.22	292.22	75.14
Socioeconomic level	NSE 1	20.97	269.89	76.35
	NSE 2	25.43	281.87	72.19
	NSE 3	31.90	298.94	70.41
	NSE 4	21.70	333.05	76.44
Ethnic population	With ethnic students	34.55	288.98	73.91
	Without ethnic students	65.45	299.56	77.96
Victims of conflict	Has conflict victim students	54.83	294.22	73.06
	Has no conflict victim students	45.17	297.96	80.97
Level-three variables (municipality-level, m = 1,009 municipalities)				
Homicide rate#	Lowest quartile (Q1)	37.95	302.24	76.89
	Q2	10.82	294.64	77.11
	Q3	24.50	290.54	75.35
	Highest quartile (Q4)	26.73	296.55	76.06
Governance index	Lowest quartile (Q1)	30.76	289.09	76.41
	Q2	19.35	287.12	75.77
	Q3	27.35	299.92	76.98
	Highest quartile (Q4)	22.54	307.88	75.87
Educ. expenditure per student, municipal-level data*	Lowest quartile (Q1)	25.44	286.24	74.40
	Q2	27.08	284.32	73.74
	Q3	39.82	308.29	74.32
	Highest quartile (Q4)	7.65	306.80	76.15
Level-four variables (department-level, d = 33 departments)				
GDP per capita	Lowest quartile (Q1)	26.22	281.40	76.47
	Q2	24.39	299.34	75.15
	Q3	24.09	288.52	75.51
	Highest quartile (Q4)	25.30	314.66	75.61
Educ. expenditure per student, department-level data	Lowest quartile (Q1)	26.02	273.12	73.37
	Q2	26.55	293.65	75.04
	Q3	25.71	308.87	77.80
	Highest quartile (Q4)	21.71	310.63	74.71

* Data for this variable is only available for n= 104,600 in j= 3,542, m= 191 and d= 33.

# Data for this variable is only available for n=248,400 in j=13,797, m=688 and d=32.

Table 10 Exploratory Data Analysis, Civic Competencies Grade 5 (country-level study)

Statistics presented at student-level		**%**	**Mean**	**Std. Dev.**
Level-one Variables (Student-level, n = 276,169 students)				
Male	Male	50.38	283.80	73.50
	Female	49.62	304.29	75.54
Level-two Variables (School-level, j = 17,533 schools)				
EN school	EN school	11.77	285.94	72.69
	Non-EN school	88.23	295.04	75.49
Session type	Complete day	10.80	307.90	78.84
	Morning	57.13	291.82	74.90
	Afternoon	32.07	293.11	74.03
Rural	Rural	23.83	278.53	71.45
	Urban	76.17	298.80	75.72
Private	Private	8.79	333.28	82.38
	Public	91.21	290.18	73.39
Socioeconomic level	NSE 1	21.26	267.82	68.77
	NSE 2	25.46	279.36	68.78
	NSE 3	31.75	297.17	71.20
	NSE 4	21.54	332.32	78.26
Ethnic population	With ethnic students	34.48	288.84	73.15
	Without ethnic students	65.52	296.67	76.15
Victims of conflict	Has conflict victim students	54.67	292.79	72.58
	Has no conflict victim students	45.33	295.39	78.26
Level-three variables (municipality-level, m = 1,010 municipalities)				
Homicide rate#	Lowest quartile (Q1)	37.89	299.16	75.86
	Q2	10.90	291.48	74.93
	Q3	24.53	289.67	74.05
	Highest quartile (Q4)	26.67	296.91	75.56
Governance index	Lowest quartile (Q1)	30.78	289.33	75.31
	Q2	19.30	282.98	71.88
	Q3	27.46	296.38	74.79
	Highest quartile (Q4)	22.45	306.83	76.38
Educ. expenditure per student, municipal-level data*	Lowest quartile (Q1)	25.38	293.91	77.03
	Q2	27.17	286.12	73.69
	Q3	39.82	308.84	76.49
	Highest quartile (Q4)	7.63	294.79	72.49
Level-four variables (department-level, d = 33 departments)				
GDP per capita	Lowest quartile (Q1)	26.31	279.33	72.20
	Q2	24.49	294.02	73.53
	Q3	27.53	295.25	76.82
	Highest quartile (Q4)	21.67	310.05	75.19
Educ. expenditure per student, department-level data	Lowest quartile (Q1)	26.05	277.66	73.71
	Q2	26.61	289.53	73.34
	Q3	25.68	302.32	74.82
	Highest quartile (Q4)	21.66	309.13	75.41

* Data for this variable is only available for n= 104,883 in j= 3,603, m= 191 and d= 10.

# Data for this variable is only available for n=250,016 in j=14070, m=688 and d=32.

In order to better illustrate the results of the exploratory data analysis, Figure 8 through Figure 11 show the mean test scores and standard deviations by subgroup/level for a selected few variables (gender, socioeconomic level of the school, session type, and departmental GDP/capita in quartiles). The standard errors for the means are not depicted in the graphs, as they would be too small to be visible. It becomes immediately clear from the graphs that the mean scores differ by subgroup/level, though the extent of the variation differs between the variables. For instance, the mean difference between boys and girls is barely visible in the graphs (even though it is statistically significant for most grade-areas), while the increase in average test scores from low to high socioeconomic levels is striking. In all cases, the standard deviations are considerable.

Put together, the data suggests that all of the proposed explanatory variables are correlated with exam scores and should thus be included in the model – with the exception of homicide rates and municipality-level data on education expenditures. However, there are good reasons to include both of these variables in the analysis nevertheless, as the effect of these variables might be concealed by intervening variables. For instance, in the case of homicides, which is used as a proxy of a peaceful and safe environment, there is a strong correlation with the location of the school, with urban areas experiencing more violence. Given that children in rural areas tend to score lower than children in urban areas, this might explain why exam scores do not seem correlate with homicide rate—even though children in areas with a higher homicide rates might score lower. Both variables (municipality-level data on education expenditures, and homicides) are only available for a smaller number of observations than the rest of the variables, hence they will be added to the main model later as a robustness check.

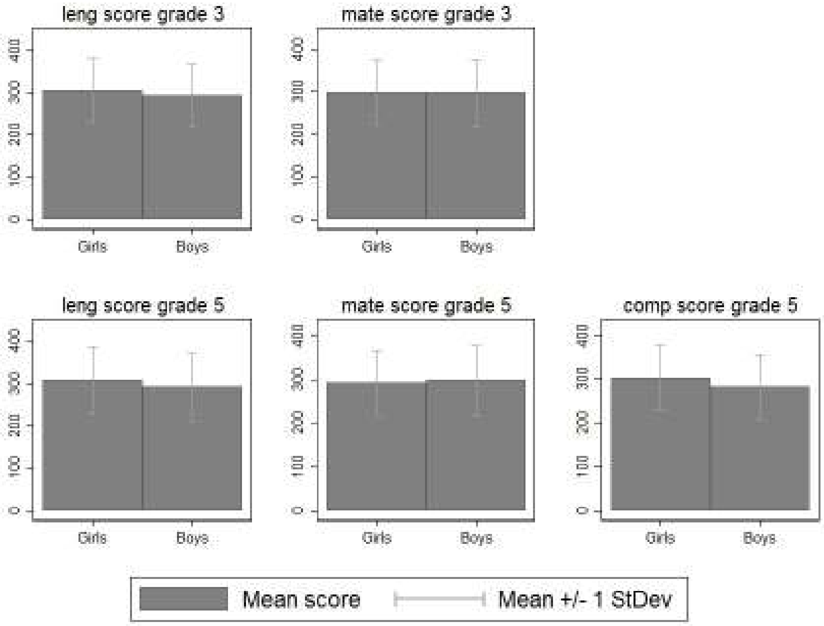


Figure 8 Exploratory Data Analysis: Mean test scores by gender (country-level study)

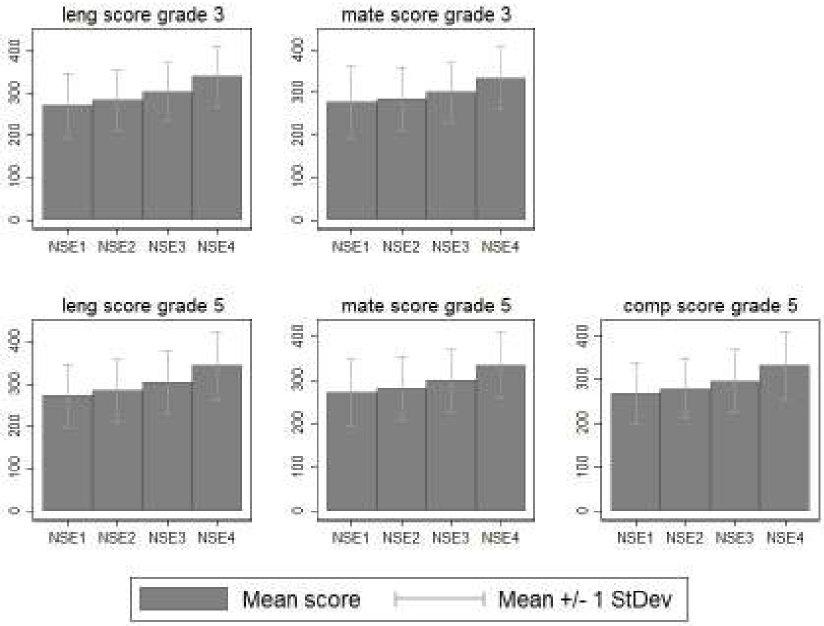


Figure 9 Exploratory Data Analysis: Mean test scores by socioeconomic level of school (country-level study)

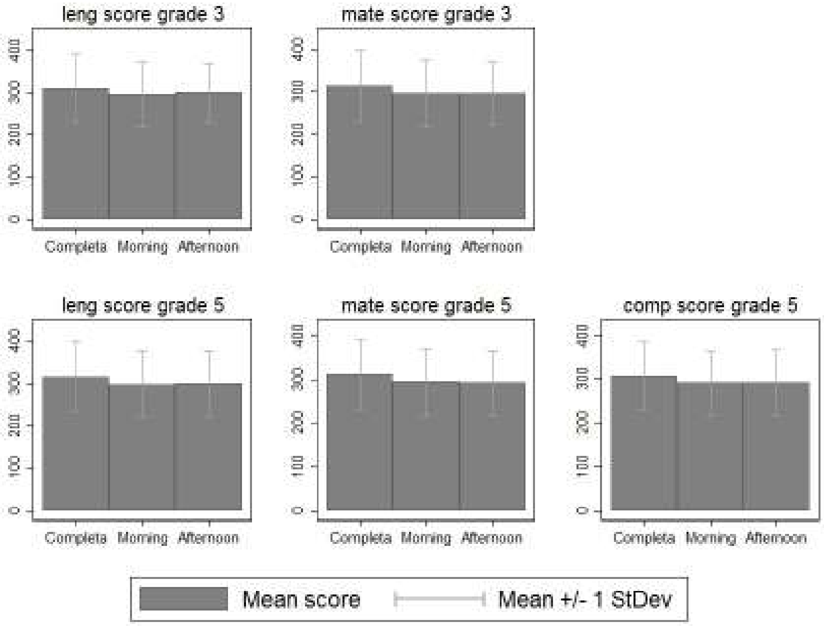


Figure 10 Exploratory Data Analysis: Mean test scores by session type (country-level study)

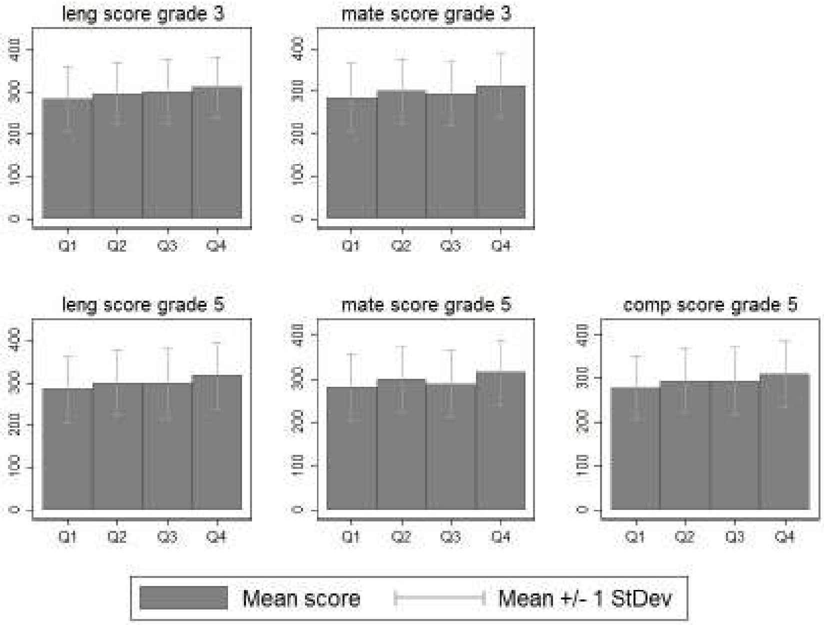


Figure 11 Exploratory Data Analysis: Mean test scores by departmental GDP per capita (quartiles) (country-level study)

4.1.2.2 School-level

The finding that there are correlations between the different regressors and the outcome variables does not necessarily mean that a multilevel model is necessary – a simple OLS model is equally capable of controlling for confounding factors. Rather, it is necessary to look for evidence for differences across schools. Is there evidence that student-level scores cluster around distinct school-level means?

Figure 12 helps to answer this question. It shows the spread of student-level grade 3 reading scores together with the respective school mean of the grade 3 reading score (based on the first plausible value). The data is ranked by the school-mean reading score; in order to make the graph easier readable, a random sample of 300 schools was selected for this exercise. Two points are immediately clear: First, the school mean scores vary significantly between schools, ranging from 110 to 603 for this specific sample, and from around -22 to 673 for the overall dataset. This suggests that the school regressions will, in fact, have different intercepts. Second, the range in the student-level scores varies as well between schools, with spreads anywhere between 0 and 348 for this specific sample and from 0 to 438 for the overall dataset. This variation suggests heteroskedastic student-level error terms. This general pattern is similar for the different grades and test areas.

A second question of interest is whether the correlation between student-level control variables and learning outcomes differs across schools. If this is the case, the slope estimates might differ between schools, which would suggest a need for predictor-dependent random effects. The only student-level control variable in this analysis is gender (and its interaction with the school model). Figure 13 shows the school mean grade 3 reading scores for each gender for the random sample of 300 schools. The means within schools are connected by lines in order to depict gender differences. The graph shows that school mean scores differ across schools, as was already illustrated in Figure 12 – this again indicates that school-level intercepts will differ. More interestingly in this context are the slopes of the lines in the graph. First, they are relatively flat, which indicates little difference between boys and girls (this mirrors the first panel of Figure 8). Second, the difference in the slopes between the schools seems moderate, which indicates that the model may not require different slopes for gender for each school.

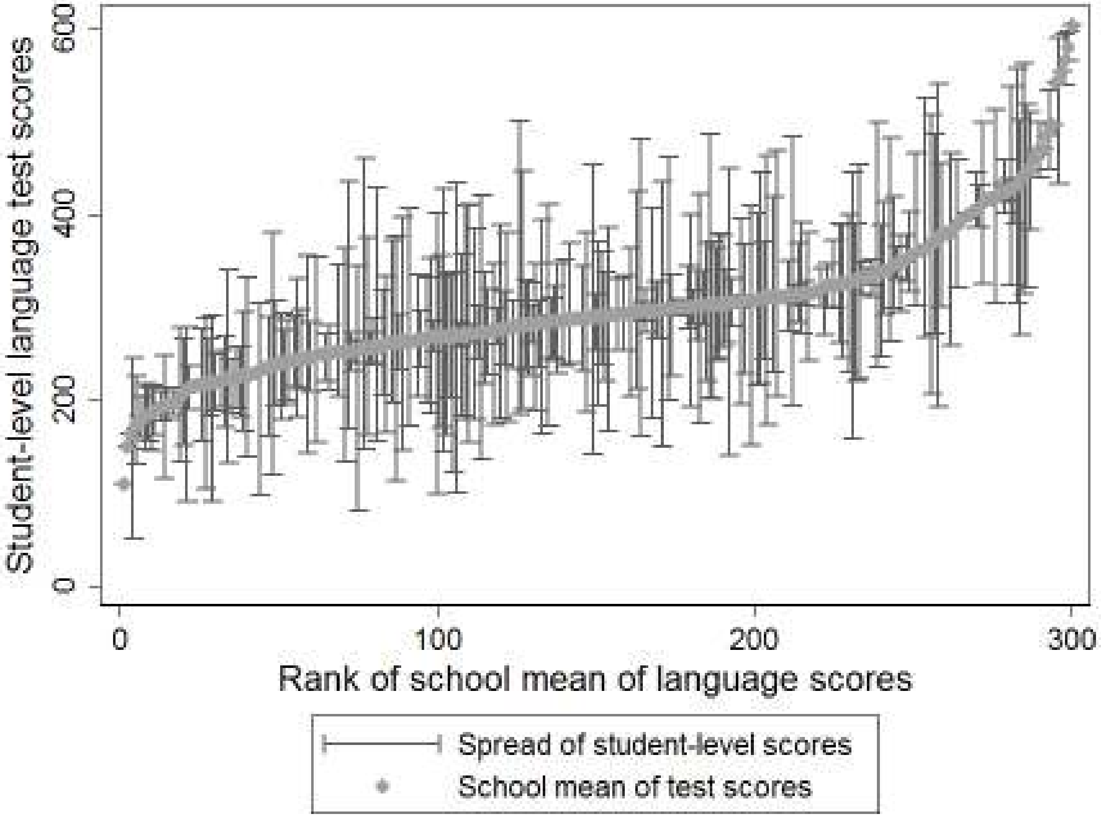


Figure 12 Variation in student-level reading scores (plausible value 1) ranked by school-mean of reading scores, for a random sample of 300 3rd grade classes (country-level study)

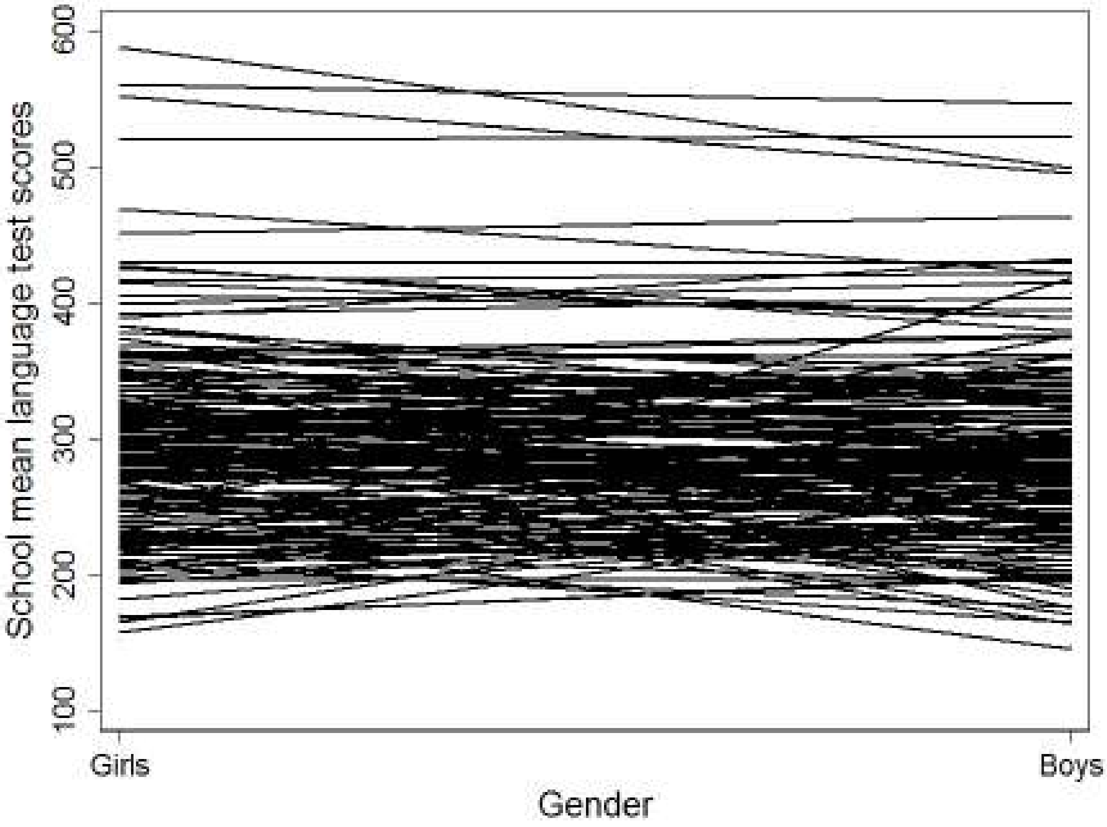


Figure 13 Mean language test scores by gender, for a random sample of 300 3rd grade classes. Lines are connecting school means (country-level study)

4.1.2.3 Municipality-level

The next step is to explore whether the introduction of a municipality-level (level 3) should be considered. As before, the question is whether students differ across municipalities, both with regard to their mean test scores and with regard to the variance in the test scores. Additionally, it is interesting to check whether school-level predictors differ across municipalities.

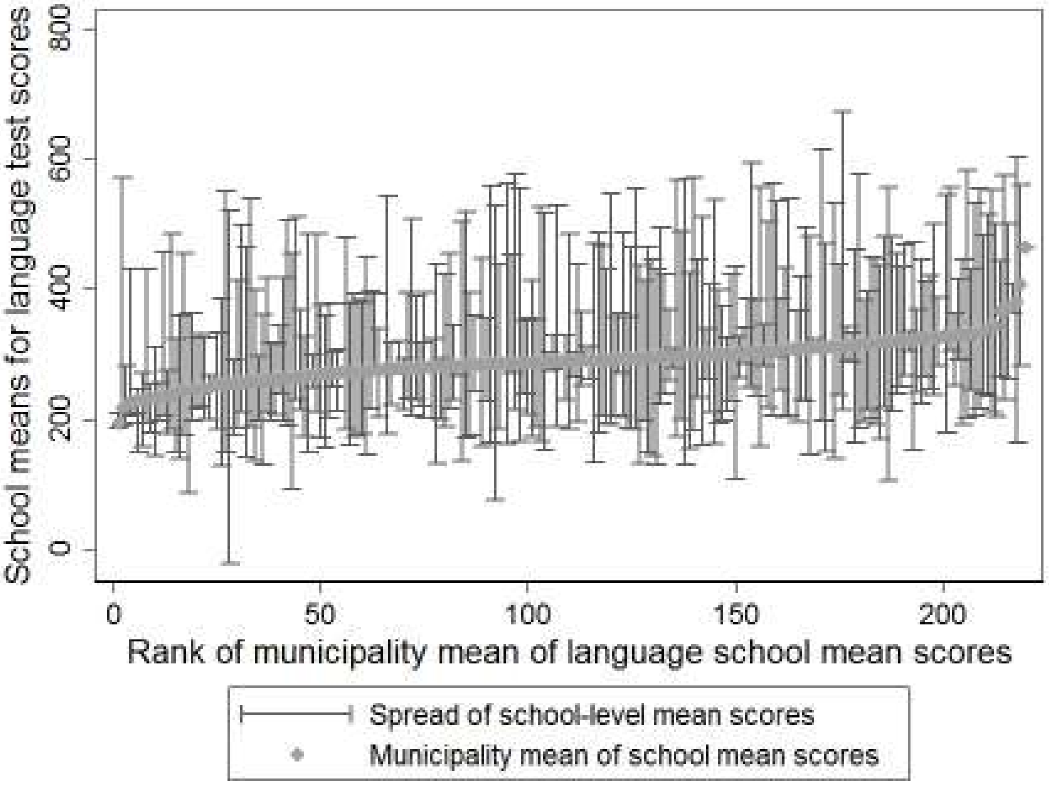


Figure 14 Variation in school-mean reading scores (plausible value 1) ranked by municipality-mean of school-mean reading scores, 3rd grade (country-level study)

Evidence for the first set of questions is presented in Figure 14. It shows the spread of school-level mean grade 3 reading scores together with the respective municipality mean of the grade 3 school-mean reading score (based on the first plausible value). The data is ranked by the municipality-mean of the school mean reading scores. The graph reveals differences across municipalities: the municipality-mean reading scores range between 185 and 464; the spread in the school-means for different municipalities range between 0 and 561. This graph indicates that municipality-level regressions will have different intercepts and error terms. Thus, the inclusion of a municipality-level into the multilevel model appears appropriate.

The second set of questions aims at exploring whether the different student-level or school-level regressors might have different slopes across municipalities. The eight panels of Figure 15 help to answer this set of questions. First, for all of the predictors there is considerable variation in average scores across municipalities, supporting the thesis that municipality-level intercepts will differ. Second, while the municipality-slopes seem mostly parallel for the predictor variables male and private school, there is some more variation across municipalities in the slopes of the other variables, especially for EN, rural areas, and session type. For the latter variables, there may thus be the need for random coefficients.

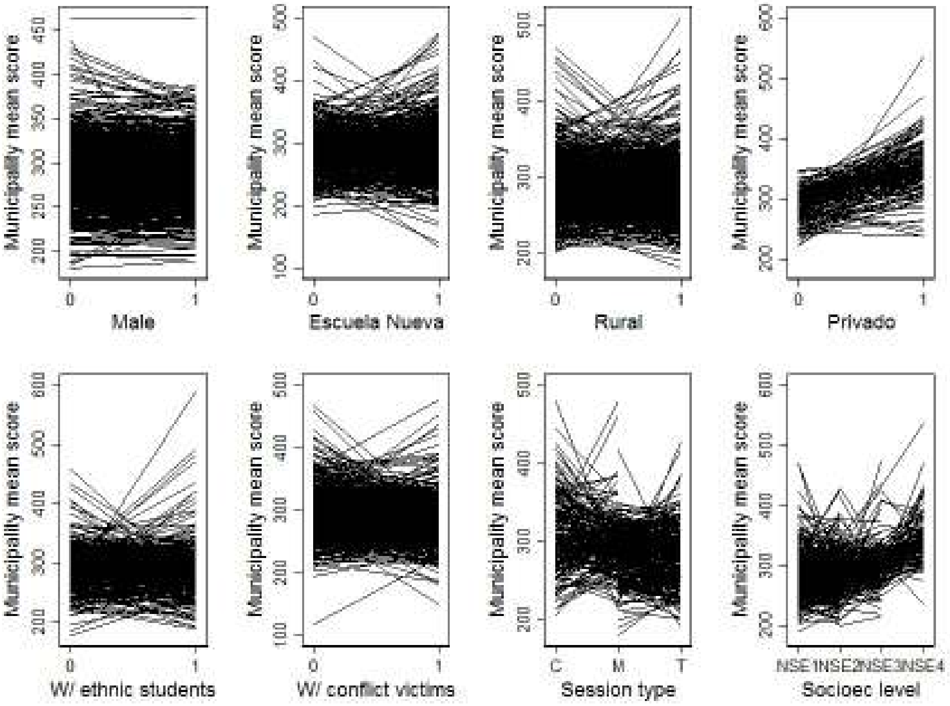


Figure 15 Municipality mean grade 3 language test scores by levels of lower-level predictors. Lines are connecting municipality means (country-level study)

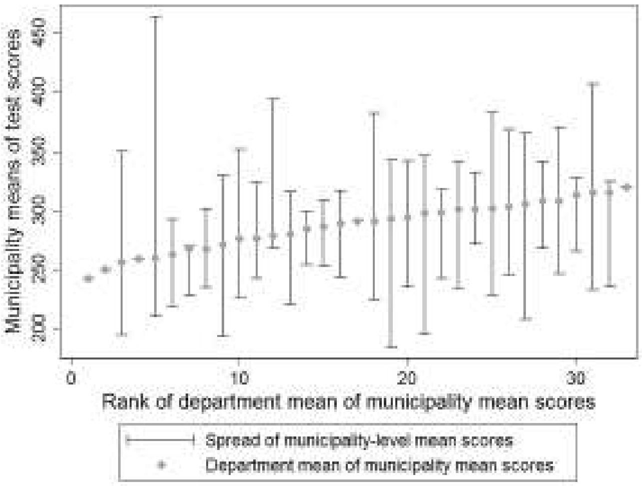


Figure 16 Variation in municipality-mean reading scores (plausible value 1) ranked by department-mean of municipality-mean reading scores, 3rd grade (country-level study)

4.1.2.4 Department-level

Finally, it remains to be explored whether the inclusion of a level 4 (departments) might be necessary. Figure 16 plots the spread of municipality-level mean grade 3 reading scores together with the respective department mean (based on the first plausible value), ranked by the department mean. The graphs show differences between departments, both with regard to the average reading score and with regard to the variance in means scores. These differences in the vertical spread suggest that the departments will have different intercepts and different error term variances. The inclusion of a department-level random effect in the multilevel might thus be necessary.

Figure 17 shows how department-level averages of test scores differ across the levels of the student-, school-, and municipality-level predictor variables. As before, the need for department-level intercepts is apparent in the vertical spread of the means. The graphs also show fairly parallel slopes for most variables, the exceptions possibly being school type (EN), schools with ethnic students, and homicides.

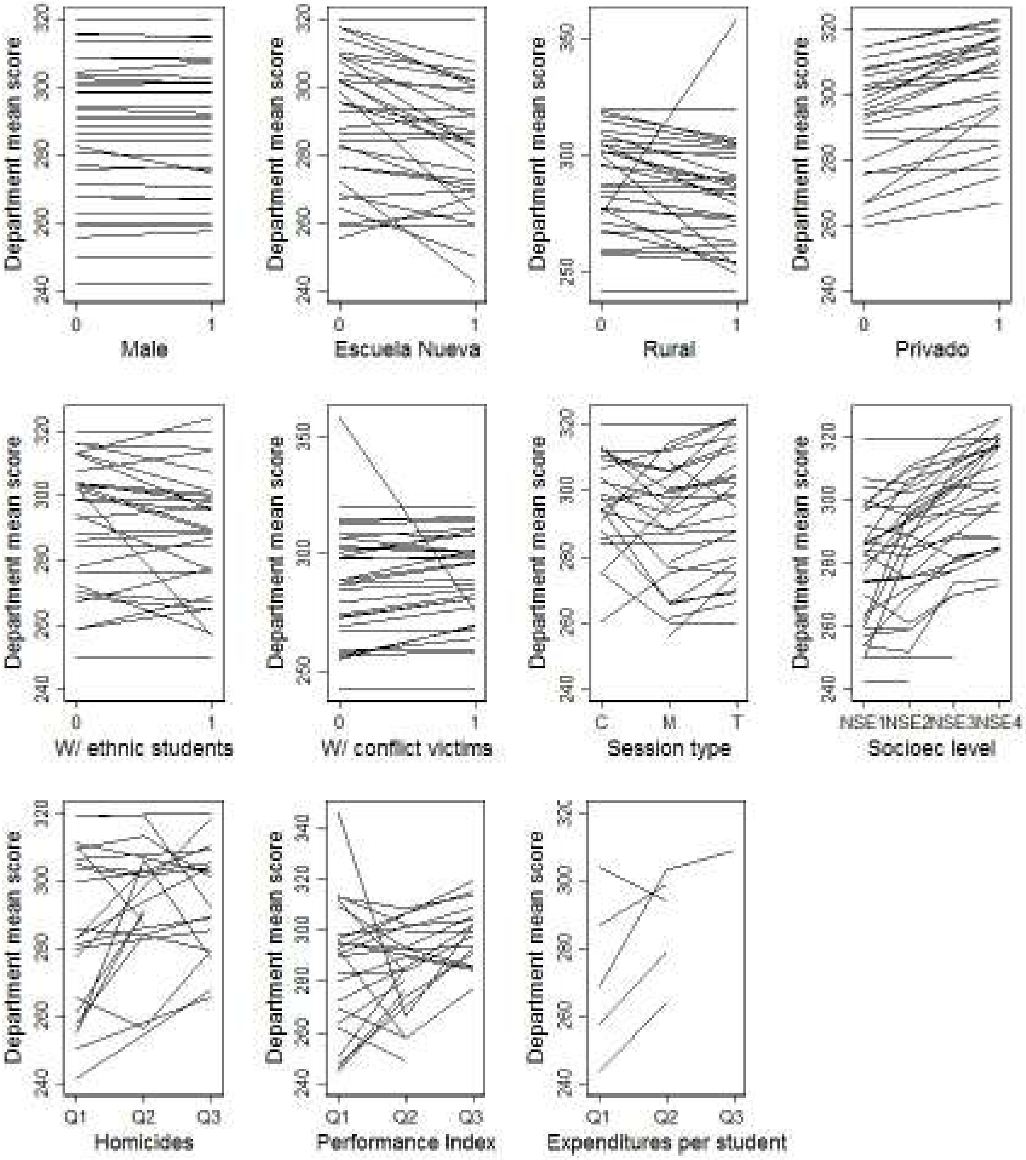


Figure 17 Department mean grade 3 language test scores by levels of lower-level predictors. Lines are connecting department means (country-level study)

4.2 The null model (ANOVA)

An analysis of variance gives insights into the sources of variance in test scores. What portion is due to differences between students, and what portion is due to differences between schools, municipalities, and departments? A random effects ANOVA, the so-called null model or variance component model, is estimated for each grade and testing area for each of these levels.

4.2.1 Two-level analysis

For the initial two-level case, the null model takes the form:

$$score_{ij} = \beta_j + \varepsilon_{ij}$$

Where: $$\beta_j = \beta + \zeta_j$$

So that $$\xi_{ij} = \zeta_j + \varepsilon_{ij} .$$

This model describes that the test score of student i in school j is composed of the school mean, β_j, and a random student-level error, ε_{ij}. The school-level mean, β_j, is itself composed of the grand mean, β, and a school-level random error term, ζ_j. (For better readability, the model does not include superscripts for grade or testing area). A different way to look at this model is to define the test score of student i in school j as the grand mean, β, and a composite error term, ξ_{ij}, which consists of the school-level error, ζ_j, and the student-level error, ε_{ij}.

Of further interest is the intra-class correlation coefficient (ICC), which denotes the percentage of the overall variance that is due to the difference at the higher level. The ICC is defined as

$$\rho = \frac{\psi}{\psi + \theta}$$

where ψ is the variance of the school-level error term ζ_j, and θ is the variance of the student-level error term ε_{ij}.

Table 11: Two-level variance components models (country-level study)

	Language, 3rd grade		**Language, 5th grade**		**Math, 3rd grade**		**Math, 5th grade**		**Civics, 5th grade**	
n (students)	197,234		277,179		195,978		274,404		276,169	
j (schools)	17,652		17,586		17,475		17,200		17,533	
Fixed part:										
Grand mean	293.39	(0.49)	293.85	(0.44)	301.24	(0.53)	294.33	(0.48)	289.57	(0.41)
Random part (sd):										
School-level	54.06	(0.44)	45.94	(0.40)	60.00	(0.46)	49.24	(0.40)	40.24	(0.35)
Student-level	59.53	(0.12)	67.55	(0.16)	61.49	(0.11)	63.92	(0.11)	65.00	(0.12)
ICC (schools)	0.45		0.32		0.49		0.37		0.28	
LR χ^2	62466.24		65521.63		59641.01		71768.92		56605.35	
p-value (χ^2)	0.000		0.000		0.000		0.000		0.000	

Standard errors in parenthesis.

Table 11, above, summarizes the estimates for these parameters for each grade and testing area. The first part of the table contains the number of observations for each model (number of students and number of schools). For the grade 3 models, there are just under 200,000 observations, while the grade 5 estimates are based on around 275,000 observations.

The second part of the table contains the fixed part of the models, that is, the estimates for the grand mean test score across all schools and students, β. For the case of the 3rd grade language exam, this grand mean is estimated as 293.39; for the 5th grade language exam, as 293.85, etc. The point estimates are precise, as indicated by the small standard errors.

The third part of the table reports the results for the random part of the model: the estimates for the school-level standard deviation, $\sqrt{\psi}$, and the student-level standard deviation, $\sqrt{\theta}$, respectively. Both are relatively high for all models, which shows that there is large variation both between schools ($\sqrt{\psi}$) and within schools ($\sqrt{\theta}$). In order to facilitate a better understanding of this variation, the table reports the intra-class correlation coefficients (ICCs), i.e. the percentage of the overall variance that is due to differences across schools. The ICCs differ considerably between grades and testing areas, but generally speaking they indicate that between one third and half of the variance is due to school-level effects. The most extreme cases are civic competencies, where only 28% of the variance stems from school-level differences, and 5th grade mathematics, where 49% of the variance is due to variance between schools. It also appears that school-level differences are more important in grade 3, and become less important in grade 5. In other words, the student background becomes more important as students grow older. This points towards a failure of Colombian primary schools to improve equity (or, indeed, even to a failure not to increase inequity). The finding that most of the variance in test results in Colombia is due to student-level effects mirrors the results from previous studies on the topic (Casassus et al. 2000; Rangel and Lleras 2010; Baron 2012; Zambrano Jurado 2013).

Before moving forward with the multilevel model, it is advisable to check whether such a model is in fact appropriate for the data. Specifically, it is necessary to test the null hypothesis that the between-school variance, ψ, is zero and that there is thus no random intercept ζ_j in the model (Rabe-Hesketh and Skrondal 2012, 88). In the absence of a random intercept, ordinary regression should be used, as it is more efficient. Formally:

$$H_0: \psi = 0$$

$$H_a: \psi > 0$$

Because standard errors of variance components tend to be unreliable in a multilevel model, they should not be used to evaluate the significance of the variance components (Kim, Anderson, and Keller 2013). Instead, the null hypothesis is tested with a likelihood-ratio test, which tests whether the model without the random intercept ($\psi = 0$) is nested within a two-level null model ($\psi > 0$). The χ^2 statistic of this test for each of the testing areas and grades is reported in the last part of

Table 11, together with the halved p-value.[14] The results clearly show that the null hypothesis of no random intercept is to be rejected, and multilevel analysis is appropriate.

4.2.2 Three-level analysis

Students are nested not only within schools, but also within geographic clusters, most notably municipalities and departments. A priori, there are good reasons to believe that these geographic clusters matter for learning outcomes. Apart from cultural differences, municipalities and departments have considerable autonomy with regard to education policy and funding. It is to be expected that observed and unobserved student-, school-, and higher-level characteristics (in particular the mean test scores, but also some of the control variables) not only vary between schools, but also between municipalities and departments. Preliminary evidence for this hypothesis was presented in the exploratory data analysis (section 4.1).

To test whether municipalities should be included as an additional level in the model, a three-level variance components model (three-level null model) is defined as:

$$score_{ijm} = \beta_{jm} + \varepsilon_{ijm}$$

Where: $\beta_{jm} = \beta_m + \zeta_{jm}$

and: $\beta_m = \beta + \zeta_m$

So that $\xi_{ijm} = \zeta_m + \zeta_{jm} + \varepsilon_{ijm}$.

[14] As ψ cannot be negative, the asymptotic sampling distribution under the null hypothesis is a 50:50 mixture of $\chi^2(0)$ and $\chi^2(1)$, i.e. a distribution with a spike a 0. The p-value obtained from the LR-test is therefore conservative, and the correct p-value is obtained by dividing this value by 2 (Rabe-Hesketh and Skrondal 2012, 88–89).

In words, these models describe that the test score of student i in school j in municipality m is composed of the school mean, β_{jm}, and a random student-level error, ε_{ijm}. The school-level mean, β_{jm}, is composed of the municipality mean β_m and a school-level random error term, ζ_{jm}. Finally, the municipality mean β_m is a combination of the grand mean β and a municipality-level error term, ζ_m. Substituting the higher-level equations into the level-one equation shows that the overall error term ξ_{ijm} has three components: the municipality-level error, ζ_m; the school-level error, ζ_{jm}; and the student-level error, ε_{ijm}.

Equivalent to the two-level model, intra-class correlation coefficients help to understand the share of variance of each level. With $\psi_{(2)}$ being the variance of the level-two error term ζ_{jm} and $\psi_{(3)}$ being the variance of the level-three error term ζ_m, there are various ICC of interest:

The share of school-level variance in overall variance: $$\rho_{(2)} = \frac{\psi_{(2)}}{\psi_{(3)} + \psi_{(2)} + \theta}$$

The share of municipality-level variance in overall variance: $$\rho_{(3)} = \frac{\psi_{(3)}}{\psi_{(3)} + \psi_{(2)} + \theta}$$

The combined share of school- and municipality-level variance: $$\rho_{(2,3)} = \frac{\psi_{(2)} + \psi_{(3)}}{\psi_{(3)} + \psi_{(2)} + \theta}$$

Of course, $1 - \rho_{(2,3)}$ is the share of student-level variance in overall variance.

Table 12 shows the estimation results for these three-level models for all grades and testing areas; the presentation of the results parallels the one in Table 11. The models were extended by a municipality-level random effect, based on just over 1,000 municipalities.

The estimates for the fixed part of the model have changed slightly compared to the two-level model, but are in the same general range (for instance, 290.24 for 3rd grade language scores and 292.04 for 5th grade language scores). More interestingly, the random part shows that the estimates for the standard deviations of the student-level error terms barely change with the introduction of a third level. Instead, the school-level error term variance changes, as a significant part of this school-level error is due to differences across municipalities. This indicates that across municipalities, school-level factors differ more than student-level factors. A look at the intra-class correlation coefficients shows that the share of the variance that is due to student-level factors has indeed hardly changed, but that between 10 and 14 percentage points of what was previously identified as school-level variance is actually due to differences across municipalities.

Table 12 Three-level variance component models (country-level study)

	Language, 3rd grade	**Language, 5th grade**	**Math, 3rd grade**	**Math, 5th grade**	**Civics, 5th grade**
n (students)	197,234	277,179	195,978	274,404	276,169
j (schools)	17,652	17,586	17,475	17,200	17,533
m (municipalities)	1,007	1,011	1,009	1,009	1,010
Fixed part:					
Grand mean	290.24 (1.02)	292.04 (1.00)	299.72 (1.20)	293.80 (1.12)	288.08 (0.90)
Random part (sd):					
Municipality-level	25.69 (0.89)	26.09 (0.84)	30.89 (1.02)	30.05 (0.93)	24.13 (0.76)
School-level	47.06 (0.43)	37.76 (0.38)	52.53 (0.44)	40.41 (0.38)	32.37 (0.33)
Student-level	59.58 (0.12)	67.60 (0.16)	61.52 (0.11)	63.97 (0.11)	65.05 (0.12)
ICC (schools)	0.34	0.21	0.37	0.25	0.18
ICC (municipalities)	0.10	0.10	0.13	0.14	0.10
ICC (j, m)	0.45	0.32	0.50	0.38	0.28
LR χ^2	1850.81	1850.81	1846.10	2637.13	2708.61
p-value (χ^2)	0.000	0.000	0.000	0.000	0.000

Standard errors in parenthesis.

In light of the considerable computational power needed to calculate three-level models, the question again becomes whether the inclusion of a municipality-level random effect is indeed necessary. That is, it is necessary to test whether $\psi_{(3)}$, the variance of the level-three error term ζ_m, is zero. Formally, the null hypothesis and the alternative hypothesis are

$$H_0\colon \psi_{(3)} = 0$$

$$H_a\colon \psi_{(3)} > 0$$

The χ^2statistics of the likelihood-radio tests carried out to test these hypotheses are again reported in the last part of Table 12, together with the halved p-values. The test results unambiguously show that the municipality-level random intercepts are in fact different from zero and should be included in the model.

4.2.3 Four-level analysis

Municipalities are nested within departments, which are the political units with major decision-making power in the Colombian education sector. Differences across departments are therefore expected. Hence, a random department-level intercept is added to the model, so that the error term is now composed of four levels. Formally:

$$score_{ijmd} = \beta_{jmd} + \varepsilon_{ijmd}$$

Where: $$\beta_{jmd} = \beta_{md} + \zeta_{jmd}$$

and: $$\beta_{md} = \beta_d + \zeta_{md}$$

and: $$\beta_d = \beta + \zeta_d$$

so that $$\xi_{ijmd} = \zeta_d + \zeta_{md} + \zeta_{jmd} + \varepsilon_{ijmd}$$

Here, ζ_d is the department-level error, ζ_{md} is the municipality-level error, ζ_{jmd} is the school-level error, and ε_{ijmd} is the student-level error; β_{jmd} is the school-level mean, β_{md} is the municipality-level mean, β_d is the department-level mean, and β is the grand mean. There are now four error-term variances: $\psi_{(4)}$ is the variance of the department-level error, ζ_d; $\psi_{(3)}$ is the variance of the municipality-level error term, ζ_{md}; $\psi_{(2)}$ is the variance of the school-level error term, ζ_{jmd}; and θ is the variance of the student-level error, ε_{ijmd}.

Table 13 Four-level variance component models (country-level study)

	Language, 3rd grade	**Language, 5th grade**	**Math, 3rd grade**	**Math, 5th grade**	**Civics, 5th grade**
n (students)	197,234	277,179	195,978	274,404	276,169
j (schools)	17,652	17,586	17,475	17,200	17,533
m (municipalities)	1,007	1,011	1,009	1,009	1,010
d (departments)	33	33	33	33	33
Fixed part:					
Grand mean	285.80 (3.40)	283.91 (3.94)	292.72 (3.89)	284.70 (4.26)	281.25 (3.59)
Random part (sd):					
Department-level	17.28 (2.56)	20.78 (2.93)	19.71 (3.00)	22.44 (3.22)	18.98 (2.65)
Municipality-level	19.46 (0.79)	17.00 (0.69)	23.76 (0.92)	20.28 (0.78)	15.84 (0.63)
School-level	46.96 (0.43)	37.59 (0.38)	52.43 (0.44)	40.28 (0.38)	32.21 (0.32)
Student-level	59.59 (0.12)	67.61 (0.16)	61.53 (0.11)	63.97 (0.11)	65.05 (0.12)
ICC (schools)	0.34	0.21	0.37	0.24	0.18
ICC (municipalities)	0.06	0.04	0.08	0.06	0.04
ICC (departments)	0.05	0.06	0.05	0.08	0.06
ICC (j, m, d)	0.45	0.32	0.49	0.38	0.28
LR χ^2	267.16	466.85	258.92	435.40	460.25
p-value (χ^2)	0.000	0.000	0.000	0.000	0.000

Standard errors in parenthesis.

As was the case for the three-level null models, different ICC can be calculated. Some of them are presented in Table 13. The most important change compared to the three-level model is that the municipality-level variance is almost equally split up between a department-level- and a municipality-level variance. The share of the school- or student-level variance in the overall variance has barely changed. The most important source of variance clearly remains at the student-level.

The last part of the table contains the results of the likelihood-ratio test which was performed to test whether the inclusion of a department-level random effect is indeed necessary ($H_0: \psi_{(4)} = 0$, versus $H_a: \psi_{(4)} > 0$). The halved p-values for the significance of χ^2 are again estimated as approaching zero, which means that a department-level random effect should be included in the model.

Based on the results of this analysis, the following sections will thus present the full model as a four-level hierarchical model.

4.3 The full model

The full model is built using the “step-up” method of model construction (Ryoo 2011; cited in Kim, Anderson, and Keller 2013): the starting point is a simple random-intercept model, and from there first the fixed effects are successively added, followed by random coefficients.

4.3.1 Random-intercept model

4.3.1.1 Development of the model

Using the null model developed in the previous section as the starting point, the random-intercept multilevel model is developed by adding control variables step by step according to the level they belong to. This step-by-step procedure helps to create a more stable model, and it also helps to better understand which part of the respective level’s remaining variance the predictors can explain. The methodological annex (Annex B) contains the details of this modeling procedure with the specific models and estimation results for each step. As it turns out, the available control variables at the school- and municipality-level increase the explanatory power of the model, while adding the department-level regressors (departmental GDP per capita and departmental public education expenditure per student) does not improve model fit. The best-fitting random-intercept model is thus model RI3, which contains control variables from the first three levels:

Model RI3: $$score_{ijmd} = \beta_0 + \beta_1 EN_{jmd} + \beta_2 male_{ijmd} + \beta_3 (male * EN)_{ijmd} + \beta_4 rural_{jmd} + \beta_5 private_{jmd} + \beta_6 NSE_{jmd} + \beta_7 (NSE * EN)_{jmd} + \beta_8 ethnic_{jmd} + \beta_9 conflict_{jmd} + \beta_{10} morning_{jmd} + \beta_{11} afternoon_{jmd} + \beta_{12} governance_{md} + \xi_{ijmd}$$

where ξ_{ijmd} is the composed error term consisting, as discussed in the previous section, of the student-level error term ε_{ijmd}, the school-level error term ζ_{jmd}, the municipality-level error term ζ_{md}, and the department-level error term ζ_d. Apart from the Escuela Nueva dummy (EN), the model includes predictors on three levels. At the student-level, $male$ is a dummy for gender, and $male * EN$ a cross-level interaction term of EN and gender. This interaction is testing the hypothesis that the effect of the EN model differs by gender. With regard to school-level regressors, the model contains the variables $rural$ (a dummy for whether the school is in a rural area); $private$ (a dummy for whether the school is private); NSE (the socioeconomic level of the school, defined as the average official socioeconomic level of the children; level 1 is coded as zero, levels 2, 3, and 4 are coded as 1, 2, and 3); the interaction of NSE and EN, which tests the hypothesis that the EN model is particularly beneficial for children from disadvantaged backgrounds; $ethnic$ (a dummy for whether there are students of ethnic background in the school); $conflict$ (a dummy for whether there are children in the school who are victims of the conflict); and $morning$ and $afternoon$ to indicate the type of the session (a full school day being the base category). Finally, on the municipality-level it includes the variable $governance$, the governance index of the municipality (see section 3 of Annex A for an explanation of the index).

GDP per capita and education expenditure per student by department are *not* included in model RI3 because they lack individual and joint statistical significance. Education expenditure per student *by municipality* is used in section 4.3.3 for testing the robustness of the results.

4.3.1.2 Results

Table 14 combines the results of the random intercept model (Model RI3) for all grades and areas. Although the random-intercept model is not the final model, a look at the results is interesting because the interpretation of the results (particularly with regard to the variances) changes with the inclusion of a random coefficient in the next section. The estimated grand mean score is around 280 for all grades and testing areas, except for grade 5 mathematics, where the mean is only 270. This mean score is important to keep in mind because it provides the reference point for the rest of the estimates. Given the control variables and their coding, the grand mean is to be interpreted as the expected test score for girls in urban, public, full-day-session non-EN schools that have no students of ethnic background nor students who are conflict victims, that serve families whose average socioeconomic level is NSE1 (i.e., the lowest level), and that are in municipalities with average governance index scores.

First and foremost, the *ceteris paribus* effect of EN on test scores is large and statistically highly significant for all models. The estimated effect is largest for grade 5 mathematics, where students in EN schools are estimated to score 19.5 points higher than comparable children in conventional schools. This is a considerable difference. The EN-effect is smallest in the civic competencies exam, where girls score “only” 8.4 points and boys 11.07 points higher than their peers in conventional schools; in 5th grade mathematics, the EN-effect for boys is also “only” 8.4 points. The random intercept models provide little support for the hypothesis that EN is particularly beneficial for girls: the gender-EN interaction is significant only for 5th grade mathematics and civic competencies, and in the latter case, boys did better than girls.

Table 14 Overview: Results of the final random intercept models RI3, all grades and testing areas (country-level study)

Overview RI3 models	**Language, 3rd grade**			**Language, 5th grade**			**Math, 3rd grade**			**Math, 5th grade**			**Civics, 5th grade**		
n (students)	197,234			277,179			195,978			274,404			276,169		
j (schools)	17,652			17,586			17,475			17,200			17,533		
m (municipalities)	1,007			1,011			1,009			1,009			1,010		
d (departments)	33			33			33			33			33		
Fixed part:															
Escuela Nueva	11.68	***	(1.59)	9.23	***	(1.43)	19.44	***	(1.89)	12.41	***	(1.43)	8.41	***	(1.23)
Male	-10.25	***	(0.32)	-13.55	***	(0.29)	0.59		(0.32)	7.22	***	(0.30)	-18.61	***	(0.30)
EN*Male	1.46		(0.93)	-0.31		(1.00)	-1.80		(1.11)	-3.06	***	(0.88)	2.66	**	(0.86)
Rural	1.70		(1.51)	4.22	***	(1.22)	4.55	**	(1.69)	4.54	***	(1.33)	3.69	***	(1.06)
Private	40.20	***	(2.00)	26.71	***	(1.57)	42.20	***	(2.19)	27.36	***	(1.74)	25.97	***	(1.43)
Socioeconomic level	13.77	***	(0.79)	17.67	***	(0.63)	11.26	***	(0.90)	15.76	***	(0.72)	15.07	***	(0.56)
EN*Socioeconomic level	-7.67	***	(1.52)	-9.42	***	(1.27)	-7.16	***	(1.68)	-9.08	***	(1.39)	-8.11	***	(1.17)
w/ ethnic students	-4.30	***	(1.28)	-4.33	***	(1.05)	-3.39	**	(1.45)	-4.50	***	(1.15)	-3.33	***	(0.91)
w/ conflict victims	-6.32	***	(1.07)	-3.40	***	(0.91)	-8.41	***	(1.23)	-2.76	**	(0.96)	-3.05	***	(0.81)
Morning session	-1.24		(1.67)	-3.91	**	(1.37)	0.98		(1.85)	-2.83		(1.49)	-3.07	*	(1.22)
Afternoon session	-7.17	***	(1.76)	-10.85	***	(1.44)	-5.14	**	(1.92)	-9.83	***	(1.52)	-7.64	***	(1.26)
Governance Index	0.27	***	(0.08)	0.23	***	(0.06)	0.29	**	(0.09)	0.27	***	(0.08)	0.16	**	(0.06)
Grand mean	280.40	***	(3.31)	280.16	***	(3.28)	278.00	***	(3.74)	269.80	***	(3.69)	280.00	***	(3.09)
Random part (sd):															
Department-level	12.56		(2.00)	14.49		(2.13)	14.23		(2.32)	16.20		(2.44)	14.00		(2.03)
Municipality-level	17.53		(0.78)	15.07		(0.65)	22.80		(0.90)	19.15		(0.77)	14.23		(0.61)
School-level	43.09		(0.43)	32.79		(0.37)	49.15		(0.44)	36.78		(0.37)	27.64		(0.31)
Student-level	59.46		(0.12)	67.37		(0.16)	61.58		(0.12)	63.93		(0.11)	64.54		(0.12)
ICC (schools)	0.32			0.18			0.35			0.22			0.14		
ICC (municipalities)	0.05			0.04			0.07			0.06			0.04		
ICC (departments)	0.03			0.03			0.03			0.04			0.04		
Total Variance	5857.393			6050.25			6929.95			6068.67			5327.67		
R^2 (Var. explained (L1-L4))	8.95%			9.76%			7.44%			8.47%			9.41%		
R^2 (Var. explained (L2-L4))	19.45%			29.15%			15.24%			21.89%			29.49%		
R^2 (Var. explained (L2))	15.81%			23.93%			12.12%			16.65%			26.36%		

Standard errors in parenthesis. ***: p≤0.001, **: p≤0.01, *: p≤0.05

Table 15 Estimated joint marginal effect of EN and socioeconomic status, based on model RI3 (country-level study)

	Language 3		**Language 5**		**Math 3**		**Math 5**		**Civics 5**	
	EN	**Non-EN**	**EN**	**Non-EN**	**EN**	**Non-EN**	**EN**	**Non-EN**	**EN**	**Non-EN**
NSE1	11.68	(base)	9.23	(base)	19.44	(base)	12.41	(base)	8.41	(base)
NSE2	17.78	13.77	17.48	17.67	23.54	11.26	19.09	15.76	15.37	15.07
NSE3	23.88	27.54	25.73	35.34	27.46	22.52	25.77	31.52	22.33	30.14
NSE4	29.98	41.31	33.98	53.01	31.74	33.78	32.45	47.28	29.29	45.21

In line with worldwide trends, boys did significantly worse in both language exams than girls, while they did do significantly better in 5th grade mathematics. In civic competencies, boys scored as much as 18.6 points less than girls (other things being equal). Equally unsurprising is the finding that students from private schools score significantly higher than children from public schools; the difference is over 40 points for both grade 3 exams, and just over 25 points for all grade 5 exams (which would suggest that the advantage of attending a public school decreases over a primary student's lifetime).

The estimation coefficients of socioeconomic level have their expected signs. First, the independent effect of socioeconomic level is strong, positive, and statistically significant, confirming that children of higher socioeconomic levels do better in all of the tests. Second and more importantly, the effect of the interaction term of EN and socioeconomic level is also highly significant, and negative. This provides support for the hypothesis that the EN model is particularly beneficial for children from disadvantaged backgrounds. In order to fully understand how socioeconomic status and EN-effect work together, all three coefficients (EN, socioeconomic level, and their interaction) have to be interpreted jointly. Table 15 helps to do that: it contains the estimated joint marginal effects of socioeconomic status and Escuela-Nueva school. The table suggests that the EN model indeed helps to close gaps between children from the different socioeconomic levels: c.p., students from EN schools generally outperform students from non-EN schools at the lowest two levels (the difference is not clear for NSE2 in grade 5 language and civics).

The results furthermore show that students from schools with children of ethnic backgrounds or with children who are conflict victims do significantly worse than students in schools without these populations. Furthermore, children from full-day-session ("*jornada completa*") schools do better than children in afternoon-session schools; there is no difference to children in morning-session schools, except in grade 5 language and civic competencies, where the latter do worse.

What may be surprising is that once the other factors are taken into account, children in rural schools score significantly higher than children in urban schools (except in the grade 3 language exam, where there is no statistical difference between rural and urban schools). This contradicts common knowledge and the results of the exploratory data analysis. However, this phenomenon has been discussed in the literature before: According to UNESCO (1998), Colombia is the only Latin American country where rural schools outperform urban ones, except in large metropolitan areas. One possible explanation may be the poor identification of EN schools by official survey data—it is possible that enough rural schools "unofficially" but successfully use the EN methodology, so that the coefficient on rural becomes positive. Around three quarters of rural schools are EN, and around 98% of EN are rural, which may mean that the estimation of the isolated effect of "rural" is unreliable given the poor identification of EN schools. Another possibility is a strong interplay between socioeconomic status and rural areas, in the sense that for students of a low socioeconomic status, rural schools provide better opportunities, while students of higher socioeconomic status do comparatively better in urban schools. After controlling for background, the effect of rural schools thus may become positive. It is possible to test this hypothesis by including an interaction term in the model. This interaction term indeed turns out to be significant and negative, confirming that rural schools outperform urban ones for low socioeconomic levels, while urban ones outperform rural ones for high socioeconomic levels (results not reported).

Finally, municipal governance has a significant and positive effect on learning outcomes, though the effect is small. Each point on the 100-points scale (which has a standard deviation of 13.5) improves test scores by only 0.16 to 0.29 points. An index score change of one standard deviation is thus only associated with an increase in the test score of between 2 and 4 points.

A look at the estimation results for the random part reveals that schools vary in their intercepts with an estimated standard deviation of between 28 and 49 points – the variation being larger for grade 3 results than for grade 5 results. Municipalities vary in their intercept with an estimated standard deviation of between 14 and 23 points. The smallest inter-municipality variation is found in civics test scores, while the largest one is found in 3rd grade mathematics. Finally, departments vary in their intercepts with a standard deviation of between 13 and 16 points. Within schools, the estimated student-specific standard deviations around the school means fall between 60 and 67 points; the estimated standard deviations are a little bit smaller in grade 3 than in grade 5.

The largest part of unexplained variance remains at the student-level, with a share of between 55% and 78% (depending on the grade and testing area). The smallest share of the unexplained variance is due to department-level effects (only around 3-4%), followed by municipality-level effects (4-7%). Unaccounted-for school-level factors are responsible for 14-35% of the remaining unexplained variance (i.e., the rest). A look at the *variance explained*, “R^2”[15], illustrates that: the

[15] “R^2” is set in quotation marks because the parameter is not the “R^2” as used in OLS analysis, but the difference between the variance in the null model and the respective model, expressed as a percentage of

variables in the model are able to explain under 10% of the variance in the null model across all levels (levels 1 to 4). The model fit is better when only considering the higher levels (L2-L4), or only the school-level (L2). Of the total variance that is due to differences across schools, the model explains between 12.1% (for grade 5 mathematics) and 26.4% (for civic competencies). Of all the variance that is *not* due to difference between students, the model explains between 15.2% (for mathematics grade 5) and 29.5% (for civic competencies).

4.3.1.3 Model diagnostics

The last step in the formulation of the random-intercept models are model diagnostics, with the aim to assess the error structure and check whether the model assumptions are valid. Figure 18 shows the empirical Bayes (EB) predictions for the random intercepts at all levels, together with the student-level residual (following Rabe-Hesketh and Skrondal 2012). The box plots depict that the variability in the unexplained variance is much higher within schools than between, within municipalities than between, and within departments than between. They also give a first idea about the distribution of the error terms: the mean seems to be zero as expected in all cases, and the quartiles seem distributed symmetrically. Furthermore, there are a number of extreme values, especially in the case of level-one residuals, as well as in the case of school-level intercepts.

For all random terms, normality is assumed. Hence, the corresponding empirical Bayes predictions should be normally distributed. Figure 19 (page 118) helps to assess whether this assumption holds for the example of language grade 3 test scores: it presents histograms showing the distribution of the respective error terms, with an added normal distribution curve for easier

the variance of the null model. This coefficient of determination is used as a measure of model fit in multilevel analysis (Raudenbush and Bryk 2002; Rabe-Hesketh and Skrondal 2012).

comparison (following Kim, Anderson, and Keller 2013). Results for other grades and test scores are similar. The histograms show an approximate normal distribution for student-level and department-level error terms. The normal distribution is somewhat less clear for school-level and municipality-level error terms, which seem to have a lighter-tailed distribution, and are slightly skewed to the right.

Finally, a key assumption is that the conditional level-one residuals are uncorrelated with the predicted exam score. Figure 20 helps to assess this assumption by plotting the standardized residuals against the fitted values (taking into account both the fixed parts and the random intercepts) (following Kim, Anderson, and Keller 2013). If the assumption of independence holds, the resulting plot should be a random collection of points without any specific patterns, and with approximately equal variance across all predicted values. Indeed, the scatter plots looks fairly random (though variance may be slightly smaller for higher predicted test scores).

All put together, the model assumptions seem tenable. Nevertheless, it may be desirable to adjust the model to remove the skew and light tails from the school- and municipality level error terms.

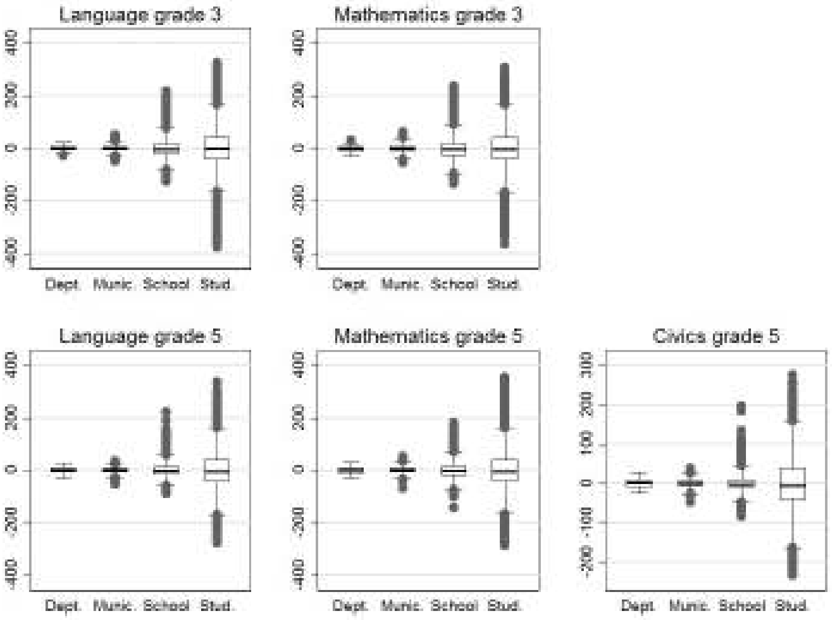


Figure 18 Box plots of empirical Bayes predictions for random intercepts at the department-level ($\tilde{\zeta}_d$), municipality-level ($\tilde{\zeta}_{md}$), and school-level ($\tilde{\zeta}_{jmd}$), and student-level residuals ($\tilde{\varepsilon}_{ijmd}$) for model RI3 (country-level study)

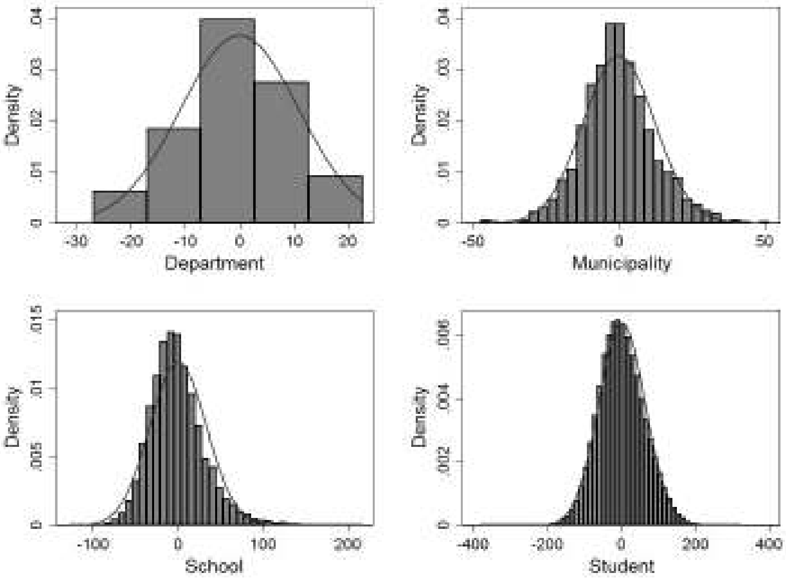


Figure 19 Histograms of empirical Bayes predictions for random intercepts at the department-level ($\tilde{\zeta}_d$), municipality-level ($\tilde{\zeta}_{md}$), and school-level ($\tilde{\zeta}_{jmd}$), and student-level residuals ($\tilde{\varepsilon}_{ijmd}$) for language grade 3 exam scores, model RI3 (country-level study)

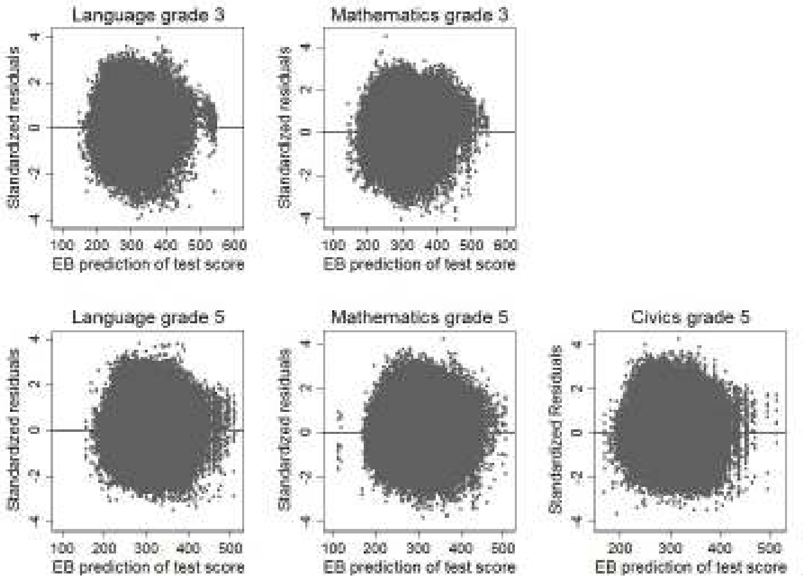


Figure 20 Scatter plots of empirical Bayes predictions for standardized level-one residuals ($\check{\varepsilon}_{ijmd}$) and fitted values ($\widehat{score}_{ijmd}$) for all grades and areas for model RI3 (country-level study)

4.3.2 Random-coefficient model

4.3.2.1 Development of the model

The final step in the macro model formulation is the relaxation of the assumption that regression lines are parallel between all schools, municipalities, or departments. By including random coefficients, it is possible to model different slopes for explanatory variables across the different levels, so that the explanatory variables are allowed to have different effects in different schools, municipalities, or departments. This may take care of the undesirable patterns in the school- and municipality-level error terms that were described in section 4.3.1.3.

Though it may be tempting to be liberal with the inclusion of random coefficients (a variation in slopes seems plausible for many variables!), the literature warns about such overzealousness

(Snijders and Bosker 2011; Rabe-Hesketh and Skrondal 2012). Because there is a variance parameter for each random effect and a covariance parameter for each pair of random effects, the complexity of the model increases rapidly with the inclusion of additional random effects, and the model may become unstable.[16] The recommendation is clear: "[R]andom slopes should be included only if strongly suggested by the subject-matter theory related to the application *and* if the data provide sufficient information" (Rabe-Hesketh and Skrondal 2012, 214).

From a theoretical standpoint, two random slopes are interesting. First, EN-effect may differ among departments. Given that the Secretaries of Education in the individual departments have a lot of decision-making power about departmental education policies and budgeting, whether or not the model can be successful will likely depend on the political priorities and resources made available in each department – for instance, on whether the Secretary provides all schools with the necessary learning guides, and whether EN teacher training workshops are regularly organized. Such a department-level random coefficient also includes the effect of a range of unobserved factors, such as the efforts of lobbying groups, influential "change makers", and personal commitment by key staff in the Secretaries. Thus, while there is no data available to directly control for all of these factors of influence, at least the presence of department-specific differences in the EN effect can be tested (and quantified) through the inclusion of a department-level random coefficient.

The second slope of interest, the effect of EN, might also vary by municipality, due to several reasons. First, municipalities have part of the decision-making power in the education sector.

[16] As Rabe-Hesketh and Skrondal (2012, 213) explain for the case of a two-level model: if there are k random slopes in a model, there are $((k+2)(k+1))/2\ +1$ parameters in the random part. For 2 random slopes, that gives a total of 7 random parameters.

While authority over budgets lies mainly with the Secretaries, other responsibilities lie with the municipalities. Second, the EN micro centers (the monthly gatherings of EN teachers to exchange experiences) are organized at the municipality-level, which means that a large part of the teacher-support network (and institutional learning) happens at this level. Third, municipalities are more homogenous with regard to unobserved environmental factors than departments. Factors such as the distance of the school to the next city or the local attitudes towards education will likely influence the efficiency of the EN model. For all these reasons, it seems likely that municipality-level effects are partly responsible for the success of the model.

The two random coefficients are added sequentially to the model in order to be able to test the contribution to the model of each of them. The first random slope to be added is the department-level coefficient of EN, given the concentration of political decision-making authority at that level. Model RC1 looks as follows:

Model RC1: $score_{ijmd} = \beta_0 + \beta_1 EN_{jmd} + \beta_2 male_{ijmd} + \beta_3 (male * EN)_{ijmd} + \beta_4 rural_{jmd} +$

$$\beta_5 private_{jmd} + \beta_6 NSE_{jmd} + \beta_7 (NSE * EN)_{jm} \quad + \beta_8 ethnic_{jmd} + \beta_9 conflict_{jmd} +$$

$$\beta_{10} morning_{jmd} + \beta_{11} afternoon_{jmd} + \beta_{12} governance_{md} + \zeta_{1d} EN_{jmd} + \xi_{ijmd}$$

It is an extension of model RI3, with the addition of the random slope $\zeta_{1d} EN_{jmd}$. ζ_{1d} is the department-level random coefficient, as indicated by the subscript $1d$. The rest of the error term remains unchanged (ξ_{ijmd} is still the composite of the four level-specific error terms).

Model RC2 includes, in addition to the department-level random coefficient, a municipality-level random coefficient, allowing the effect of EN to vary both between departments and between municipalities. The model is specified as:

Model RC2: $score_{ijmd} = \beta_0 + \beta_1 EN_{jmd} + \beta_2 male_{ijmd} + \beta_3(male * EN)_{ijmd} + \beta_4 rural_{jmd} + \beta_5 private_{jmd} + \beta_6 NSE_{jmd} + \beta_7(NSE * EN)_{jmd} + \beta_8 ethnic_{jmd} + \beta_9 conflict_{jmd} + \beta_{10} morning_{jmd} + \beta_{11} afternoon_{jmd} + \beta_{12} governance_{md} + \zeta_{1d} EN_{jmd} + \zeta_{2md} EN_{jmd} + \xi_{ijmd}$

The random slope coefficient ζ_{2md} was added to the model; the rest remains unchanged. For the estimation of the model, no assumptions are made regarding the covariance between the random coefficients and the random intercepts.

4.3.2.2 Results

Table 17 and Table 18 show the estimation results for models RC1 and RC2, respectively. Before interpreting the results, it is useful to take a look at the last row of each table, which contains the results for the likelihood-ratio tests. The test statistics are highly significant, indicating that RC1 fits the data better than RI3, and RC2 better than RC1. In other words: there is evidence that there are, in fact, different slopes for the effect of EN on learning outcomes for different departments and municipalities. As model RC2 was identified as the best-fitting model, the interpretation will focus on that model.

The estimated effect of EN is clearly and significantly positive for all grades and areas, ranging on average across departments, municipalities, and schools from 10.5 (for civic competencies) to 23.2 (for mathematics grade 3). The effect is stronger for grade 3- than for grade 5 students, which is not what would have been expected (longer exposure to the model should increase, not decrease, the advantage). There is only little support for the hypothesis that the EN model can help close gender gaps. The corresponding interaction is significant only for grade 5 mathematics and civic competencies, in the first case favoring girls, in the second case favoring boys. There is, however, strong evidence for the hypothesis that the EN model is particularly beneficial for children from disadvantaged backgrounds. The interaction term of EN and the socioeconomic level is clearly significant and negative. Table 16 summarizes the estimated difference in learning outcomes between EN and non-EN students at different socioeconomic levels based on model RC2. It shows that, other things being equal, students in EN schools can expect to do better than students in conventional schools for the socioeconomic levels NSE1 and NSE2, while students in schools with an average socioeconomic level NSE3 or NSE 4 generally do better if the school is not an EN.

Table 16 Estimated joint marginal effect of EN and socioeconomic status, based on model RC2 (country-level study)

	Language 3		**Language 5**		**Math 3**		**Math 5**		**Civics 5**	
	EN	**Non-EN**	**EN**	**Non-EN**	**EN**	**Non-EN**	**EN**	**Non-EN**	**EN**	**Non-EN**
NSE1	15.0	(base)	11.6	(base)	23.2	(base)	15.4	(base)	10.5	(base)
NSE2	19.4	14.8	17.1	18.9	25.5	12.0	19.7	16.5	15.8	15.7
NSE3	23.9	29.7	22.6	37.7	27.7	24.0	24.1	33.0	21.2	31.5
NSE4	28.4	44.5	28.1	56.6	30.0	36.0	28.4	49.5	26.5	47.2

Table 17 Results of the random-coefficient models RC1 (country-level study)

	Language				Mathematics				Civic Competencies	
	Grade 3		Grade 5		Grade 3		Grade 5		Grade 5	
n (students)		197,234		277,179		195,978		274,404		276,169
j (schools)		17,652		17,586		17,475		17,200		17,533
m (municipalities)		1,007		1,011		1,009		1,009		1,010
d (departments)		33		33		33		33		33
Fixed part:										
Escuela Nueva	13.10 ***	(2.73)	9.70 ***	(2.54)	20.58 ***	(3.00)	12.42 ***	(2.19)	8.65 ***	(2.02)
Male	-10.24 ***	(0.32)	-13.55 ***	(0.29)	0.59	(0.32)	7.22 ***	(0.30)	-18.61 ***	(0.30)
EN*Male	1.45	(0.93)	-0.27	(1.00)	-1.82	(1.11)	-3.03 ***	(0.88)	2.69 **	(0.86)
Rural	1.41	(1.52)	4.12 ***	(1.22)	4.28 *	(1.71)	4.54 ***	(1.34)	3.78 ***	(1.07)
Private	39.67 ***	(2.00)	26.12 ***	(1.57)	41.80 ***	(2.19)	27.12 ***	(1.74)	25.72 ***	(1.43)
Socioeconomic level	14.53 ***	(0.81)	18.55 ***	(0.64)	11.78 ***	(0.92)	16.30 ***	(0.73)	15.54 ***	(0.57)
EN*Socioecon. level	-9.33 ***	(1.61)	-12.22 ***	(1.33)	-7.76 ***	(1.74)	-10.10 ***	(1.46)	-9.22 ***	(1.23)
w/ ethnic students	-4.22 ***	(1.28)	-4.40 ***	(1.05)	-3.40 *	(1.46)	-4.50 ***	(1.15)	-3.26 ***	(0.91)
w/ conflict victims	-6.26 ***	(1.07)	-3.34 ***	(0.90)	-8.21 ***	(1.23)	-2.70 **	(0.96)	-2.93 ***	(0.81)
Morning session	-1.74	(1.69)	-4.82 ***	(1.41)	0.06	(1.90)	-3.27 *	(1.53)	-3.71 **	(1.25)
Afternoon session	-7.66 ***	(1.79)	-11.75 ***	(1.48)	-6.03 **	(1.96)	-10.26 ***	(1.56)	-8.25 ***	(1.29)
Governance Index	0.27 ***	(0.07)	0.24 ***	(0.06)	0.29 ***	(0.09)	0.27 ***	(0.08)	0.16 **	(0.06)
Grand mean	279.62 ***	(3.21)	279.88 ***	(3.15_	277.77 ***	(3.71)	269.58 ***	(3.69)	279.72 ***	(3.09)
Random part (sd):										
EN (department)	10.59	(2.27)	10.28	(1.92)	11.06	(2.34)	7.87	(1.73)	7.58	(1.58)
Intercept Department	11.56	(1.96)	13.53	(2.04)	13.75	(2.39)	16.01	(2.44)	13.88	(2.03)
Correlation (EN, dep.)	-0.11	(0.26)	0.08	(0.25)	-0.29	(0.24)	-0.02	(0.28)	-0.10	(0.26)
EN (municipality)										
Intercept (Municipality)	17.33	(0.78)	14.78	(0.66)	22.68	(0.90)	18.94	(0.77)	14.09	(0.62)
Correlation (EN, muni)										
Intercept (School)	42.96	(0.44)	32.60	(0.37)	49.00	(0.44)	36.68	(0.37)	27.52	(0.31)
Residual (Student)	59.46	(0.12)	67.37	(0.16)	61.58	(0.12)	63.93	(0.11)	64.54	(0.12)
LR χ^2 (RI3 vs. RC1)	36.37***		60.15***		38.57***		31.78***		37.84***	

Standard errors in parenthesis. ***: p≤0.001, **: p≤0.01, *: p≤0.05. LR-test statistics reported for calculations based on plausible value 1 (results of other plausible values not qualitatively different)

Table 18 Results of the random-coefficient models RC2 (country-level study)

	Language				**Mathematics**				**Civic Competencies**	
	Grade 3		**Grade 5**		**Grade 3**		**Grade 5**		**Grade 5**	
n (students)	197,234		277,179		195,978		274,404		276,169	
j (schools)	17,652		17,586		17,475		17,200		17,533	
m (municipalities)	1,007		1,011		1,009		1,009		1,010	
d (departments)	33		33		33		33		33	
Fixed part:										
Escuela Nueva	14.97 ***	(2.64)	11.64 ***	(2.46)	23.24 ***	(2.78)	15.38 ***	(2.22)	10.47 ***	(1.96)
Male	-10.24 ***	(0.32)	-13.55 ***	(0.29)	0.59	(0.32)	7.22 ***	(0.30)	-18.61 ***	(0.30)
EN*Male	1.49	(0.93)	-0.28	(1.01)	-1.84	(1.11)	-3.03 ***	(0.88)	2.72 **	(0.86)
Rural	1.36	(1.51)	3.88 ***	(1.21)	3.88 *	(1.68)	4.03 **	(1.31)	3.58 ***	(1.06)
Private	39.20 ***	(1.97)	25.52 ***	(1.56)	41.34 ***	(2.15)	26.42 ***	(1.71)	25.33 ***	(1.42)
Socioeconomic level	14.84 ***	(0.79)	18.85 ***	(0.64)	12.00 ***	(0.89)	16.51 ***	(0.72)	15.74 ***	(0.57)
EN*Socioecon. level	-10.37 ***	(1.69)	-13.38 ***	(1.41)	-9.76 ***	(1.86)	-12.17 ***	(1.57)	-10.40 ***	(1.30)
w/ ethnic students	-4.48 ***	(1.26)	-4.57 ***	(1.04)	-3.75 **	(1.43)	-4.85 ***	(1.13)	-3.49 ***	(0.91)
w/ conflict victims	-5.98 ***	(1.06)	-3.23 ***	(0.89)	-7.84 ***	(1.21)	-2.44 *	(0.96)	-2.86 ***	(0.81)
Morning session	-2.39	(1.70)	-5.53 ***	(1.42)	-0.74	(1.91)	-4.01 **	(1.53)	-4.06 ***	(1.27)
Afternoon session	-8.34 ***	(1.79)	-12.46 ***	(1.48)	-6.87 ***	(1.96)	-11.02 ***	(1.56)	-8.61 ***	(1.30)
Governance Index	0.24 ***	(0.07)	0.18 **	(0.06)	0.23 **	(0.08)	0.19 **	(0.07)	0.13 *	(0.06)
Grand mean	279.48 ***	(3.10)	280.00 ***	(3.13)	277.52 ***	(3.59)	269.09 ***	(3.63)	279.67 ***	(3.06)
Random part (sd):										
EN (department)	9.01	(2.33)	8.95	(2.07)	7.67	(2.46)	6.60	(1.97)	6.37	(1.68)
Intercept Department	11.12	(1.85)	13.56	(2.01)	13.48	(2.27)	16.02	(2.38)	13.82	(1.99)
Correlation (EN, dep.)	0.02	(0.30)	0.14	(0.29)	-0.13	(0.34)	0.11	(0.33)	-0.05	(0.30)
EN (municipality)	18.21	(1.89)	16.96	(1.65)	26.87	(1.99)	21.03	(1.69)	14.85	(1.39)
Intercept (Municipality)	12.78	(0.91)	12.22	(0.80)	15.97	(1.05)	14.24	(0.88)	11.82	(0.71)
Correlation (EN, muni)	0.14	(0.15)	-0.08	(0.14)	0.08	(0.11)	0.02	(0.12)	-0.01	(0.12)
Intercept (School)	42.60	(0.43)	32.25	(0.37)	48.20	(0.44)	36.14	(0.37)	27.24	(0.31)
Residual (Student)	59.46	(0.12)	67.36	(0.16)	61.58	(0.12)	63.92	(0.11)	64.53	(0.12)
LR χ^2 (RC1 vs. RC2)	96.78***		103.13***		206.24***		162.66***		109.57***	

Standard errors in parenthesis. ***: p≤0.001, **: p≤0.01, *: p≤0.05. LR-test statistics reported for calculations based on plausible value 1 (results of other plausible values not qualitatively different)

The positive c.p.-effect for rural schools is still present, though it is a little bit smaller than in the random intercept model. Students in private schools do considerably better on average than students from public schools, and students in schools with ethnic minorities or with conflict victims do, on average, worse. The full school day continues to benefit students, especially in comparison with the afternoon session (obviously, the school population in these sessions is very different). Finally, better municipal governance continues to have a positive effect on learning outcomes.

The interpretation of the random part of the model changes compared to the random intercept model. The reported standard deviations are now conditional on the school being an EN, and the different parameters must be interpreted jointly. The standard deviation of the random intercepts for the department- and municipality-levels is the estimated standard deviation for conventional schools ($EN_{jmd} = 0$). With the introduction of random slopes, the effect of EN now varies across departments and municipalities. This is most easily assessed graphically. For the case of grade 3 language scores, Figure 21 (page 129) shows stark differences in the fitted regression lines within and between departments (the results for other grades and areas are similar). Each of the plots represents one department, and within each plot, each line represents one municipality. Most importantly, the direction of the slopes is not uniform: though the estimated average effect of the EN model is positive (14.97 in the case of language grade 3, as per Table 18), the effect is negative for some municipalities. The graph also illustrates the presence of a department-level random coefficient: the individual plots differ from each other with regard to the average slope of the lines. At the department-level, two thirds of the departments have an average effect of EN in the range of 14.97±9.01 (i.e., between 5.96 and 23.98); at the municipality-level, 95% of the municipalities have an average effect of the EN model in the range of 14.97±1.96*18.21 (i.e., between -20.7 and 50.7). This is a very wide range of slopes, which strongly indicates that the success of the model depends on the specific circumstances in the department or municipality – presumably factors such as political support and resource availability, yet unfortunately there are no data to test this.

The average variance in exam scores for conventional schools and EN schools are presented in Table 19. The main result of the table is that the total variance in test scores is larger for EN schools than for conventional schools for all grades and areas.

Table 19 Estimated total variance in scores for students in different school types based on model RC2 (country-level study)

	Language		**Mathematics**		**Civics**
	Grade 3	**Grade 5**	**Grade 3**	**Grade 5**	**Grade 5**
Escuela Nueva	6121.21	6278.39	7371.23	6374.48	5485.05
Conventional Schools	5637.01	5910.66	6552.82	5851.79	5236.29

Two other interesting questions that can be answered based on the results from model RC2 are if and how mean test scores across municipalities or departments are correlated with differences between school models. For instance, a positive correlation between the random slope and the random intercept at the department-level would indicate that departments with larger mean test scores in conventional schools tend to have larger differences between EN schools and conventional schools. A negative correlation would suggest the opposite: Departments with large mean test scores in conventional schools tend to have smaller differences between the school types. Table 20 summarizes the estimates for correlations and covariances of the department-

and municipality-level intercepts and slopes. The table indicates that there is not much correlation between the random intercepts and the random slopes, and that the correlation that exists, while not clearly systematic, tends to be positive. That indicates that across departments or municipalities, the size of the effect of the EN model generally does not depend on the average effectiveness of conventional schools; though, if anything, the effect of EN is stronger in departments with larger mean test scores for conventional schools. In other words: EN schools are almost equally likely to be successful in municipalities or departments with low average exam scores as in departments with high average exam scores, with a potential slight advantage in the latter. This relationship is depicted in Figure 22 for civics (as an example for low correlation) and in Figure 23 for language grade 3 (as an example for a weak to moderate correlation at the municipality-level).

While it would be interesting to calculate intraclass-correlation coefficients, this is unfortunately not easily possible for random coefficient models. The reason is that the magnitude of the variance estimate depends on the scale and outcome of the underlying covariates. For the same reason, it does not make sense to compare the magnitude of random-intercept and random-slope variances (Rabe-Hesketh and Skrondal 2012, 191).

Table 20 Correlations and covariances for department- and municipality-level random effects (country-level study)

		Language		**Mathematics**		**Civics**
		Grade 3	**Grade 5**	**Grade 3**	**Grade 5**	**Grade 5**
Department	**Correlation**	0.02	0.14	-0.13	0.11	-0.05
	Covariance	1.99	16.67	-13.31	12.14	-4.11
Municipality	**Correlation**	0.14	-0.08	0.08	0.02	-0.01
	Covariance	33.74	-16.60	32.23	6.32	-2.01

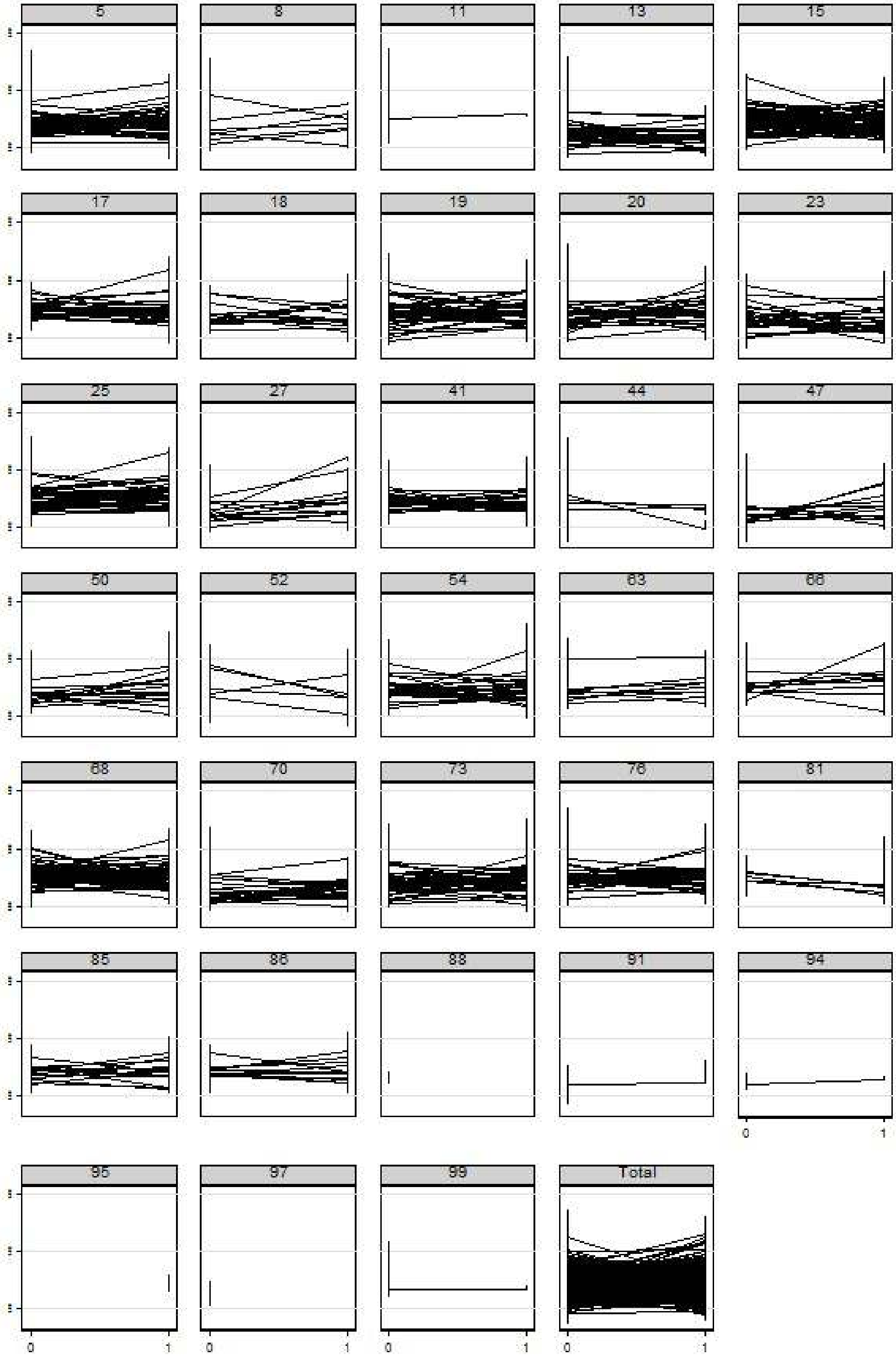


Figure 21 Spaghetti plots of empirical Bayes predictions of municipality-specific regression lines for model RC2 by department, language grade 3. Fitted municipality-mean on the left side of each panel for conventional schools and on the right side for Escuela Nueva schools (country-level study)

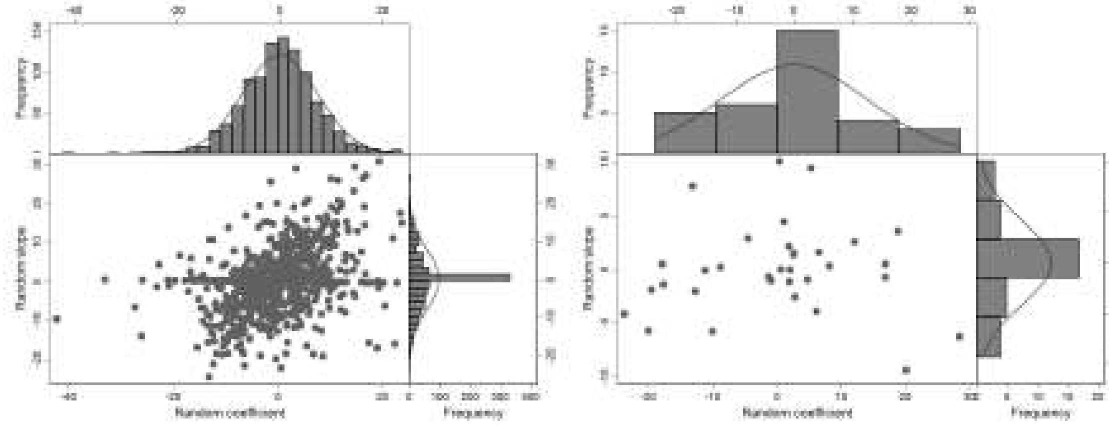


Figure 22 Correlation between random coefficient and random slope at the municipality-level (left panel) and department-level (right panel) for grade 5 civics scores (country-level study)

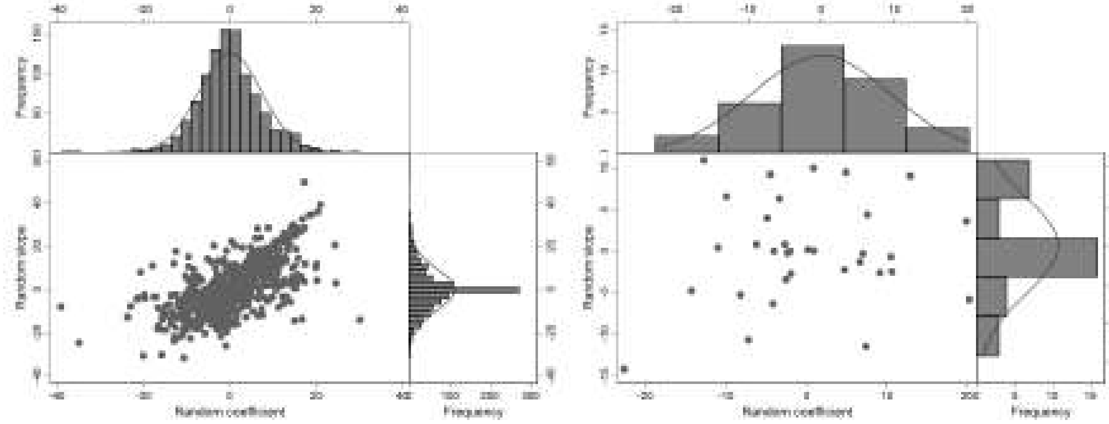


Figure 23 Correlation between random coefficient and random slope at the municipality-level (left panel) and department-level (right panel) for grade 3 language scores (country-level study)

4.3.2.3 Model diagnostics

The last step is to assess whether the model assumptions hold. As in section 4.3.1.3, this is done by plotting the error terms. For the case of the language grade 3 model, the empirical Bayes (EB) predictions for the random intercepts and slopes at all levels and the student-level residual are plotted in Figure 24 (the depicted variability in the intercepts is conditional on $\mathrm{EN} = 0$). The box plots on the left show that the variability in the unexplained variance is much higher within schools than between, and within municipalities than between. Variability in the random intercepts and slopes of departments and municipalities is very limited. The box plots also suggest that the means of all error terms are zero as expected, and the quartiles are distributed symmetrically. There remain a number of extreme values, especially in the case of level-one residuals. The right side of the graph shows histograms depicting the distribution of all error terms. The histograms suggest an approximate normal distribution for student-level residuals and random intercepts, but a large spike in the estimated random coefficients at zero at the municipality-level. This spike is due to the fact that almost a quarter of municipalities have only one type of school, which makes it impossible to obtain a municipality-level random coefficient on EN for these cases. The respective graphs for other grades and test areas look similar.

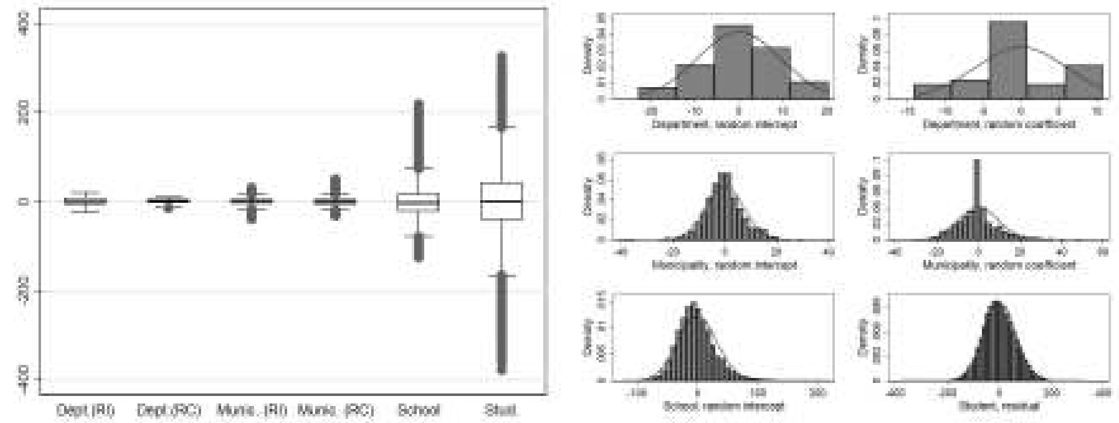


Figure 24 Box plots and histograms of empirical Bayes predictions for random intercepts and random slopes at the department-level ($\tilde{\zeta}_d$ and $\tilde{\zeta}_{1d}$) and at the municipality-level ($\tilde{\zeta}_{md}$ and $\tilde{\zeta}_{2md}$), random intercepts at the school-level ($\tilde{\zeta}_{jmd}$), and student-level residuals ($\tilde{\varepsilon}_{ijmd}$) for language grade 3 exam scores, based on model RC2. Estimates for variability of random intercepts are conditional on $EN = 0$ (country-level study)

Figure 25 presents the standardized residuals, plotted against the fitted values, for all grades and test areas. The plots seem random: there are no apparent patterns, and the variance seems approximately equal across all levels of predicted values.

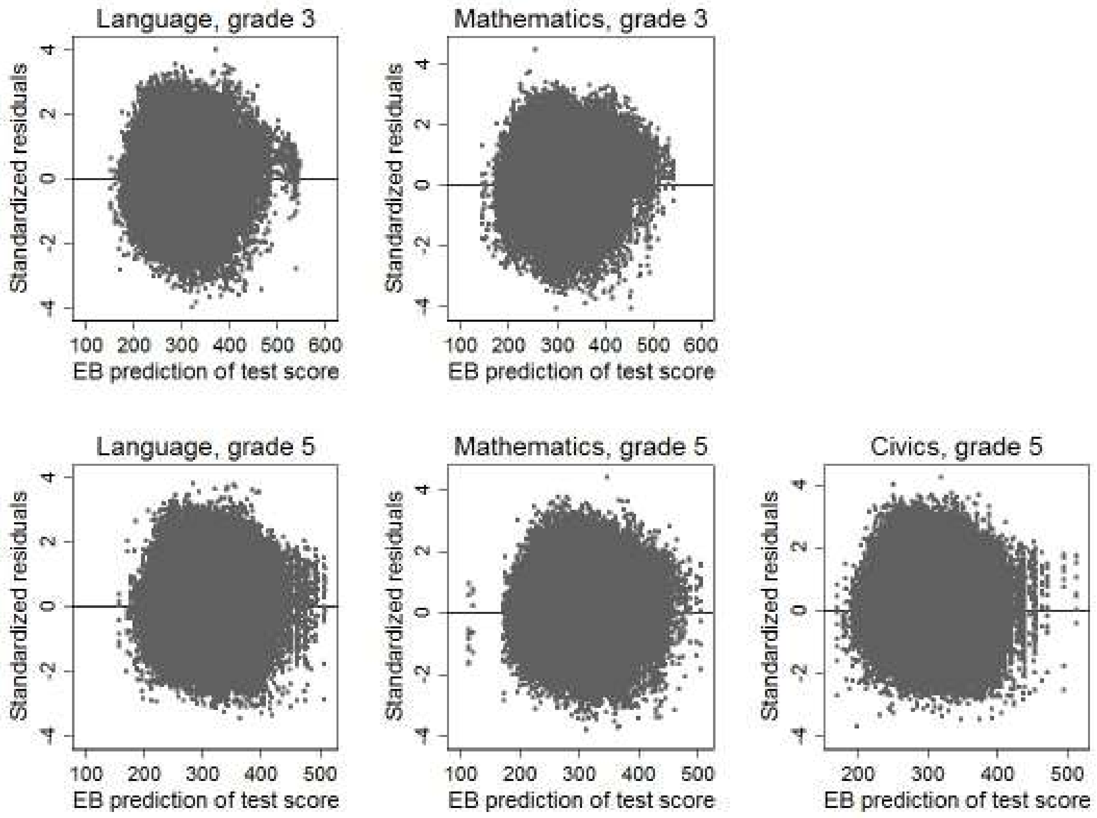


Figure 25 Scatter plots of empirical Bayes predictions for standardized level-one residuals ($\check{\varepsilon}_{ijmd}$) and fitted values ($\widehat{score}_{ijmd}$), based on model RC2 (country-level study)

4.3.3 Robustness

In order to make sure that the results are not driven by some isolated modeling decision or assumption, the model's robustness is tested in the following ways: First, the calculations are rerun using DANE's classification of rural/urban and public/private schools instead of ICFES's. As shown in Annex A, the classification for some schools differs between the two databases, and the robustness of the result to the classification source needs to be tested. Second, expenditure per student and homicide rates are re-introduced into the model to account for the large importance that the empirical literature gives to the former for general education outcomes, and to the latter (as a proxy for a peaceful environment) especially in the case of civic competencies.

Table 21 and Table 22 compare the results for the model based on DANE's classifications of rural and private school with the model based on ICFES' respective classification, for language and for mathematics and civic competencies, respectively. The change does not have any effect on the significance or substantial size of any of the effects (changes, if any, are within a fraction of a standard error). Cleary, the results are robust to the source of definition.

The second robustness test—the re-introduction of homicide rates and education expenditure—is more challenging computationally. The first choice for this robustness check was the introduction of municipality-level expenditure data as a control variable into the model RC2. However, for at least one of the underlying plausible values, the models for each grade/area did not converge with any model fitting method (maximum likelihood or restricted maximum likelihood) or maximization process (using matrix square roots or logarithms to parameterize variance components). As this data is only available for ten departments, the second choice was to remove the department-level random effect. Because the models still failed to converge, the third choice was the use of department-level expenditure data, which is available for 32 departments but was found to be non-significant in the random-intercept models (see section 1 of Annex B). Model RC3 thus expands RC2 by including data on homicides and education expenditure, but the calculations are based on a smaller sample due to data constraints in homicide rates. For the case of mathematics grade 5, model RC3 does not include homicide data, because that indicator prevented the model from converging.

Table 21 Robustness Analysis: Comparison of model RC2 based on classification data from ICFES and DANE, language (country-level study)

	Language Grade 3				Language Grade 5			
	ICFES		DANE		ICFES		DANE	
n (students)	197,234		197,234		277,179		277,179	
j (schools)	17,652		17,652		17,586		17,586	
m (municipalities)	1,007		1,007		1,011		1,011	
d (departments)	33		33		33		33	
Fixed part:								
Escuela Nueva	14.97***	(2.64)	14.87***	(2.64)	11.64***	(2.46)	11.53***	(2.46)
Male	-10.24***	(0.32)	-10.25***	(0.32)	-13.55***	(0.29)	-13.55***	(0.29)
EN*Male	1.49	(0.93)	1.49	(0.93)	-0.28	(1.01)	-0.28	(1.01)
Rural	1.36	(1.51)	1.56	(1.54)	3.88***	(1.21)	4.20***	(1.23)
Private	39.20***	(1.97)	39.13***	(1.97)	25.52***	(1.56)	25.61***	(1.56)
Socioec. level	14.84***	(0.79)	14.92***	(0.80)	18.85***	(0.64)	18.95***	(0.64)
EN*Socioec. level	-10.37***	(1.69)	-10.39***	(1.69)	-13.38***	(1.41)	-13.45***	(1.41)
w/ ethnic students	-4.48***	(1.26)	-4.56***	(1.26)	-4.57***	(1.04)	-4.63***	(1.04)
w/ conflict victims	-5.98***	(1.06)	-5.96***	(1.06)	-3.23***	(0.89)	-3.17***	(0.90)
Morning session	-2.39	(1.70)	-2.42	(1.70)	-5.53***	(1.42)	-5.52***	(1.42)
Afternoon session	-8.34***	(1.79)	-8.35***	(1.79)	-12.46***	(1.48)	-12.44***	(1.48)
Governance Index	0.24***	(0.07)	0.24***	(0.07)	0.18**	(0.06)	0.19**	(0.06)
Grand mean	279.48***	(3.10)	279.41***	(3.10)	280.00***	(3.13)	279.75***	(3.13)
Random part (sd):								
EN (department)	9.01	(2.33)	9.07	(2.34)	8.95	(2.07)	8.99	(2.07)
Intercept Dep.	11.12	(1.85)	11.09	(1.84)	13.56	(2.01)	13.53	(2.01)
Correl. (EN, dep.)	0.02	(0.30)	0.02	(0.30)	0.14	(0.29)	0.13	(0.29)
EN (municipality)	18.21	(1.89)	18.21	(1.89)	16.96	(1.65)	17.02	(1.65)
Intercept (Muni.)	12.78	(0.91)	12.79	(0.91)	12.22	(0.80)	12.21	(0.80)
Correl. (EN, muni)	0.14	(0.15)	0.14	(0.15)	-0.08	(0.14)	-0.08	(0.14)
Intercept (School)	42.60	(0.43)	42.62	(0.43)	32.25	(0.37)	32.25	(0.37)
Residual (Student)	59.46	(0.12)	59.45	(0.12)	67.36	(0.16)	67.36	(0.16)

Standard errors in parenthesis. ***: p≤0.001, **: p≤0.01, *: p≤0.05.

Table 22 Robustness Analysis: Comparison of model RC2 based on classification data from ICFES and DANE, mathematics and civic competencies (country-level study)

	Mathematics Grade 3		**Mathematics Grade 5**		**Civic Competencies Grade 5**	
	ICFES	**DANE**	**ICFES**	**DANE**	**ICFES**	**DANE**
n (students)	195,978	195,978	274,404	274,404	276,169	276,169
j (schools)	17,475	17,475	17,200	17,200	17,533	17,533
m (municipalities)	1,009	1,009	1,009	1,009	1,010	1,010
d (departments)	33	33	33	33	33	33
Fixed part:						
Escuela Nueva	23.24 *** (2.78)	22.97 *** (2.79)	15.38 *** (2.22)	15.24 *** (2.23)	10.47 *** (1.96)	10.26 *** (1.97)
Male	0.59 (0.32)	0.59 (0.32)	7.22 *** (0.30)	7.22 *** (0.30)	-18.61 *** (0.30)	-18.61 *** (0.30)
EN*Male	-1.84 (1.11)	-1.84 (1.11)	-3.03 *** (0.88)	-3.03 *** (0.88)	2.72 ** (0.86)	2.73 ** (0.86)
Rural	3.88 * (1.68)	4.71 ** (1.72)	4.03 ** (1.31)	4.42 *** (1.35)	3.58 *** (1.06)	4.14 *** (1.08)
Private	41.34 *** (2.15)	41.55 *** (2.15)	26.42 *** (1.71)	26.43 *** (1.72)	25.33 *** (1.42)	25.50 *** (1.42)
Socioeconomic level	12.00 *** (0.89)	12.22 *** (0.90)	16.51 *** (0.72)	16.64 *** (0.72)	15.74 *** (0.57)	15.89 *** (0.57)
EN*Socioecon. level	-9.76 *** (1.86)	-9.92 *** (1.87)	-12.17 *** (1.57)	-12.26 *** (1.58)	-10.40 *** (1.30)	-10.50 *** (1.30)
w/ ethnic students	-3.75 ** (1.43)	-3.81 ** (1.43)	-4.85 *** (1.13)	-4.92 *** (1.13)	-3.49 *** (0.91)	-3.54 *** (0.91)
w/ conflict victims	-7.84 *** (1.21)	-7.70 *** (1.21)	-2.44 * (0.96)	-2.38 * (0.96)	-2.86 *** (0.81)	-2.77 *** (0.81)
Morning session	-0.74 (1.91)	-0.70 (1.91)	-4.01 ** (1.53)	-4.00 ** (1.53)	-4.06 *** (1.27)	-4.02 ** (1.26)
Afternoon session	-6.87 *** (1.96)	-6.81 *** (1.96)	-11.02 *** (1.56)	-11.01 *** (1.56)	-8.61 *** (1.30)	-8.55 *** (1.30)
Governance Index	0.23 ** (0.08)	0.23 ** (0.08)	0.19 ** (0.07)	0.19 *** (0.07)	0.13 * (0.06)	0.13 * (0.06)
Grand mean	277.52 *** (3.59)	276.88 *** (3.60)	269.09 *** (3.63)	268.80 *** (3.64)	279.67 *** (3.06)	279.26 *** (3.06)
Random part (sd):						
EN (department)	7.67 (2.46)	7.70 (2.46)	6.60 (1.97)	6.65 (1.98)	6.37 (1.68)	6.42 (1.69)
Intercept Department	13.48 (2.27)	13.45 (2.27)	16.02 (2.38)	15.99 (2.38)	13.82 (1.99)	13.80 (1.99)
Correlation (EN, dep.)	-0.13 (0.34)	-0.13 (0.34)	0.11 (0.33)	0.11 (0.33)	-0.05 (0.30)	-0.05 (0.30)
EN (municipality)	26.87 (1.99)	26.88 (1.98)	21.03 (1.69)	21.11 (1.69)	14.85 (1.39)	14.91 (1.39)
Intercept (Muni.)	15.97 (1.05)	15.95 (1.05)	14.24 (0.88)	14.20 (0.88)	11.82 (0.71)	11.82 (0.71)
Correl. (EN, muni)	0.08 (0.11)	0.07 (0.11)	0.02 (0.12)	0.02 (0.12)	-0.01 (0.12)	-0.01 (0.12)
Intercept (School)	48.20 (0.44)	48.20 (0.44)	36.14 (0.37)	36.15 (0.37)	27.24 (0.31)	27.23 (0.31)
Residual (Student)	61.58 (0.12)	61.58 (0.12)	63.92 (0.11)	63.92 (0.11)	64.53 (0.12)	64.53 (0.12)

Standard errors in parenthesis. ***: p≤0.001, **: p≤0.01, *: p≤0.05.

Table 23 Robustness Analysis: Comparison of model RC2 with model including expenditure and homicides, language (country-level study)

	Language Grade 3				Language Grade 5			
	RC2		RC3		RC2		RC3	
n (students)	197,234		178,634		277,179		250,813	
j (schools)	17,652		14,176		17,586		14,094	
m (municipalities)	1,007		686		1,011		689	
d (departments)	33		32		33		32	
Fixed part:								
Escuela Nueva	14.97***	(2.64)	14.54***	(2.74)	11.64***	(2.46)	13.26***	(2.76)
Male	-10.24***	(0.32)	-10.12***	(0.34)	-13.55***	(0.29)	-13.53***	(0.32)
EN*Male	1.49	(0.93)	0.62	(1.07)	-0.28	(1.01)	-0.27	(1.03)
Rural	1.36	(1.51)	2.83	(1.62)	3.88***	(1.21)	5.82***	(1.34)
Private	39.20***	(1.97)	38.58***	(1.96)	25.52***	(1.56)	24.64***	(1.58)
Socioec. level	14.84***	(0.79)	15.57***	(0.82)	18.85***	(0.64)	19.93***	(0.67)
EN*Socioec. level	-10.37***	(1.69)	-11.92***	(1.91)	-13.38***	(1.41)	-14.46***	(1.62)
w/ ethnic students	-4.48***	(1.26)	-4.46***	(1.30)	-4.57***	(1.04)	-4.14***	(1.07)
w/ conflict victims	-5.98***	(1.06)	-5.85***	(1.12)	-3.23***	(0.89)	-3.43***	(0.96)
Morning session	-2.39	(1.70)	-1.36	(1.83)	-5.53***	(1.42)	-5.18***	(1.53)
Afternoon session	-8.34***	(1.79)	-7.39***	(1.90)	-12.46***	(1.48)	-12.18***	(1.60)
Governance Index	0.24***	(0.07)	0.26***	(0.08)	0.18**	(0.06)	0.25***	(0.07)
Homicides			0.00	(0.02)			-0.02	(0.02)
Expend. p. student			0.02*	(0.01)			0.02	(0.01)
Grand mean	279.48***	(3.10)	278.92***	(3.19)	280.00***	(3.13)	280.14***	(3.16)
Random part (sd):								
EN (department)	9.01	(2.33)	8.86	(2.51)	8.95	(2.07)	9.73	(2.15)
Intercept Dep.	11.12	(1.85)	8.02	(1.73)	13.56	(2.01)	11.20	(1.88)
Correl. (EN, dep.)	0.02	(0.30)	-0.05	(0.36)	0.14	(0.29)	0.00	(0.33)
EN (municipality)	18.21	(1.89)	16.87	(2.06)	16.96	(1.65)	15.72	(1.78)
Intercept (Muni.)	12.78	(0.91)	12.82	(0.99)	12.22	(0.80)	11.89	(0.86)
Correl. (EN, muni)	0.14	(0.15)	0.09	(0.16)	-0.08	(0.14)	-0.10	(0.16)
Intercept (School)	42.60	(0.43)	41.69	(0.44)	32.25	(0.37)	31.94	(0.39)
Residual (Student)	59.46	(0.12)	59.66	(0.13)	67.36	(0.16)	67.58	(0.17)

Standard errors in parenthesis. ***: p≤0.001, **: p≤0.01, *: p≤0.05.

Table 24 Robustness Analysis: Comparison of model RC2 with model including expenditure and homicides, mathematics and civic competencies (country-level study)

	Mathematics Grade 3				**Mathematics Grade 5**				**Civic Competencies Grade 5**			
	RC2		**RC3**		**RC2**		**RC3b**		**RC2**		**RC3**	
n (students)	195,978		177,602		274,404		274,404		276,169		250,016	
j (schools)	17,475		14,059		17,200		17,200		17,533		14,070	
m (municipalities)	1,009		687		1,009		1,009		1,010		688	
d (departments)	33		32		33		33		33		32	
Fixed part:												
Escuela Nueva	23.24 ***	(2.78)	24.12 ***	(3.08)	15.38 ***	(2.22)	15.33 ***	(2.22)	10.47 ***	(1.96)	11.87 ***	(2.17)
Male	0.59	(0.32)	0.84 *	(0.33)	7.22 ***	(0.30)	7.22 ***	(0.30)	-18.61 ***	(0.30)	-18.58 ***	(0.31)
EN*Male	-1.84	(1.11)	-2.10	(1.22)	-3.03 ***	(0.88)	-3.03 ***	(0.88)	2.72 **	(0.86)	2.86 **	(1.02)
Rural	3.88 *	(1.68)	4.68 **	(1.82)	4.03 **	(1.31)	4.04 **	(1.31)	3.58 ***	(1.06)	4.77 ***	(1.17)
Private	41.34 ***	(2.15)	41.18 ***	(2.16)	26.42 ***	(1.71)	26.41 ***	(1.71)	25.33 ***	(1.42)	24.86 ***	(1.44)
Socioeconomic level	12.00 ***	(0.89)	12.54 ***	(0.92)	16.51 ***	(0.72)	16.51 ***	(0.72)	15.74 ***	(0.57)	16.51 ***	(0.59)
EN*Socioecon. level	-9.76 ***	(1.86)	-11.53 ***	(2.11)	-12.17 ***	(1.57)	-12.17 ***	(1.57)	-10.40 ***	(1.30)	-11.63 ***	(1.49)
w/ ethnic students	-3.75 **	(1.43)	-2.85	(1.47)	-4.85 ***	(1.13)	-4.92 ***	(1.13)	-3.49 ***	(0.91)	-3.31 ***	(0.93)
w/ conflict victims	-7.84 ***	(1.21)	-7.55 ***	(1.28)	-2.44 *	(0.96)	-2.43 *	(0.96)	-2.86 ***	(0.81)	-3.00 ***	(0.84)
Morning session	-0.74	(1.91)	-0.67	(2.09)	-4.01 **	(1.53)	-4.03 **	(1.53)	-4.06 ***	(1.27)	-3.82 **	(1.43)
Afternoon session	-6.87 ***	(1.96)	-6.80 ***	(2.14)	-11.02 ***	(1.56)	-11.04 ***	(1.56)	-8.61 ***	(1.30)	-8.35 ***	(1.45)
Governance Index	0.23 **	(0.08)	0.30 ***	(0.09)	0.19 **	(0.07)	0.20 **	(0.07)	0.13 *	(0.06)	0.18 **	(0.06)
Homicides			0.01	(0.03)							-0.02	(0.02)
Expendit. per student			0.01	(0.01)			0.01	(0.01)			0.02 *	(0.01)
Grand mean	277.52 ***	(3.59)	276.79 ***	(3.81)	269.09 ***	(3.63)	270.17 ***	(3.59)	279.67 ***	(3.06)	280.75 ***	(2.98)
Random part (sd):												
EN (department)	7.67	(2.46)	8.01	(3.10)	6.60	(1.97)	6.60	(1.97)	6.37	(1.68)	6.34	(1.82)
Intercept Department	13.48	(2.27)	10.71	(2.42)	16.02	(2.38)	15.06	(2.47)	13.82	(1.99)	11.12	(1.83)
Correlation (EN, dep.)	-0.13	(0.34)	-0.09	(0.44)	0.11	(0.33)	0.12	(0.35)	-0.05	(0.30)	0.03	(0.37)
EN (municipality)	26.87	(1.99)	26.48	(2.18)	21.03	(1.69)	21.05	(1.70)	14.85	(1.39)	14.53	(1.48)
Intercept (Muni.)	15.97	(1.05)	16.53	(1.16)	14.24	(0.88)	14.27	(0.89)	11.82	(0.71)	11.50	(0.76)
Correl. (EN, muni)	0.08	(0.11)	-0.01	(0.12)	0.02	(0.12)	0.02	(0.12)	-0.01	(0.12)	-0.11	(0.12)
Intercept (School)	48.20	(0.44)	47.25	(0.49)	36.14	(0.37)	36.15	(0.37)	27.24	(0.31)	26.96	(0.32)
Residual (Student)	61.58	(0.12)	61.77	(0.13)	63.92	(0.11)	63.92	(0.11)	64.53	(0.12)	64.82	(0.12)

Standard errors in parenthesis. ***: p≤0.001, **: p≤0.01, *: p≤0.05.

Table 23 and Table 24 show the result for the robustness analysis, comparing model RC2 with model RC3. Again, model RC2 does not change with the changes in the specification. The coefficient on educational expenditure per student is positive but only marginally significant in two of the five models, suggesting that an increase in educational expenditures does not necessarily improve learning outcomes. The coefficient on homicides is not significant in any testing area or grade. Overall, the results discussed seem robust to changes in the model.

4.4 Discussion

The results of the country-level analysis are unambiguous: even after controlling for cluster-effects at the school-, municipality-, and department-levels, students in schools that are officially classified as EN do significantly better than students in other schools, other things being equal. In the final model, students in EN schools score, on average, between 11.6 and 23.2 points higher on the Pruebas SABER test. The estimated c.p. effects are summarized again in Table 25: A girl in an urban full-day public school of the lowest socioeconomic level, without any ethnic students or students who are victims of the conflict, in a municipality with an average governance index, can expect to score 279.5 points on the Pruebas SABER grade 3 language exam if her school does not identify as an EN, but 294.5 points if her school does identify as an EN. Thus, the null hypothesis that there is no difference in learning outcomes between students in EN and non-EN schools can be clearly rejected based on this analysis.

A variation of the alternative hypothesis stated that the effect of socioeconomic status on learning outcomes is smaller in EN schools than in conventional schools, other things being equal. The answer is affirmative: Given the highly significant and negative coefficient on the interaction term of EN and socioeconomic level, the null hypothesis of no difference can be rejected, in favor of the alternative hypothesis that EN is particularly beneficial for children from poor families. Table

16 on page 123 already summarized the estimated joint marginal effects of EN and socioeconomic background: In general, children in schools with an average socioeconomic level of NSE1 to NSE2 do better if their school is an EN, while children from the socioeconomic levels NSE3 and NSE4 do better in non-EN schools. In that sense, the school model helps to bridge the gaps between the groups. A note of caution here is that the socioeconomic level is measured at the school-level, not at the level of the individual student. It is possible that the positive effect exists only at the aggregate (school-) level, but not at the student-level: The average student in a school with a lower average socioeconomic level may be benefitting more from the model, but the effect need not be the same for students from a specific socioeconomic level in schools with a higher (or lower) average level. While there is no data to separate the effect of the socioeconomic background into a between-school and within-school component, the fact that the effect of EN is strongest in schools with an *average* socioeconomic level of NSE1 (i.e., the lowest) means that these schools really do cater to the most disadvantaged students.

Table 25 Expected exam scores in the Pruebas SABER based on model RC2 (country-level study)

	Language Grade 3	**Language Grade 5**	**Mathematics Grade 3**	**Mathematics Grade 5**	**Civics Grade 5**
Escuela Nueva	294.5	291.6	300.8	284.5	290.1
Difference	15.0	11.6	23.2	15.4	10.5
Non-Escuela Nueva	279.5	280.0	277.5	269.1	279.7
Standard deviation	75.2	79.7	77.5	76.8	75.2
Effect size	0.20	0.15	0.30	0.20	0.14

The third research question that was assessed in this chapter is whether the effect of the EN model differs by gender, the alternative hypothesis being that gender gaps are smaller in EN schools. The results suggest that this is only partially the case. The interaction term in the final model was only significant in two cases (5th grade mathematics, and civic competencies), in the former case favoring girls, and in the latter, boys. As far as the model results are concerned, the school model has thus some limited impact on existing differences between genders. In grade 5 mathematics, where boys tend to score higher, the model decreases the difference between boys and girls by improving girls' scores relative to boys'. In civic competencies, where girls tend to score higher, the same reduction in gender-differences takes place: Even though boys perform worse than girls in both school types, the difference is smaller in EN schools. However, this equalizing effect could not be found for language, or 3rd grade mathematics.

EN improves learning outcomes, especially for poor children; in some instances, it also helps to decrease the difference between genders. That being said, the question becomes whether or not the value added by the EN model is of any practical significance – that is, whether the magnitude of the effect is large enough to argue that the model can make an actual and noticeable difference. In order to facilitate interpretation, Table 25 (page 139) also includes information about the standard deviation in the exam results for each grade and testing area and the effect size based on the difference between the models and the standard deviation. The estimated effect sizes are anywhere between 0.14 (in the case of civic competencies) and 0.3 (in the case of 3rd grade mathematics). Effect sizes are slightly higher for grade 3 exams than for grade 5 exams.

These effects may not appear impressive by general standards (Cohen (1988) and Lipsey (1990) would classify them as small) – but when interpreting effect sizes it is important to take into account the context of the research. There is no universal guideline for judging the practical importance of an effect size. Instead, the nature of the intervention, the target population, and the outcome measure have to be taken into consideration (Hill et al. 2007).

The logical way to assess the practical importance of the effect of EN is to compare it with the effects estimated for other control variables. A good candidate is the socioeconomic level. Moving from a school of the lowest average socioeconomic level to a school with the second-lowest level is associated with an expected increase in exam results of between 57% and 191% of the effect of moving from a non-EN school to an EN school, depending on the area. In other words: for children from the poorest families, on average across testing areas and grades, being an EN school appears to make up for the disadvantage of as much as an entire socioeconomic level. This surely is a remarkable effect (Table 16 on page 123 helps to understand the differences at other socioeconomic levels).

A second helpful comparison is to contrast the estimated *ceteris paribus* effect of EN with the number of points necessary to move from one achievement level to the next. ICFES (2015a) defines for each grade and testing area a four-step scale of achievement levels. For instance, in the case of the language grade 3 exam, a score between 100 and 238 is considered insufficient; a score from 239 to 300 is considered poor; a score between 301 and 376 is considered satisfactory; and a score between 377 and 500 is advanced. The specific ranges for other grades and testing areas are different, but comparable in size. The marginal effect of the EN model, estimated between 10 and 23 points, is clearly not enough to bridge a whole achievement level – but the added value of the model is of noticeable size on that scale (up to a third of the distance from one achievement level to the next).

Put together, it seems thus safe to say that the effect of the EN model on learning outcomes is significant not only in a statistical, but also in a practical sense. However, the statistical model developed in this chapter can only explain a relatively small share of the differences in exam scores between students. By far the largest share is due to differences between students, which cannot be controlled for with the available database.

As discussed in section 3.3.1 (as well as section 1.2 of Annex A), it may be necessary to adjust the estimates for a potential sample selection bias. The database only contains schools that reported results correctly, which are likely to be schools of higher overall quality. Under the assumption of positive correlations between school quality and learning outcomes as well as between school quality and use of the EN model, the estimated effects are likely biased upwards. This is true for the main effect of the EN model as well as for its interaction with socioeconomic status, provided the latter is positively correlated with school quality. Unfortunately, the extent of this confounding cannot be established based on the available data.

The analysis based on the available secondary data cannot address the question of program implementation, yet the multilevel model does give some possible leads. First, the random coefficient model showed that the total variance in test scores is larger for EN schools than for conventional schools. This can certainly be attributable to more diverse student populations, but it also could be due to the fact that schools that are officially classified as EN are actually very heterogeneous in their teaching practices. As every school is observed only as being or not being an EN school, it is not possible to calculate the variance in the effect of the use of the EN model between schools with different characteristics.

Second, it is possible to estimate differences in the effect of EN across municipalities and departments, which was done by introducing a random coefficient. Indeed, the model confirmed differences in the effect of the model between these respective clusters, which may be due to differing levels of program implementation and support. Unfortunately, no data is available to test this hypothesis further, but Table 26 summarizes the estimates for the department-level random coefficients in all grades and testing areas. Highlighted in dark grey are estimates that are negative. While the picture is not uniform, there are some patterns: Departments tend to have negative effects in either all/most or none/almost none of the five grade-area combinations. The departments with consistent positive department-level slope estimates are Caldas, Cesar, Meta, Norte de Santander, Quindío, Santander, Tolima, and Valle del Cauca—which includes most of the departments of the coffee growing region where the model gained a foothold early on.

Table 26 Estimated random coefficient of EN by department. Negative effects marked in grey (country-level study)

	Language Grade 3	**Language Grade 5**	**Mathematics Grade 3**	**Mathematics Grade 5**	**Civics Grade 5**
Amazonas	0.76	-1.00	1.12	-1.13	-0.09
Antioquia	0.02	-4.48	-1.84	-0.67	-0.99
Arauca	-2.25	-4.46	-1.76	0.25	-2.59
Atlántico	0.41	0.38	0.09	-3.18	-1.45
Bogota	-1.33	1.59	-2.12	0.79	0.27
Bolivar	-14.20	-10.12	-4.40	-4.96	-5.77
Boyacá	-11.47	-3.02	-12.39	-6.73	-9.39
Caldas	9.01	8.82	3.93	1.90	2.52
Caquetá	-5.29	-0.21	2.52	0.24	2.88
Casanare	-2.79	-2.74	-1.14	-2.98	1.45
Cauca	-2.62	-1.34	-0.29	0.53	1.31
Cesar	9.18	7.95	7.63	7.95	7.69
Choco	10.93	0.43	2.71	-3.22	0.45
Cordoba	-10.69	-9.11	-6.82	-6.24	-5.80
Cundinamarca	-2.52	5.31	-4.74	0.83	0.47
Guainía	-0.31	-0.41	0.07	0.26	-0.74
Guaviare	0.15	-1.43	-0.36	-0.23	0.01
Huila	-3.47	-4.57	-4.60	-1.41	-3.93
La Guajira	-6.37	-3.77	-1.82	-1.42	-1.93
Magdalena	6.54	-10.61	6.24	-3.90	-4.18
Meta	3.90	3.83	4.88	0.02	4.42
N. de Santander	4.36	2.60	3.45	7.49	1.56
Nariño	-5.84	-6.43	-6.88	-0.52	-6.30
Putumayo	0.73	-3.90	-1.92	-0.92	-1.09
Quindío	9.40	17.08	3.50	7.84	9.39
Risaralda	-0.37	-1.54	-2.52	-1.50	-0.78
San Andres	-0.03	-0.79	0.15	-0.49	0.19
Santander	3.56	5.25	7.47	4.56	3.55
Sucre	-4.77	-5.66	0.10	-3.62	-2.05
Tolima	6.27	5.73	5.32	0.43	2.07
Valle del Cauca	9.91	17.17	4.68	11.11	10.05
Vaupes	-0.06	-0.24	0.86	-0.51	-0.04
Vichada	-0.74	-0.33	-1.11	-0.57	-1.18

Departments where the model has received the least political attention include, according to Fundación Escuela Nueva, Amazonas, Atlántico, Bolivar, and La Guajira; these tend to have negative random slope estimates. However, other departments where the model has generally been well supported in recent years (again according to Fundación Escuela Nueva: Cundinamarca, Boyacá, and several municipalities of Antioquia) turn out to have mixed or even negative department-level slopes (not necessarily negative *overall* EN slopes). All put together, there is some indication that department-level political support matters, but more research is necessary in order to test that hypothesis.

A crucial caveat of this chapter remains that the variable used to identify EN schools is known to be imprecise: Not all schools that are officially classified as EN actually use the model, and many of the schools that are officially classified as "not EN" use at least parts of it. That identification problem cannot be solved based on the available data. The next two chapters address that problem by first gathering evidence on the level of program implementation in one department, and then analyzing how this implementation affects learning outcomes.

5 Escuela Nueva Implementation in Quindío

The goal of this chapter is to find answers to the first set of research questions: To what extent have the Escuela Nueva model and its components been implemented in Colombia? How does the official classification compare to the actual status of implementation? How clearly are EN schools distinguishable from conventional schools? And what determines the extent of program implementation? The questions are answered based on quantitative and qualitative primary data collected in the department of Quindío, in Colombia's coffee growing region.

The chapter is structured as follows. First, the Escuela Nueva implementation index is introduced (section 5.1). Based on this index, program implementation in Quindío's rural schools is analyzed. The results show large variation in the way that the model is implemented, and a weak correlation between the official EN classification and the extent of program implementation (section 5.2). Subsequently, the results of the qualitative study are presented, organized by the indicator dimensions. The qualitative analysis confirms large differences in implementation, not just between schools, but also within schools with regard to different components of the model (section 5.3). Finally, section 5.4 discusses the results and draws conclusions with regard to the research questions.

5.1 The Escuela Nueva implementation index

In order to capture the differences in classroom practices across schools (which may or may not be formally classified as EN), an index of EN implementation was developed. The index is based on the conceptual framework of the school model that was already presented in section 1.2 (Figure 3 on page 13).

5.1.1 Construction of the index

The EN implementation index is a combination of a “teacher index” and a “student index”, both equally weighted. Each of these indices is based on the classroom practices as reported by the respective group. Using information from both teachers and students helps to triangulate the information, as there are specific error sources for both groups. For instance, teachers may be more likely to report their classroom practices as more closely following the guidelines than what is actually the case. Additionally, information about an aspect of program implementation is not always available from both teachers and students. On the one hand, this is because there are some aspects of the EN model (such as the organization of the teacher training workshops, or students’ regular participation in the student government) that can only be reliably answered by one group or the other. On the other hand, the student questionnaire had to be considerably shorter than the teacher questionnaire, given the reading level of the students who were asked to fill out the survey. The section on the statistical properties of the index (section 5.1.2) will compare the student index with the teacher index.

Table 27 shows the simplified structure of the implementation index. It is based on four (for students) or five (for teachers) dimensions. These are:

1. Teacher Training (teachers only)
2. Classroom Organization
3. School and Community
4. Learning Guides
5. Roles of Students.

Table 27 The Escuela Nueva Implementation Index

Dim.	Sub-dimension	Component
1 Teacher Training	1 Pre-service training	1 Received pre-service training
		2 Participation in Initiation Workshop
		3 Participation in Learning Guide Workshop
		4 Workshops followed EN methodology
		5 Feels that managed to put content into practice
	2 In-service training and support	1 Participates in micro centers / experience exchange
		2 Visits model schools
		3 Receives regular mentoring visits
2 Classroom setup	1 Learning corners	1 Set up in classroom
		2 Stocked with appropriate materials
		3 Stocked by teachers, students, community
		4 Are continually expanded
		5 Are set up for all subject areas
		6 Are often used
	2 Flexible Furniture	
	3 Classroom library	1 Set up in classroom
		2 Is often used
		3 Contains wide range of materials
	4 EN instruments	1 Exist
		2 Are visibly displayed in classroom
	5 Multigrade classrooms	
3 School & Community	1 Flat administrative structure	
	2 Parental involvement	1 Frequent contact with families
		2 Use of travelling journal
		3 Families participate in school work (application exercises)
	3 Community Involvement	1 Use of community map/croquis
		2 Use of community monograph
		3 Family book
		4 Celebration of achievement Day
4 Learning Guides	1 Teacher guides	1 Teacher has teacher's guide
		2 Uses guides regularly
	2 Student guides	1 Guides available in all subjects
		2 Guides frequently used
		3 One guide per student
	3 Proper use of guides by students	1 Complement guide with other materials
		2 Do activities in own note books
		3 Do not write in guide book
		4 Use alone, in pairs, and in groups
	4 Proper use of guides by teachers	1 Makes modifications to guides
		2 Holistic use/all activities
		3 Used to promote active learning
5 Roles of Students	1 Student-centered/active learning	1 Students work alone, in pairs, and groups
		2 Teachers as guides, not lecturers
		3 Flexible promotion
		4 Assistance self-reported
		5 Peer-to-peer tutoring
		6 Progress report
	2 School democracy and Shared responsibility for classroom	1 Student government in place
		2 Committees in place
		3 Shared responsibilities in classroom

Light grey: Component only in the teacher index; Dark grey: component only in the student index

Within each of these dimensions, there are different numbers of sub-dimensions, components, and sub-components, which consist of a varying number of indicators. A detailed overview of the index is provided in Annex C .

The index was calculated based on primary data that was collected in 78 schools in the Colombian department of Quindío between April and November of 2016. More information on the data collection process can be found in section 3.3.3.

The calculation of the index was done in several steps. First, at the level of the individual respondent, the indicators are defined as dichotomous variables taking the value 1 if the respective aspect of the model is implemented (according to the respondent), and 0 if it is not. In a few cases, partial scores for implementation are possible instead of a dichotomous coding. For instance, the indicator on the frequency of contact with parents takes the value 1 if there is reported contact every week, the value 0.5 if there is contact every month, and 0.25 if there is contact each term. While calculations were done separately for teachers and students, the indicator definitions parallel each other where possible and appropriate in order to maximize comparability between the students' and teachers' answers.

Second, the weighted sum of the indicators is calculated for each respondent for all dimensions, based on a multiple-level equal-weighting system: Within each dimension, all sub-dimensions are equally weighted; within each sub-dimension, all components are equally weighted; on so on, until the last level. This equal weighting assures that the weight of an indicator is not determined by the level of detail with which information on a particular aspect of implementation is available. Given that there are no *a priori* reasons to think that one aspect of the model is more important than the others—in fact, publications of Fundación Escuela Nueva always stress the holistic nature of the approach that "rethinks" primary education in its entirety (Colbert 2009; Colbert 2015;

Caballero Rojas 2009)—adopting equal weights for the major aspects of the model, and equal weights of all constituting elements at the lower levels, seems to be the most conservative approach to take.

Third, for each school and dimension, the mean score of all respondents of a group (i.e., all students or all teachers) is calculated as a simple average. The resulting number is multiplied by 100. For each school, there is thus a student dimension implementation score between 0 and 100 for each of the four dimensions in the student index, and a teacher dimension implementation score between 0 and 100 for each of the five dimensions in the teacher index.

Fourth, for each school and dimension, the average of the dimension implementation scores of teachers and students is calculated, resulting in a total implementation percentage for the respective dimension. In dimension 1, where information is available only for teachers, their score becomes the total implementation percentage. For the two cases where no teacher data is available, the percentage of dimension 1 is imputed so that the ratio of the implementation score in dimension 1 to the implementation score in the other dimensions is the same as it is, on average, in schools where teacher data is available. That way, the overall assessment relative to other schools is not changed.

Fifth, for each school the dimension implementation scores are averaged in order to obtain a total implementation percentage for the school. At the same time, school-level teacher and student implementation indices are calculated by taking the average dimension score across all respondents of the same type (i.e., the average of all teacher dimension implementation scores gives a school-level teacher implementation index, and the average of all student dimension implementation scores given a school-level student implementation index). Because of the equal weighting process described in step two, each of the indices can thus take theoretical values between 0 and 100.

Constructing the index in this way has several advantages. First, the index values fall per definition in a range between 0 and 100, which can be interpreted as the implementation percentage – both at the aggregate level, and for individual dimensions. Second, by considering both students' and teachers' responses, the approach makes it possible to analyze agreement and disagreement between the groups, and check whether results change if only one group's experience is considered. Third, the approach also allows for disaggregating the index scores (and their variation) into the different elements of the EN methodology.

5.1.2 Statistical properties

Table 27 summarizes key descriptive statistics for the implementation index, student index, and teacher index. The mean score is 62.1 on the overall index, and similar on the underlying group indices (65.3 for students, and 62.6 for teachers). The respective standard deviations are 11.9, 14.1, and 13.5. The index ranges from a minimum value of 22.5 to a maximum value of 83.8 (20.6 to 92.6 for students, and 19.8 to 85.5 for teachers). Note that the values of the EN implementation index are not the average of the student and teacher indices, because information from these groups was averaged at the level of the index dimensions, not at the level of totals.

Figure 26 shows how the scores are distributed. The histograms on the left panel show a fairly good overlap in the scores between respondent groups, and, as a result, the overall index. The right panel shows the student and teacher index histograms separated, together with a scatter plot that indicates how scores on the teacher and student index are correlated (each dot represents one school). The scatter plot shows that the agreement between the respondent groups within the same school is only moderate. This is also demonstrated by the correlation coefficient between the two parts of the index, which is 0.46 (p<0.001) and lies thus at the upper end of the "moderate" spectrum (Cohen 1988). However, the correlation between the student index and the overall index, as well as between the teacher index and the overall index, is with 0.77 (p<0.001) and 0.92 (p<0.001), respectively, high.

Table 28 Descriptive Statistics on Implementation Index and Student- and Teacher Index

	N	Mean	Std. Dev.	Min	Max
Implementation Index	78	62.1	11.9	22.5	83.8
Student Index	68	65.3	14.1	20.6	92.6
Teacher Index	76	62.6	13.5	19.8	85.5

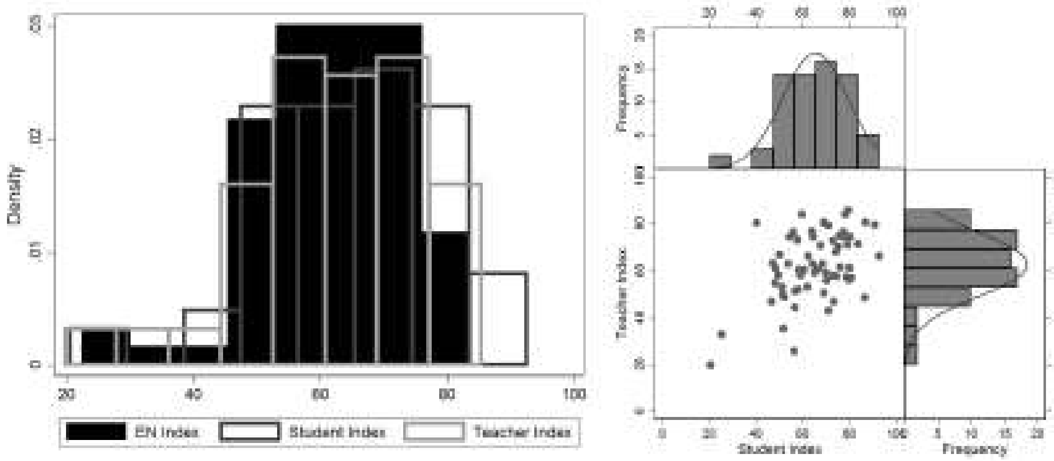


Figure 26 Histograms of EN Implementation index, Student- and Teacher Index (left side); variance and covariance of teacher and student indices (right side)

Figure 27 shows how student and teacher index are correlated for the individual dimensions (as there is only teacher information on dimension 1 [teacher training and support], this dimension is omitted). The graph shows that the level of agreement differs between the dimensions. The correlation between students' and teachers' assessments of classroom practices is high for dimension 2 (classroom organization), where r = 0.62 (p<0.001); moderate for dimensions 3 (school and community) and 5 (roles of students), with correlation coefficients of 0.44 (p<0.001) and 0.38 (p=0.002), respectively; and low for dimension 4 (learning guides), with r = 0.17 (p=0.161).

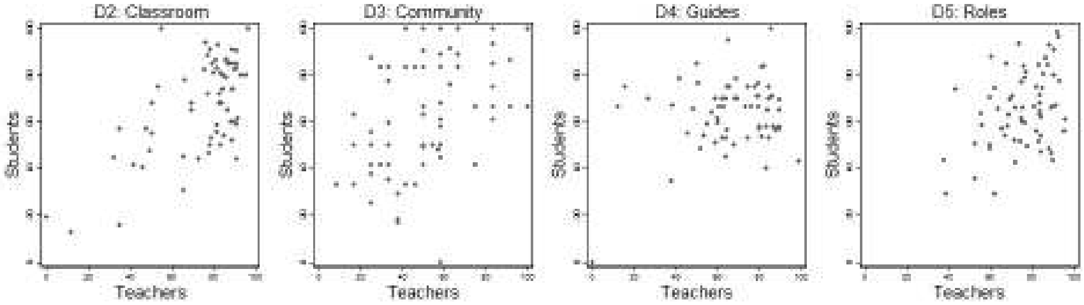


Figure 27 Correlation between student- and teacher indices for individual dimensions

A closer look at the original data reveals that this is not necessarily due to a faulty index. For some important questions, teachers' and students' responses do, in fact, contradict each other. For instance, in many schools, teachers say that there are learning guides for a given subject area, but students say there are not; and in almost half of the schools, students and teachers disagree about whether teachers are acting more as lecturers or as guides.[17]

Finally, Table 29 displays the correlation matrix for the index dimension scores and total index scores. It shows high correlations between the individual dimensions and the overall index (last row). It shows a mixed picture for the correlations between the indicator components, from very weak correlations (for instance, between dimension 1 [teacher training] and dimension 5 [roles of students], or dimension 3 [community relations] and dimension 4 [learning guides]) to strong correlations (between dimension 2 [classroom organization] and dimension 4 [learning guides] or 5 [roles of students]). This finding is positive for the overall index: It indicates that the individual index elements are well connected to the overall index, but each element captures a different concept from the ones captured by other elements.

Table 29 Correlation Matrix of EN index dimensions

	Dim. 1	**Dim. 2**	**Dim. 3**	**Dim. 4**	**Dim. 5**	**EN Index**
Dim. 1	1.00					
Dim. 2	0.24 *	1.00				
Dim. 3	0.28 *	0.39 ***	1.00			
Dim. 4	0.34 **	0.56 ***	0.16	1.00		
Dim. 5	0.18	0.65 ***	0.25 *	0.41 ***	1.00	
EN Index	0.68 ***	0.78 ***	0.65 ***	0.66 ***	0.65 ***	1.00

*** p≤0.001; ** p<0.01; * p<0.05

[17] The respective questions are question 58 in the teacher questionnaire and question 24 in the student questionnaire, which can both be found in the annex.

5.2 Escuela Nueva implementation in Quindío

Based on the implementation index, the first set of research questions relating to program implementation can be answered: To what extent is the EN model being implemented in Quindío's rural schools? Does the official classification of EN schools reflect observed differences in the application of the model's elements? And how different are EN classrooms from conventional classrooms? These questions will first be answered based on the overall implementation index, and then for the individual dimensions of the index. A note of caution right away: The final sample includes only four schools that are not officially classified as EN schools (and 72 that are). This, of course, limits the power of the comparisons.

5.2.1 Overall implementation

Table 30 shows the mean index score, standard deviation, range, standard error, and confidence interval first for all schools, and then separated into schools that are officially classified as EN or non-EN (DANE classification). The statistics are estimated using finite population correction to reflect that data is available for half of all rural primary schools. First, as the first row of Table 30 shows, while there is no school that implements 100% of the model's elements, there is also no school that does not implement any of them. Instead, all schools implement between 22% and 83% of the EN elements. This is not surprising, given that some key elements are the existence and use of classroom libraries, or regular contact with the parents or community – both may well be common practice in many schools. If the underlying population implementation index is normally distributed, two thirds of Quindío's rural schools implement between 55% and 74% of the elements of the EN model (mean ± one standard deviation), and 95% of schools implement anywhere between 39% and 85% of the elements (mean ± 1.96*standard deviation). This is a relatively high lower bound of the score, yet it also means that barely any school is truly faithful to the model in its entirety.

Second, the first row also shows a relatively narrow error range for the estimated mean implementation score across all rural primary schools of the department: The estimated mean lies within a 95% confidence interval of 60.3% to 63.3%. The error range is even smaller for EN schools, but considerable larger for non-EN schools, because only four schools in the sample are in that latter category. A t-test shows that the difference in the means of the two groups is borderline significant (t=2.05, p=0.044). Thus, with regard to the overall implementation of the model, the conclusion is that schools that are officially classified as EN do indeed have a higher implementation index – though the difference in the mean index score is smaller than a standard deviation.

Table 30 Estimated Implementation Index mean, standard deviation, range, standard error, and confidence interval for schools officially classified as EN and non-EN

	n	Mean	Std. Dev.	Min	Max	Std. Err. of mean	95% Conf. Interval of mean	
Combined	76	61.8	11.8	22.5	83.1	0.74	60.4	63.3
Conventional Schools	4	51.4	22.4	22.5	74.1	5.33	40.8	62.0
Escuela Nueva Schools	72	62.4	10.9	28.7	83.1	0.70	61.0	63.8

5.2.2 Implementation of program elements

The left panel of Figure 28 shows how well each of the five dimensions of the model is implemented, according to the teacher index, student index, and overall index. Across all schools, the dimensions of the EN model that are implemented to the largest extent are classroom organization and student roles. The dimension with clearly the lowest implementation scores is teacher training, followed by community relations. The right side of Figure 28 shows the distribution of implementation percentages for each dimension. The histograms show a wide spread of scores for each dimension, each with a distinct spike. The dimension implementation scores are far from normally distributed.

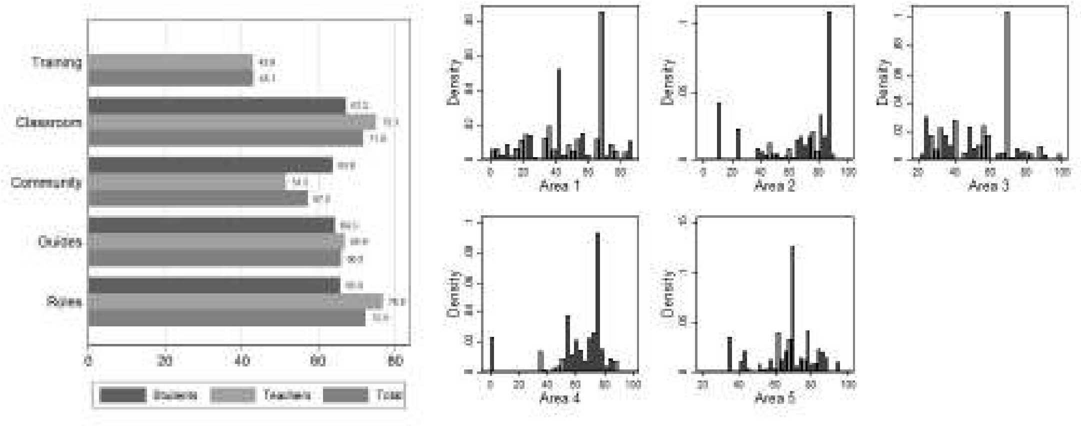


Figure 28 Percentage of program implementation by dimension (according to student-, teacher-, and overall index) (left panel); distribution of the implementation scores by dimension (right panel)

Figure 29 and Table 31 show that the biggest difference between EN and conventional schools lies in the organization of the classroom (dimension 2), where official EN schools implement 72.6% of the model's elements, while conventional schools only implement 51.2% (a difference of 21 percentage points). The second biggest difference can be found in community relations, where EN schools implement 57.5% of the model's elements, while conventional schools implement only 44.5% (a difference of 13 percentage points). In these two dimensions, the difference between conventional schools and EN schools is statistically significant, while this is not the case in the other dimensions.

Table 31 Percentage of average program implementation by dimension and school type

	Conventional Schools				**Escuela Nueva**				**Difference**	
	Mean	**Std. Err.**	**95% CI**		**Mean**	**Std. Err.**	**95% CI**			
D1	43.65	4.95	33.78	53.51	42.63	1.54	39.57	45.69	1.01	
D2	51.16	9.33	32.58	69.74	72.59	0.94	70.72	74.47	-21.4	*
D3	44.51	4.74	35.07	53.96	57.48	1.27	54.96	60.01	-13.0	**
D4	54.39	9.34	35.79	72.99	66.69	0.71	65.27	68.11	-12.3	
D5	63.38	5.45	52.51	74.24	72.70	0.79	71.13	74.27	-9.3	

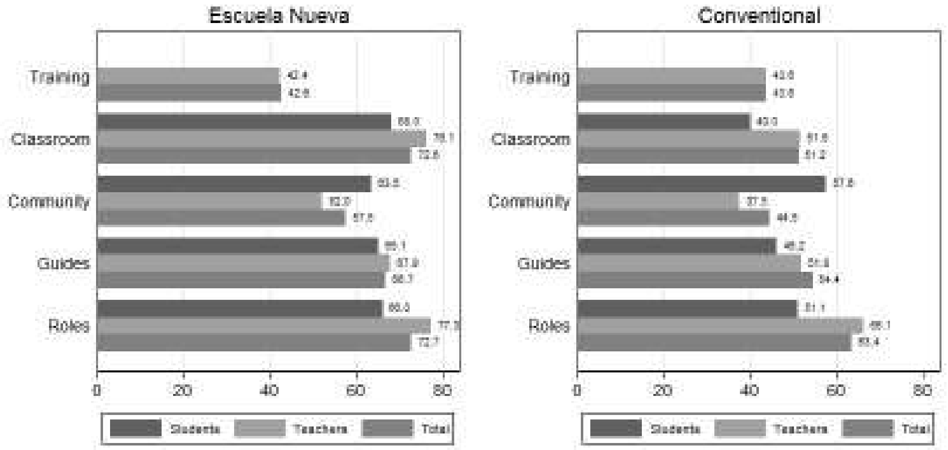


Figure 29 Percentage of average program implementation by dimension and school type (according to student index, teacher index, and overall index)

The conclusions based on the implementation of the individual elements of the model parallel those from the previous section: There is a wide range of classroom practices, which means that the EN model is far from being implemented homogenously. Schools that are officially classified as EN schools tend to implement more elements of the model, but the difference to conventional schools is not clear-cut – in fact, only for classroom organization and community relations is there a statistically significant difference in the implementation scores. This suggests that the official EN classifier should be seen, at best, as a "proxy" for EN implementation rather than as an objective identifier.

5.3 Qualitative evidence

The quantitative field work on EN implementation in Quindío was complemented by a qualitative component. Three schools with a high implementation index and four schools with a low implementation index were visited for additional, qualitative data collection through interviews and observations. The interviews were then transcribed and coded (more information on the process is provided in section 3.2.2). The results of the analysis are summarized here, sorted by indicator dimension. This is an exploratory, qualitative analysis, and the results should not be mistaken as a statistically representative description of EN implementation in the department. Rather, this section aims at providing context for the EN index, to triangulate the findings, and to obtain a better understanding of the range of implementation practices and the way in which the model is being used or not used.

5.3.1 Teacher training and support

The results of the quantitative analysis suggest low implementation in the dimension teacher training and support – a finding that is confirmed by the qualitative interviews. None of the younger teachers have participated in any of the week-long, topic-focused teacher training workshops designed to help teachers in the correct implementation of the program. The reason is simple: There has not been such a workshop in the department in many years, presumably due to a lack of funding. EN methods are also not taught mandatorily at university in the course of formal teacher training, though one of the interviewed teachers reported being instructed to use the method in the practice semester. As a result, it seems to be the norm that young, unexperienced teachers arrive at their first teaching assignment, typically at a distant rural school, without knowing much about the model, and scrambling to deal with the reality of a multigrade classroom. In some municipalities, teacher support networks (*micro centros*) are organized regularly. Teachers starting their career in such municipalities reported learning about the model in these informal monthly meetings. Other strategies include self-instruction through books and teacher's guides, and experience exchange with other teachers in the school. However, several of the interviewees described their first year as a teacher as a "lost year". Thus, in light of a lack of a unified training program, it is little surprising that model implementation in the classrooms differs widely, and that not all program elements may be interpreted, understood, or used in the same way across the department (or country).

In the ideal-typical EN model, teachers not only receive a specialized pre-service training, but career-long support through teacher micro centers and mentoring visits. According to the interviewed teachers, the implementation of this element varies considerably between municipalities. While the micro centers are active arenas of experience exchange and peer learning in some municipalities, they are not organized at all in others. Teachers who participate voiced that the gatherings are helpful. Only one teacher talked about receiving mentoring visits.

Overall, the qualitative interviews confirm the conclusion based on the implementation index: The teacher training and support component of the model is poorly implemented. The qualitative interviews add to this finding that this seems to be a source of stress for teachers, who feel poorly prepared for the context in which they are working, and as a result, at least at the beginning of their careers, sometimes experience the EN methodology as an additional burden rather than as a helpful tool set.

5.3.2 Classroom setup

The implementation index suggests that classroom setup is one of the better-implemented dimensions of the model. The dimension includes the use of multigrade teaching; the existence and proper use of learning corners; flexible classroom furniture; the existence and proper use of a classroom library; and the existence and proper use of a series of specific EN instruments, such as attendance self-control, a responsibility board, or a suggestion box.

Of the schools visited for qualitative interviews, only one (which has a low implementation index value and many more students than the other schools) did not implement multigrade teaching. All other schools were much smaller, so that two or more grades are co-taught by one teacher. Independent of the implementation level, teachers tended to describe multigrade teaching as a challenge (or at best a necessity), and not as a pedagogical tool. Many voiced frustrations about the lack of support for teachers in multigrade schools. While the EN model tries to provide that support, a lack of proper training in multigrade methods in general and EN methods in particular reduces the potential that this support can be properly taken advantage of.

Flexible classroom furniture was available in all schools. Two of the schools that were visited (one with a high index and one with a low index) do not use the specific EN furniture, trapezoid tables to facilitate combining the tables for group work. Though only a snapshot and not statistically representative, it was interesting to note that in all of the schools with a high index the tables were arranged for group work, while all but one of the low-index schools had arranged their tables in the conventional way (rows facing the front of the class).

All classrooms counted with a classroom library, though the interviews with teachers and students showed differences with regard to the available materials and usage: In this small sample, students in high-index schools said they use the library regularly, while students in low-index schools say they rarely use it. More interesting for the qualitative analysis is the way in which the libraries were described. One teacher, who reports making little use of the library, labelled it as "obsolete" due to the age of the available books. Additionally, most of the schools today dispose of computers or even tablets that are used as reference material instead of books. For instance, one teacher in a remote school in the mountains explained that the school is about to be connected to the Internet via satellite. In an area without landline network or stable cell phone coverage, this will surely provide incommensurable advantages compared to a small classroom library. Information on access and use of technology was not collected in the implementation survey, hence it is not possible to determine how such new technologies are being used compared to conventional classroom libraries.

Independent of the overall implementation level, none of the interviewees reported using the learning corners in the way suggested by Fundación Escuela Nueva. They were described as "dust traps" and as taking up a lot of valuable space. That being said, one teacher described using materials found in and around the school as didactic materials (oranges from a tree in the yard to teach division, stones from the yard to teach addition, etc.), and encouraging students to bring materials from their homes to school when needed. Thus, while learning corners may not be implemented in the way they were intended, the basic idea of using didactic materials originating in the students' familiar environment may be implemented in different ways.

Finally, an EN classroom should be equipped with a range of specific instruments that help support the overall mission of the model. For instance, the attendance self-control is a poster or board openly visible in the classroom, where each student marks each day whether they are present.

This instrument is meant to teach students self-responsibility and honesty. However, even in some of the schools where the quantitative data suggested that the instrument is being used, classroom observations and probing students' (and teachers') answers showed some adaptations to that concept – for instance, students taking turns each day in marking everybody's attendance, or self-reporting attendance into a little book. It is not clear to what extent these changes happen across the department, and at what point the core idea of the instrument is changed "too much" in order to still consider the instrument as being implemented. This is a recurring tension encountered at various occasions during the field work, and a similar story could be told about the other EN instruments (responsibility board, suggestion box, friendship mailboxes…): The EN model tries to be a flexible toolbox for teachers and to encourage them to adapt curriculum and teaching to the local context. However, it is not clear where the line between "local adaptation" and [partial] "implementation failure" should be drawn.

5.3.3 School and community

The third index dimension tries to capture the relationship between school, community, parents, and administrators. According to the student index, which only captures parental involvement, 64% of EN elements are implemented in this dimension across the department. According to the teacher index, which also includes the administrative structure and community involvement, only 52% of EN elements are implemented.

In the small group of schools that were visited for the qualitative study, the involvement of the community in general and of parents in particular was generally described as more active in schools with a high implementation index. That being said, the forms of parental involvement differ widely between the schools, and are not always tied to the "ideal-typical" EN instruments. For instance, schools do organize festivals or events, but none of the interviewees confirmed organizing a Día de Logros (Day of Achievements). Ways of community engagement include a parent's school, which by law is to be organized once per term yet has varying success; health days where doctors visit the school so that the entire community can come see them; or school improvement activities, such as planting a school garden or collecting money for classroom TV sets.

Other EN family engagement tools like community maps or family sheets are used to a varying degree as well. The traveling journal—a book that goes from family to family and everyone adds a story, recipe, prayer, joke, or similar contribution—is a beloved bonding tool in some schools, while another teacher described it as a *bobada* (nonsense). Even in schools where it is being used, the contributions seem to be sometimes added by the children themselves instead of by the family, which calls into question the purpose of the tool.

EN encourages parents to participate in their children's education through application exercises in the learning guides that ask children to do certain activities with their parents. The use of learning guides is discussed below. In this context, it is interesting to note that in the small qualitative sample, both teachers and students talked about the fact that parents are oftentimes too busy to do these exercises. Still, according to one teacher, if parents do engage in the activities, they are a strong tool to strengthen the relationship between parents and the school.

Overall, the (anecdotal) impression gained through the school observations and interviews is that ties with families are indeed stronger in schools with a high EN index. It is, however, not clear to what extent that is attributable to the EN methodology.

5.3.4 Learning guides

Learning guides are a central tool of the EN model. The implementation index captures the availability and proper use of both teacher guides and student guides, and suggests that on average, schools implement 66% of this component.

Both students and teachers had a lot to say about the learning guides. All of the schools visited use the guides in one way or another, though the extent and way of using them varies. Some teachers use them in some subject areas, but not in others; others use them for some of the exercises but complement them extensively with other books. In the interviews, challenges and benefits of the learning guides emerged.

On the positive side, learning guides were described as key to the development of self-control and a sense of self-responsibility. The guide books "put the students to think", as one teacher described it. Because of how the guide books are structured, students have to practice not only their reading, but also their reading *comprehension*. They would not be able to independently carry out the exercises if they did not understand what they are being asked to do.

The learning guides also promote group work. Students who described how their school day is structured spoke about coming together in groups to carry out the exercises in the learning guides; students describing a more lecture-based setting typically made references to additional materials brought in by the teacher.

Furthermore, the learning guides are key to achieving a flexible learning environment where each student can work at their own pace. Teachers and students described how the guide books enable students who fall behind for one reason or another to catch up with their peers.

That being said, the interviews also provided some hints that this flexibility may not always be achieved in practice. First, availability of materials is a challenge. In many schools, two or even three or four students share one learning guide, which makes it much harder if not impossible for each student to work at their own pace. Second, both teachers and students reported that the aim is usually to have all students in one grade (or even in different grades) work on the same topic at the same time. This can promote collaborative social skills and the concept of taking responsibility for each other when students who learn faster help students who learn slower to understand the material, so everyone can move at the same pace. However, there is a clear tension with the concept of self-paced learning. The reason for trying to have all students work at the same pace is also a practical one. In the words of one teacher, one "would have to turn into a wizard" (*volverse un mago*) in order to fully implement the concept of self-paced learning—it is hard enough to work with several grade levels at the same time, without also having to promote internal differentiation within each grade level.

Another challenge to the proper use of learning guides is the availability of *up-to-date* materials. Learning guides are often many years old and thus outdated. Some of the guides were described as too complex for the population they are designed for. For instance, one teacher said it was impossible to use the early-grade learning guides in Spanish, because the students lacked the required reading skills. Another teacher found the social sciences learning guides to be too dry and too full of long, complicated, and boring texts that could not engage the students. While these are only anecdotes, they show the range of challenges that teachers experience.

Overall, learning guides were both described as an EN element that works well, and as a source of challenges. In either case, though, the teachers shared how they adapt and complement the guides using materials they see as appropriate for the classroom, which in general terms is the role of the teacher that the EN model wants to promote. Again, the open question is where adaptation *within* the EN model ends.

5.3.5 Roles of students

The last dimension of the implementation index is the special role that the EN model assigns to students: The model wants to promote active, student-centered learning and gives students responsibility not just for their own learning, but also for their classrooms and schools. The quantitative analysis suggests that besides classroom organization, this is the dimension with the highest implementation percentage: Across schools, more than 72% of the model's elements in this dimension are being implemented.

Some of the aspects related to the roles of students have already been discussed above in different contexts (such as the use of group work). The brief discussion here will focus on other aspects of student-centered education and on school democracy and student governments.

Flexible promotion is an interesting element of the EN model in that it is a strategy to respond to a very real and common problem: School drop-out is known to be linked to grade repetition, which is a common practice around the world despite being remarkably inefficient. EN's response is to abolish grade repetition and instead promote students on a flexible schedule, that is, to make them graduate from one grade to the next in a given subject area whenever they fulfill all the requirements, regardless of the time in the academic year. The qualitative interviews suggested that this concept is not being adopted widely, and it is generally seen as impractical or even detrimental. First, as one teacher explained, there is an inherent contradiction in the school system: The Secretary of Education requires paperwork at the end of every school year documenting which students passed and which students failed; this is not compatible with the flexible, student- and subject-depending system of EN. Second, at least some teachers seem to perceive flexible promotion as detrimental to a student's academic career. One teacher described how this system covers up a student's lack of academic performance, which may lead to even larger problems later in the academic career. Another teacher said that students' parents would not accept not knowing whether or not their child passed a grade at the end of a year.[18] A third, practical problem is that different grades may be taught by different teachers. Several of the schools visited have two multigrade classrooms, each one combining two or more grades. In this setting, it is not clear how flexible promotion can work if, for instance, the grade 3 teacher is different from the grade 4 teacher and a student has passed the requirements of grade 3 in some subjects but not in others. Lastly, there is the problem that was already discussed above, that teachers feel overwhelmed teaching students in the same grade different content depending on how far along in the learning process they are. Only one of the interviewed teachers reported using the system and seeing it as appropriate for the context, though it was not clear how the school was dealing with the outlined practical problems.

The last important element is school democracy, promoted, among other techniques, through student governments and committees. All of the schools visited had a student government of some sort, though the way in which it was set up and worked differed. The institutions seem to be mostly formalities in some schools, where meetings and activities are rare or only held if called for by teachers or by students from higher grades (lower secondary, in schools with both levels). In other schools, students talked about how the conflict-solving committee helps to break up quarrels between students, how the red cross committee helps if a student injures herself, or how the learning guide committee makes sure that the learning guides are available when needed and

[18] As with many of the views gathered in the qualitative studies, whether or not this is actually the case is an empirical question that needs more focused research – the important message in this context is the reluctance of teachers to implement the method based on its perceived shortcomings.

otherwise properly stowed away. Thus, as in the other EN elements discussed above, the depth of implementation varied widely among schools that all formally implemented the elements.

5.3.6 Lessons learned

The qualitative interviews and school observations gave some deeper insight into the working of the model in the field. There are three central insights. First, there is the confirmation of the quantitative finding that EN implementation varies widely, not only among schools, but also among the different dimensions of the model within the same school.

Second, the qualitative data suggests a wide range of working definitions for some of the model's key elements. For one thing, that explains why the official self-definition of whether or not a school is an EN is not coherent. Another implication is that once not just the "if" but also the "how" and "why" of program implementation are considered, the distinction between EN schools and conventional schools becomes even fuzzier.

And third, the interviews gave some indirect insight into the driving factors for model implementation. Not surprisingly, the role of the teacher appears crucial: Whatever the political guidelines or practical limitations, the person who determines in the end what happens in the classroom is the teacher. If a teacher is, for instance, not convinced that classroom committees are serving their purpose, they will not be used in the proper way. That being said, the interviews also showed the importance of external (political) support. Providing a sufficient number of learning guides is essential for implementing the model as intended, yet it lies outside of the teacher's power to secure that. In the same way, teachers showed creativity in learning about the EN model even without a full series of targeted teacher workshops, yet it is possible that the model's elements would be understood and used in a different (hopefully, better) way if these workshops did exist as intended.

5.4 Discussion

Together, the evidence collected in the quantitative and the qualitative analysis helps to answer the research questions about program implementation.

The first two research hypotheses stated that EN model implementation is incomplete and irregular. These hypotheses are supported both by the quantitative and by the qualitative data: None of the schools in the sample implements over 90% of the model's elements, and the average school implements just over 60%. Implementation is irregular in the sense that schools may implement one dimension of the model with a high level of accuracy, while making many changes to other aspects of the model.

The data also supports the claim that there are major differences between the official classification and the actual degree of program implementation. Within each of the official groups (EN or not), the variance in observed implementation scores is very large. The range in the percentage of program implementation for official EN schools goes from 29% to 81%; the range for conventional schools goes from 23% to 74%. Thus, the official classifier is a poor predictor for classroom practices. While there is some correlation between being officially classified as EN and the program implementation index score, the difference is not large, and significant only for the overall index and two of the four model dimensions. Therefore, the official classification is a very rough approximation of actual classroom practices.

Finally, the third alternative hypothesis stated that decisions about program implementation are taken mostly on political grounds, and that actual on-the-ground implementation is driven mostly by individual teachers. The field work did not provide enough data for the first part of the hypothesis, but the analysis provides preliminary support for the second part. Teachers seem to have considerable leeway in the way that they structure their classes, and they decide which of the model’s elements to use based on what they see as fitting for their classes. However, more research is necessary in order to provide sufficient evidence to come to a conclusion regarding this last part of the hypothesis.

The evidence collected in this chapter will be used in the next chapter to answer the second set of research questions with regard to program outcomes.

6 Department-Level Analysis of Learning Outcomes

Chapter 4 analyzed learning outcomes in Colombia using the official Escuela Nueva classifier and concluded that the EN model indeed helps students improve their test scores. Chapter 5, however, showed that this official EN classifier does not reflect the reality of actual classroom practices very well: Schools officially classified as EN or as non-EN may be much closer in the methods they apply than what the dichotomous variable would suggest. Hence, the question arises whether the strong effect of the EN model on learning outcomes persists when using the EN implementation index instead of the official classifier to identify EN schools. Though it would be ideal to carry out that analysis for the country in its entirety, the analysis has to be confined to the small department of Quindío due to data constraints. Data for this is available for a total of 1,068 students (506 3rd graders, 562 5th graders) in 76 schools.[19]

Studying the effect of the EN model on learning outcomes in a small department has advantages and disadvantages. On the positive side is the fact that a statistically representative sample is easier obtained, and that there may be less heterogeneity in the data than in a larger department (due to, say, cultural differences within the department). On the negative side, the smaller sample also means less statistical power and limitations in the types of analysis that can be done. To illustrate the reduced statistical power, Table 32 shows the results of estimating the random intercept model RI3 from the country-level study based solely on data from Quindío (the random

[19] Data from two schools could not be matched with the DANE database (see section 4 of Annex A).

coefficient models could not converge or produce standard errors for the random effects). Although the effect of EN was estimated to be above-average in Quindío (see page 143), when restricting the sample to Quindío, the EN effect becomes non-significant – as do most other effects. Many variables also have to be omitted due to a lack of variance. Furthermore, the last part of the table suggests a lack of between-municipality variance. All of this does not necessarily indicate a lack of a positive effect of the EN model (or, for that matter, most other variables) in Quindío, but it does mean that changes to the modeling and estimation strategy will be necessary.

The rest of the chapter is structured as follows. In section 6.1, an exploratory data analysis provides descriptive statistics for outcome and control variables used in the department-level analysis and provides some first insights into how the variables are correlated. Next, an analysis of variance decomposes the total variance in learning outcomes into student-, school-, and municipality effects (section 6.2). Section 6.3 develops a multilevel model based on these findings, but concludes that the noise in the data limits the conclusions that can be drawn from the analysis; only some indicators, but not the EN implementation index, have significant coefficients. Section 6.4 offers an alternative approach to analyze the available data through survey data analysis, and suggests that schools that implement more parts of the EN model do have better learning outcomes on average. Section 6.5 discusses the results and concludes.

6.1 Exploratory data analysis

This section provides descriptive statistics and an exploratory analysis of outcome and predictor variables for the department-level study. Unless otherwise stated, all statistics are based on the sample of observations with non-missing values in the respective exam score, gender, socioeconomic level, and EN implementation index, as these are the main variables of interest that will be included in all models.

Table 32 Results of Model RI3, based on Quindío sample

	Language Grade 3		**Language Grade 5**		**Mathematics Grade 3**		**Mathematics Grade 5**		**Civic Competencies Grade 5**	
n (students)	245		369		246		312		371	
j (schools)	65		71		62		59		74	
m (municipality)	10		10		10		10		10	
Fixed part:										
Escuela Nueva	1.32	(101.49)	26.65	(79.16)	71.22	(160.71)	25.13	(90.22)	13.66	(84.36)
Male	-1.75	(12.41)	-42.23 ***	(11.56)	0.24	(13.97)	-13.91	(12.82)	-34.77 **	(11.57)
EN*Male	-21.03	(16.88)	19.28	(14.68)	6.05	(20.08)	28.89	(14.93)	-1.59	(14.14)
Rural	(omitted)		(omitted)		(omitted)		(omitted)		(omitted)	
Private	(omitted)		(omitted)		(omitted)		(omitted)		(omitted)	
Socioeconomic level	5.06	(54.49)	28.18	(41.11)	36.40	(86.81)	14.12	(52.64)	46.50	(44.96)
EN*Socioeconomic level	-6.58	(58.67)	-10.17	(45.11)	-19.95	(92.18)	-14.89	(56.67)	-9.46	(48.73)
w/ ethnic students	-28.13	(26.37)	-32.92	(21.52)	15.32	(40.79)	-16.90	(25.78)	-40.69	(21.19)
w/ conflict victims	-24.76	(15.81)	-11.42	(13.05)	13.56	(22.63)	-44.02 **	(14.42)	-15.33	(13.08)
Morning	(base)		(omitted)		(base)		(omitted)		(omitted)	
Afternoon	-24.30	(29.77)	(omitted)		-6.30	(29.01)	(omitted)		(omitted)	
Governance	0.64	(1.73)	-0.63	(1.73)	1.50	(2.55)	-1.74	(1.69)	-0.51	(1.47)
Grand mean	332.87 ***	(100.71)	287.41	(78.76)	214.97 ***	(160.06)	296.01 **	(89.00)	298.20 ***	(83.64)
Random part (sd):										
Municipality-level	0.000	(0.000)	0.91	(6.72)	0.000	(0.000)	0.000	(0.003)	0.000	(.)
School-level	39.84	(9.16)	28.81	(10.57)	66.26	(10.91)	35.47	(7.90)	32.59	(6.78)
Student-level	56.01	(3.42)	66.48	(2.84)	61.40	(3.91)	56.20	(3.64)	60.05	(2.72)

Standard errors in parenthesis. ***p≤0.001; ** p<0.01; * p<0.05

6.1.1 Plausible value outcomes

Table 33 shows summary statistics for exam scores in language, mathematics, and civic competencies, both for the individual plausible values and for the imputed estimate. In each grade and area, estimated mean and standard deviations are relatively similar in the aggregate, though the estimates based on the individual plausible values differ more from each other than was the case in the country-level study.

Table 33 Exploratory Data Analysis: Plausible values in department-level study

	Language				**Mathematics**				**Civic Competencies**	
	Grade 3		**Grade 5**		**Grade 3**		**Grade5**		**Grade 5**	
	mean	**sd**	**mean**	**sd**	**mean**	**sd**	**mean**	**sd**	**mean**	**sd**
PV1	292.25	68.88	297.35	75.50	297.88	87.56	283.20	70.16	304.96	74.02
PV2	296.54	73.34	300.61	77.93	296.41	85.56	278.97	70.83	307.13	74.73
PV3	292.92	70.69	298.68	77.26	298.39	84.28	281.51	71.41	306.92	73.41
PV4	295.90	69.32	297.41	76.11	298.53	86.66	279.63	68.75	305.75	77.27
PV5	293.65	67.46	299.21	77.51	297.31	85.48	280.14	70.76	306.01	75.15
Imputed	294.25	69.80	298.66	76.76	297.70	85.74	280.69	70.27	306.15	74.82
Skewness*	0.38		0.14		0.45		0.57		0.38	
Kurtosis*	3.06		2.68		2.95		3.34		2.88	
N	252		376		254		318		378	

* These statistics are provided exemplarily for plausible value 1.

The difference in the individual mean estimates is still small compared to the average within-person standard deviation of the five plausible values (which lies between 22.5 and 27.7, depending on the area and grade). The table also shows that the sample is smaller than the country-level sample by many orders: Data is available only from between 252 and 378 students in each area-grade combination.

The distribution of scores for the different plausible values is shown for the case of language exam scores in Figure 30 below. Because of the small sample size, the distribution is less smooth than for the country-sample. The histograms still show an approximate normal distribution, especially when aggregating across the individual plausible values. The picture is similar for mathematics and civic competencies.

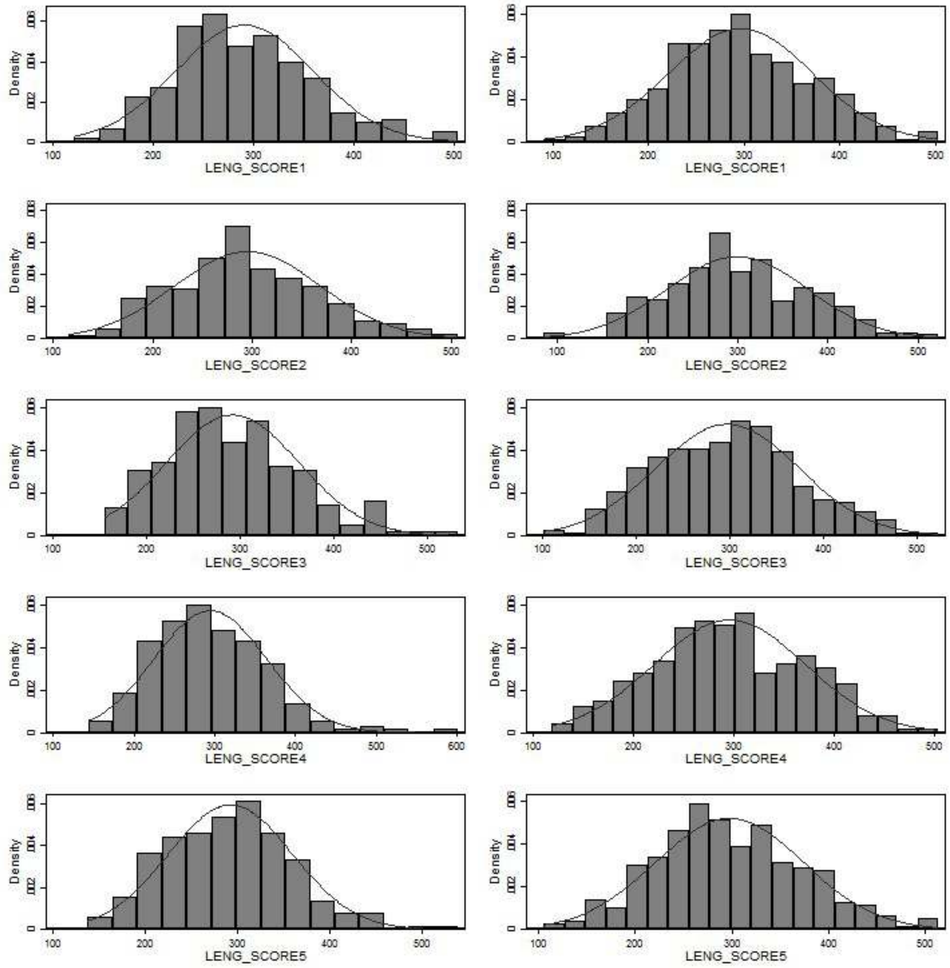


Figure 30 Histograms of Plausible Values in the sample for language exam scores, grade 3 (left) and grade 5 (right)

6.1.2 Predictors

Table 34 through Table 38 show the mean and standard deviation of the language, mathematics, and civic competencies scores for the different levels of explanatory variables, together with the percentage of students belonging to each category. The table also includes standard errors of the mean estimates to better assess the differences between the different categories (the standard errors are calculated using a finite population correction factor).

The tables suggest that students in EN schools have, on average, higher exam scores, and that students in schools with a high implementation index also score highest. However, the correlation between exam score and implementation index seems far from linear, with students in the second-lowest implementation quartile generally scoring higher than students in the third quartile. The correlation between gender and exam scores as well as between socioeconomic level and exam scores differs between the grades and testing areas, and is in most cases what is to be expected, though the relationship with socioeconomic level is also non-linear. Figure 31 through Figure 34 visualize the mean scores and standard deviations for these categories of the four key variables (EN, EN index, gender, and socioeconomic level).

Furthermore, the table suggests that some of the variables that were hypothesized to influence learning outcomes—and are thus used as control variables in the country-level study—have little systematic correlation at the level of the department of Quindío, among them homicides, governance, or education expenditure. Finally, there is barely any variance in the session type of Quindío's sampled primary schools (most of them are on a morning schedule).

Everything put together, the data indicates that the department-level model will look different than the country-level model, with some of the hypothesized explanatory variables potentially doing little to decrease unexplained variance. Additionally, the mean estimates are much less precise, which means that finding statistically significant regression coefficients is less likely.

Table 34 Exploratory Data Analysis for department-level study, Language Grade 3

Statistics presented at student-level		**%**	**Mean**	**Std. Err.**	**Std.Dev.**
Level-one Variables (Student-level, n = 252 students)					
Male	Male	55.56	285.82	5.60	68.03
	Female	44.44	304.80	6.35	70.49
Level-two Variables (School-level, j =66 schools)					
EN School (Official)	EN school	58.73	300.20	5.47	72.24
	Non-EN school	41.27	285.79	9.77	65.21
Implementation Index	Lowest quartile (Q1)	25.4	286.46	11.16	68.56
	Q2	26.19	315.68	8.96	76.08
	Q3	39.68	278.23	5.68	61.22
	Highest quartile (Q4)	8.73	325.48	11.65	62.09
Session type	Complete day	--	--	--	--
	Morning	96.73	293.89	5.71	70.22
	Afternoon	3.27	261.02	.	35.27
Socioeconomic level	NSE 1	10.71	304.48	10.08	77.34
	NSE 2	55.16	295.58	5.61	69.99
	NSE 3	31.35	283.97	12.35	64.80
	NSE 4	2.78	344.49	.	54.61
Ethnic population	With ethnic students	38.89	268.36	3.92	58.79
	Without ethnic students	61.11	310.73	5.14	71.20
Victims of conflict	Has conflict victim students	78.17	287.02	6.06	68.39
	Has no conflict victim students	21.83	320.15	6.97	68.48
Level-three variables (municipality-level, m = 10 municipalities)					
Homicide rate	Lowest quartile (Q1)	29.76	282.66	7.37	66.38
	Q2	46.03	288.30	8.74	64.95
	Q3	--	--	--	--
	Highest quartile (Q4)	24.21	319.84	8.80	75.79
Governance index	Lowest quartile (Q1)	30.95	303.50	5.51	65.20
	Q2	53.57	282.81	7.30	65.81
	Q3	--	--	--	--
	Highest quartile (Q4)	15.48	315.36	12.89	82.80
Educ. expenditure per student, municipal-level data*	Lowest quartile (Q1)	27.20	293.15	8.45	72.02
	Q2	24.27	303.33	9.22	72.69
	Q3	48.54	288.30	8.74	64.95
	Highest quartile (Q4)	--	--	--	--

* Data for this variable is only available for n=239 in j= 60, and m= 9.

"." Indicates that no standard error could be computed (data available from one cluster only)

Table 35 Exploratory Data Analysis for department-level study, Language Grade 5

Statistics presented at student-level		**%**	**Mean**	**Std.Err**	**Std.Dev.**
Level-one Variables (Student-level, n = 376 students)					
Male	Male	47.87	284.40	6.13	75.45
	Female	52.13	311.75	5.14	75.56
Level-two Variables (School-level, j = 72 schools)					
EN School (Official)	EN school	60.64	312.23	4.04	74.54
	Non-EN school	39.36	277.75	6.50	75.36
Implementation Index	Lowest quartile (Q1)	25.2	295.35	8.92	76.44
	Q2	24.8	316.16	5.70	73.29
	Q3	37.01	279.23	5.00	74.50
	Highest quartile (Q4)	12.99	336.89	10.73	69.68
Session type	Complete day	--	--	--	--
	Morning	100.00	297.18	4.86	76.16
	Afternoon	--	--	--	--
Socioeconomic level	NSE 1	10.11	295.05	10.04	73.71
	NSE 2	56.12	308.22	5.56	75.16
	NSE 3	31.91	278.45	5.87	74.81
	NSE 4	1.86	376.35	.	66.26
Ethnic population	With ethnic students	37.50	270.50	4.31	66.23
	Without ethnic students	62.50	315.55	3.71	77.66
Victims of conflict	Has conflict victim students	78.99	292.40	5.31	76.33
	Has no conflict victim students	21.01	322.18	5.93	73.61
Level-three variables (municipality-level, m = 10 municipalities)					
Homicide rate	Lowest quartile (Q1)	25.53	314.32	5.91	71.25
	Q2	51.06	282.50	5.35	74.38
	Q3	--	--	--	--
	Highest quartile (Q4)	23.40	316.81	6.80	79.82
Governance index	Lowest quartile (Q1)	25.27	313.14	7.03	75.53
	Q2	61.17	289.11	5.84	76.71
	Q3	--	--	--	--
	Highest quartile (Q4)	13.56	314.73	8.40	71.84
Educ. expenditure per student, municipal-level data*	Lowest quartile (Q1)	33.15	307.55	6.08	76.51
	Q2	20.06	329.78	6.64	71.79
	Q3	46.80	277.99	5.46	73.69
	Highest quartile (Q4)	--	--	--	--

* Data for this variable is only available for n= 359 in j=76, and m=9.

"." Indicates that no standard error could be computed (data available from one cluster only)

Table 36 Exploratory Data Analysis for department-level study, Mathematics Grade 3

Statistics presented at student-level		**%**	**Mean**	**Std.Err.**	**Std.Dev.**
Level-one Variables (Student-level, n = 254 students)					
Male	Male	61.81	299.63	7.89	86.40
	Female	39.19	294.59	9.19	84.50
Level-two Variables (School-level, j = 63 schools)					
EN School (Official)	EN school	59.45	307.29	5.64	86.91
	Non-EN school	40.55	283.65	17.08	81.92
Implementation Index	Lowest quartile (Q1)	25.53	307.44	13.71	72.21
	Q2	25.27	312.73	8.84	91.18
	Q3	38.83	261.08	6.62	78.72
	Highest quartile (Q4)	10.37	354.45	12.76	72.18
Session type	Complete day	--	--	--	--
	Morning	96.34	296.26	8.27	86.58
	Afternoon	3.66	278.00	.	278.00
Socioeconomic level	NSE 1	10.24	300.94	14.61	92.53
	NSE 2	56.69	303.98	5.55	81.29
	NSE 3	29.92	277.87	21.95	87.72
	NSE 4	3.15	362.52	.	67.33
Ethnic population	With ethnic students	36.61	261.91	7.20	72.25
	Without ethnic students	63.39	318.38	5.95	86.05
Victims of conflict	Has conflict victim students	81.10	294.23	9.34	84.33
	Has no conflict victim students	18.90	312.61	9.42	89.89
Level-three variables (municipality-level, m = 10 municipalities)					
Homicide rate	Lowest quartile (Q1)	29.53	295.39	8.08	77.25
	Q2	44.09	282.10	14.14	79.64
	Q3	3.54	290.31	24.79	104.35
	Highest quartile (Q4)	22.83	331.98	9.13	93.96
Governance index	Lowest quartile (Q1)	27.95	305.55	6.50	77.92
	Q2	57.48	282.24	10.83	78.42
	Q3	--	--	--	--
	Highest quartile (Q4)	14.57	343.64	14.72	106.41
Educ. expenditure per student, municipal-level data*	Lowest quartile (Q1)	26.64	292.46	9.16	86.96
	Q2	27.46	325.36	8.47	82.92
	Q3	45.90	282.10	14.14	79.64
	Highest quartile (Q4)	--	--	--	--

* Data for this variable is only available for n=244 in j=59, and m=9.

"." Indicates that no standard error could be computed (data available from one cluster only)

Table 37 Exploratory Data Analysis for department-level study, Mathematics Grade 5

Statistics presented at student-level		**%**	**Mean**	**Std.Err.**	**Std.Dev.**
Level-one Variables (Student-level, n = 318 students)					
Male	Male	48.74	284.74	7.40	74.88
	Female	51.26	276.84	5.37	65.29
Level-two Variables (School-level, j = 60 schools)					
EN School (Official)	EN school	63.84	294.83	4.70	71.72
	Non-EN school	36.16	255.72	5.04	59.88
Implementation Index	Lowest quartile (Q1)	26.1	271.74	6.62	62.96
	Q2	24.53	300.93	8.49	82.07
	Q3	39.94	266.52	6.50	61.95
	Highest quartile (Q4)	9.43	312.83	11.80	66.26
Session type	Complete day	--	--	--	--
	Morning	100.00	279.13	5.43	69.51
	Afternoon	--	--	--	--
Socioeconomic level	NSE 1	12.26	312.07	9.30	82.03
	NSE 2	57.86	284.99	5.31	68.74
	NSE 3	27.99	252.59	.	54.38
	NSE 4	1.89	361.58	.	58.25
Ethnic population	With ethnic students	42.14	254.45	2.50	54.61
	Without ethnic students	57.86	299.80	5.42	74.09
Victims of conflict	Has conflict victim students	79.25	268.05	4.66	63.50
	Has no conflict victim students	20.75	328.96	6.89	73.77
Level-three variables (municipality-level, m = 10 municipalities)					
Homicide rate	Lowest quartile (Q1)	26.1	301.67	7.52	73.75
	Q2	48.74	257.18	3.43	57.48
	Q3	3.77	298.66	17.74	63.58
	Highest quartile (Q4)	21.38	305.49	8.11	75.51
Governance index	Lowest quartile (Q1)	27.99	308.42	6.96	73.15
	Q2	57.86	263.27	4.73	62.06
	Q3	--	--	--	--
	Highest quartile (Q4)	14.15	297.06	9.44	74.02
Educ. expenditure per student, municipal-level data*	Lowest quartile (Q1)	25.97	297.44	7.11	72.15
	Q2	31.17	303.15	6.15	69.16
	Q3	42.86	253.81	3.25	57.35
	Highest quartile (Q4)	--	--	--	--

* Data for this variable is only available for n=308 in j=56, and m=9.

"." Indicates that no standard error could be computed (data available from one cluster only)

Table 38 Exploratory Data Analysis for department-level study, Civic Competencies Grade 5

Statistics presented at student-level		%	Mean	Std.Err.	Std.Dev.
Level-one Variables (Student-level, n = 378 students)					
Male	Male	49.74	290.30	7.97	73.96
	Female	50.26	321.84	4.96	72.29
Level-two Variables (School-level, j = 75 schools)					
EN School (Official)	EN school	62.43	311.33	3.99	73.37
	Non-EN school	37.57	297.55	14.12	76.36
Implementation Index	Lowest quartile (Q1)	25.13	316.72	14.76	83.89
	Q2	26.19	316.31	6.30	73.62
	Q3	37.3	282.59	4.15	65.00
	Highest quartile (Q4)	11.38	336.69	7.69	63.52
Session type	Complete day	--	--	--	--
	Morning	100.00	305.28	5.86	74.67
	Afternoon	--	--	--	--
Socioeconomic level	NSE 1	10.58	289.19	8.62	66.55
	NSE 2	56.88	310.68	5.00	73.69
	NSE 3	30.69	300.83	18.64	77.86
	NSE 4	1.85	352.38	.	66.69
Ethnic population	With ethnic students	36.77	275.24	2.69	62.16
	Without ethnic students	63.23	324.13	5.12	75.66
Victims of conflict	Has conflict victim students	78.57	300.80	6.79	74.68
	Has no conflict victim students	21.43	325.78	5.02	71.98
Level-three variables (municipality-level, m = 10 municipalities)					
Homicide rate	Lowest quartile (Q1)	26.19	312.96	5.18	67.17
	Q2	49.21	297.36	10.08	74.43
	Q3	--	--	--	--
	Highest quartile (Q4)	24.60	316.50	6.29	80.90
Governance index	Lowest quartile (Q1)	25.93	308.38	5.31	67.97
	Q2	60.05	304.57	9.28	77.60
	Q3	--	--	--	--
	Highest quartile (Q4)	14.02	308.80	9.68	74.40
Educ. expenditure per student, municipal-level data*	Lowest quartile (Q1)	25.69	309.48	8.19	85.05
	Q2	29.83	308.31	5.41	67.50
	Q3	44.48	299.01	11.78	73.17
	Highest quartile (Q4)	--	--	--	--

* Data for this variable is only available for n= 362 in j=69, and m=9.

"." Indicates that no standard error could be computed (data available from one cluster only)

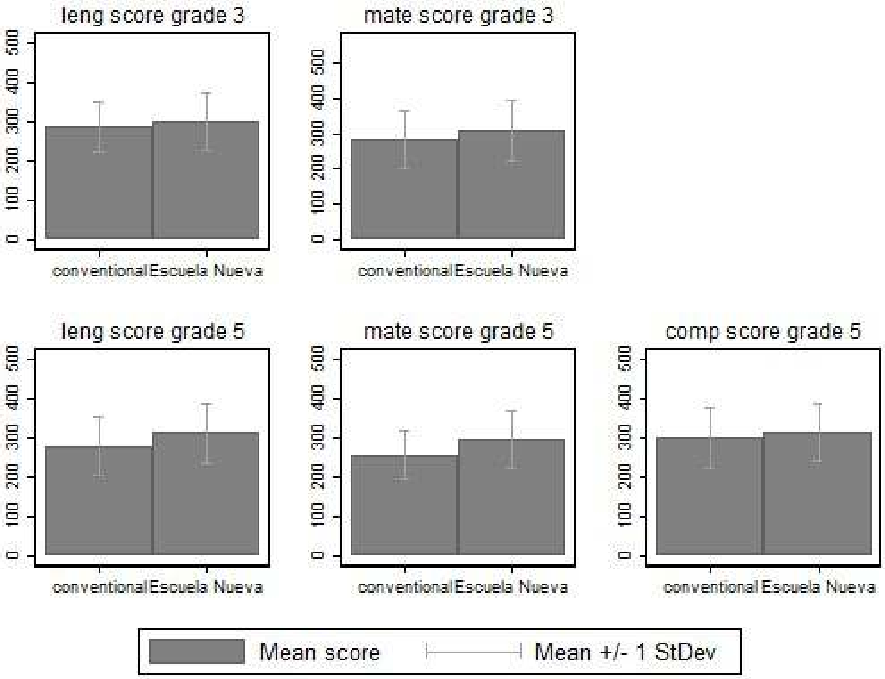


Figure 31 Exploratory Data Analysis: Mean test scores by school type (department-level study)

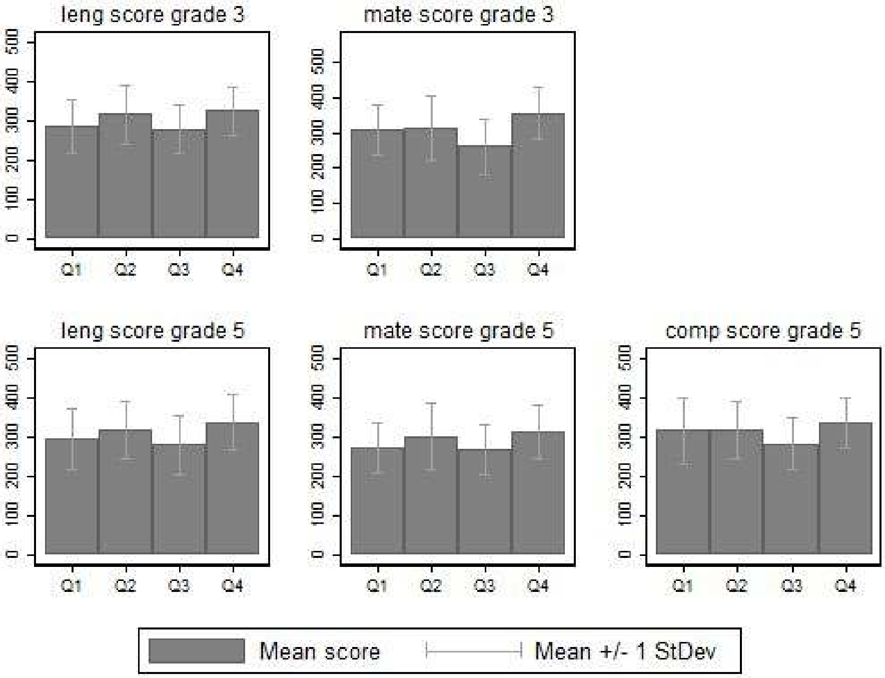


Figure 32 Exploratory Data Analysis: Mean test scores by implementation index quartile (department-level study)

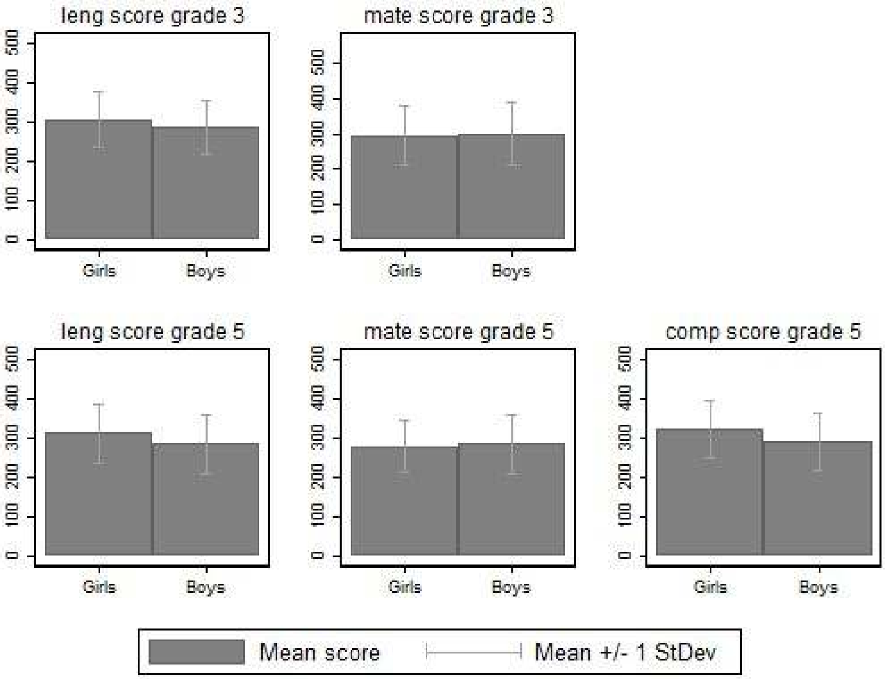


Figure 33 Exploratory Data Analysis: Mean test scores by gender (department-level study)

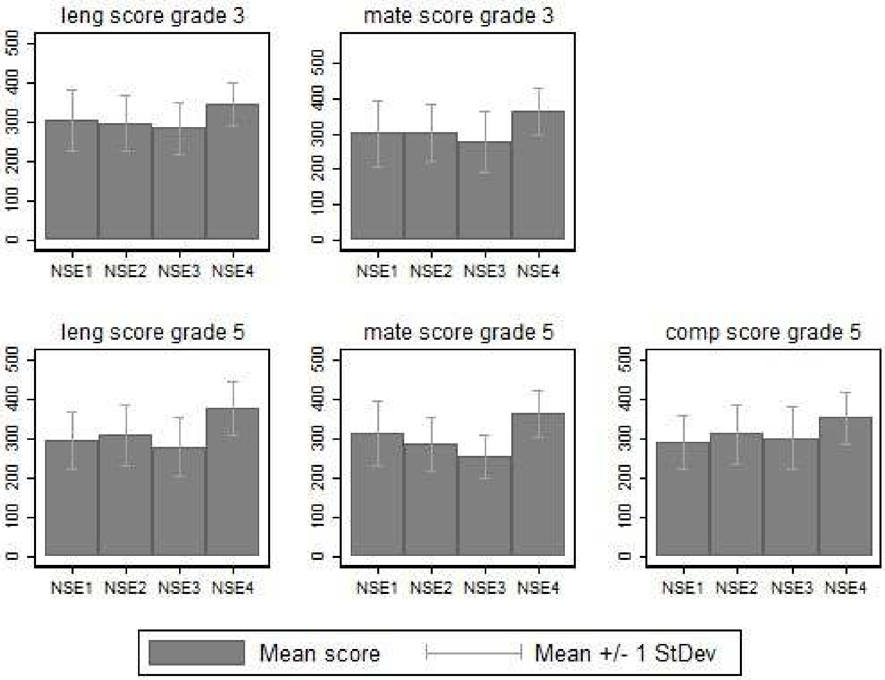


Figure 34 Exploratory Data Analysis: Mean test scores by socioeconomic level (department-level study)

6.1.2.1 School-level

In order to assess the need for a multilevel model, the variation of exam scores within and between schools needs to be analyzed. A multilevel model is likely to be called for if exam scores cluster around distinct school-level means. Figure 35 helps to assess this for the case of grade 3 language scores by plotting school-level mean scores and the associated spread in school-level scores for all schools. The schools are ranked by the school-level mean. Clearly, school mean scores differ widely between schools, ranging from under 200 to up to 500. Additionally, the graph suggests that the variation of scores within schools differs as well, with some schools producing much more homogenous outcomes than others. The difference in means indicates the need for school-specific intercepts, while the difference in the range of student-level results between schools indicates that student-level errors may be heteroskedastic. The picture looks similar for other grades and testing areas.

Differences in the mean of the outcome variable between schools indicates the need for school-level random intercepts, but differences in control variables between schools are equally important. If the correlation between student-level control variables and learning outcomes differs across schools, slope estimates might differ across schools, which would suggest the need for random coefficients. Figure 36 plots the school-level mean grade 3 reading scores by gender for all plausible values. Lines connect school means, so that each line represents the estimated difference between boys and girls for one school. The vertical spread of the lines again shows that mean reading scores differ between schools, as was noted above. Additionally, it seems from the graph that the effect of gender differs widely between schools – in some schools, girls outperform boys, while in others, boys outperform girls. Unless these differences across schools can be explained by the available control variables (including an interaction with the implementation index), it may thus be interesting to check for random school-level gender coefficients.

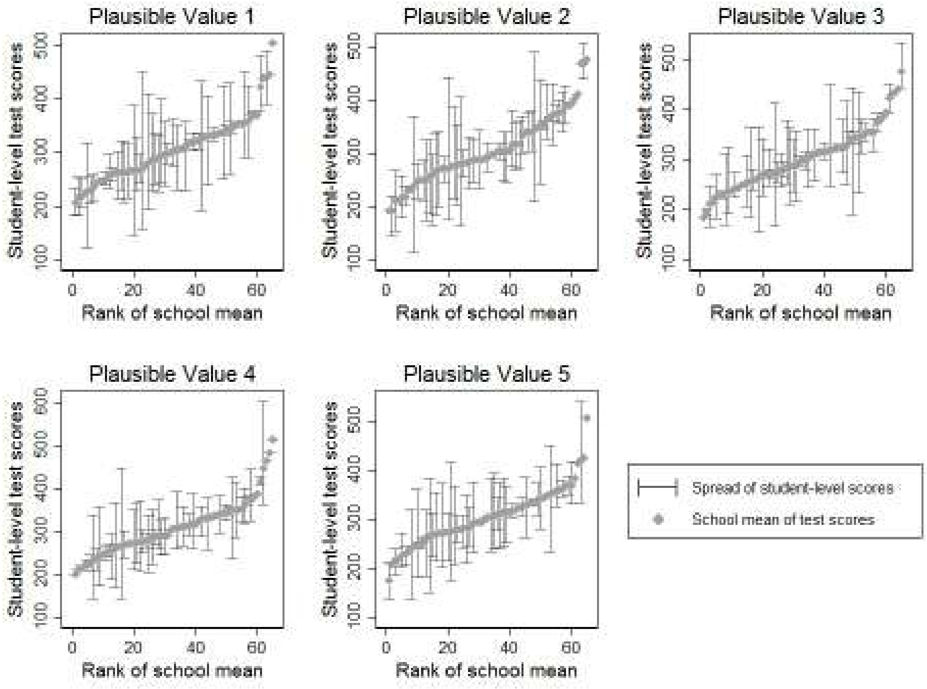


Figure 35 Variation in student-level reading scores, ranked by school-mean of reading scores, for all grade 3 plausible values (department-level study)

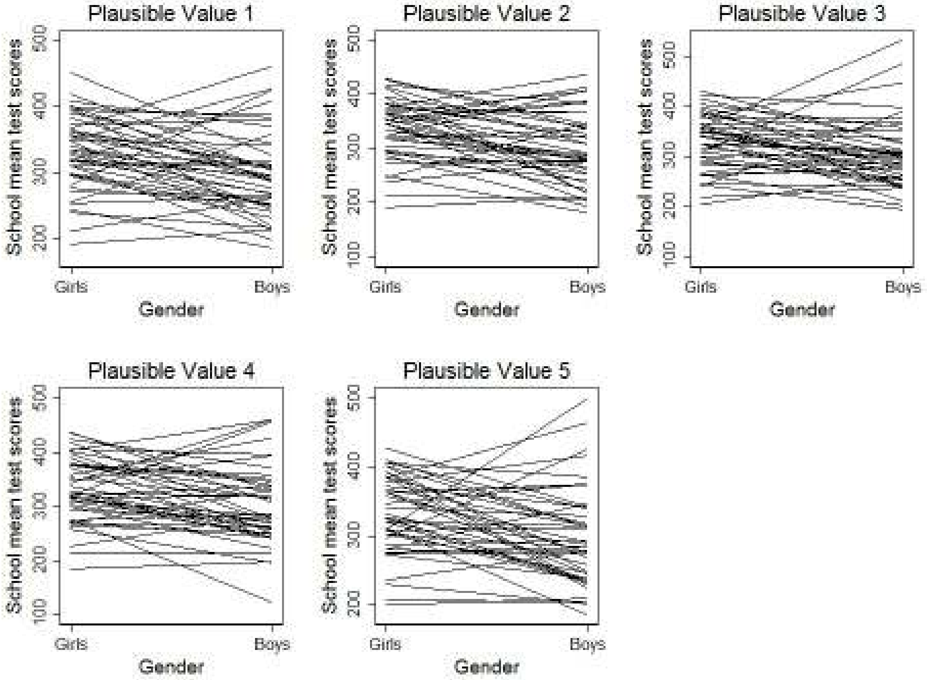


Figure 36 Mean grade 3 language test score by gender, for all plausible values. Lines are connecting school means

6.1.2.2 Municipality-level

The last exercise in this context is to assess differences across municipalities: Do students differ between municipalities? Do explanatory variables differ across municipalities, both with regard to their mean, and with regard to their correlation with the outcome variable?

Figure 37 plots the average school mean exam scores and school mean spreads for the ten municipalities in the study, ranked by the average school mean score, for all grades and areas. There clearly is variation between the municipalities, both with regards to the mean scores and with regard to the spread. Random municipality-level intercepts may thus be necessary. Given the small sample, however, it is not clear whether the differences between the municipalities are indeed statistically significant.

Figure 38 shows how lower-level predictors and municipality-level mean test scores are correlated, for the case of language grade 3 scores (results for other grades and areas are qualitatively similar). There are a few important insights to be gained from the plots. The first thing that strikes the eye is the small number of lines in some of the plots – in particular, no lines at all for session types, and only one line for the official EN classifier. The reason is that within most municipalities, there is no variation in these indicators – which is why the within-municipality effect of the given variable cannot be calculated. This is an indication for the limitations of the data with respect to the research question at hand.

Second, when there is within-municipality variation for most (or all 10) municipalities, the graphs indicate differences between the municipalities, both in the average exam scores, and in the slopes. For instance, in some municipalities, boys seem to be scoring higher on average, while in others, girls appear to score higher.

Third, the effect of the EN implementation index remains unclear from the plots; though, it seems to be the case that average exam scores are highest for children who are in schools in the highest quartile of the implementation index, which is what the research hypothesis predicts. However, the picture is far from clear, with some municipality-lines rising and some municipality-lines falling when moving from the lower to the highest implementation quartile. This may be an indication of different effects across municipalities – though, again, the sample may well be too small for statistically significant effects.

Overall, the plots suggest some differences between municipalities, but also some challenges in modeling these differences. Whether or not a municipality-level should be included thus has to be tested formally, which will be done in the next section.

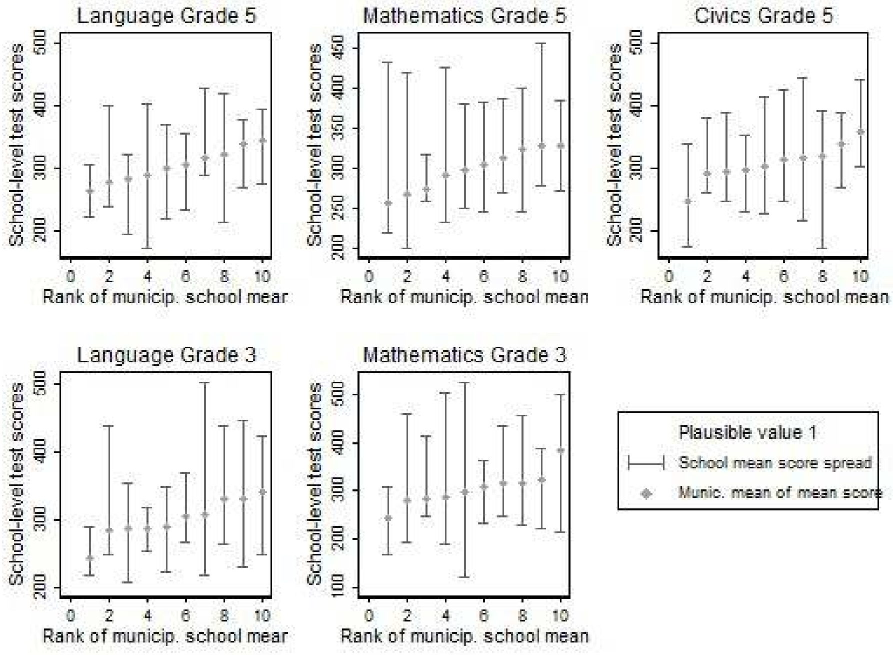


Figure 37 Variation in school-mean exam scores, ranked by municipality mean of school-mean scores, for plausible value 1 of all grades and areas (department-level study)

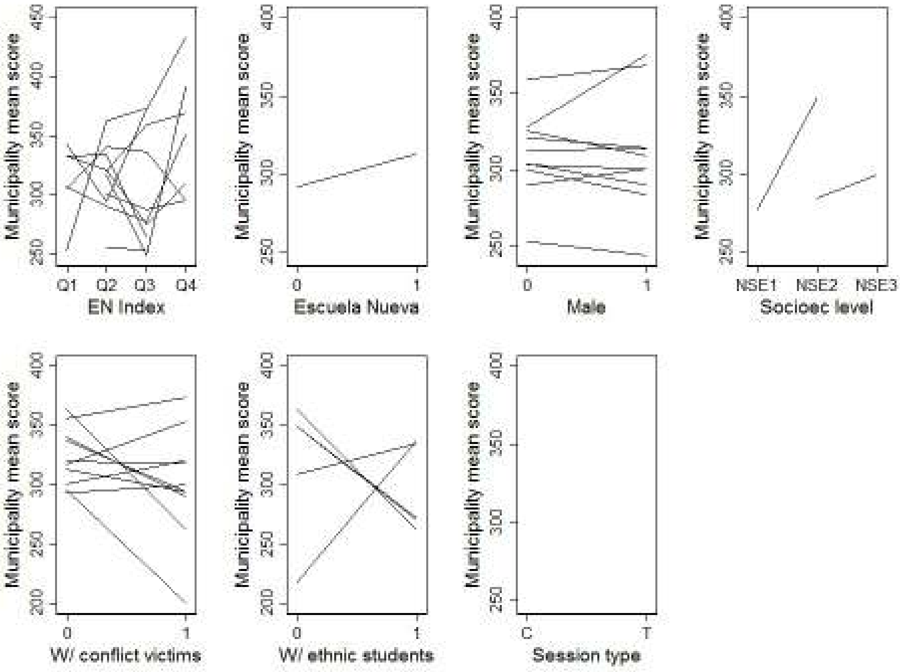


Figure 38 Municipality mean grade 3 language test scores by levels of lower-level predictor, for plausible value 1. Lines are connecting municipality means (department-level study)

6.2 The null model (ANOVA)

An analysis of variance (ANOVA, null-model, or variance component model) helps to understand the main drivers of differences in exam scores at the department-level: What share of the variance in test scores can be explained by differences between students, schools, and municipalities, respectively? The variance component model is calculated for all grades and test scores and helps to determine at which levels, if any, the variance calls for a random intercept.

6.2.1 Two-level analysis

The ANOVA model for the initial two-level case is defined as:

$$score_{ij} = \beta_j + \varepsilon_{ij}$$

Where: $$\beta_j = \beta + \zeta_j$$

So that $$\xi_{ij} = \zeta_j + \varepsilon_{ij}\ .$$

This model describes that the test score of student i in school j is composed of the school mean, β_j, and a random student-level error, ε_{ij}. The school-level mean, β_j, is itself composed of the grand mean, β, and a school-level random error term, ζ_j. (For better readability, the model does not include superscripts for grade or testing area). A different way to look at this model is to define the test score of student i in school j as the grand mean, β, and a composite error term, ξ_{ij}, which consists of the school-level error, ζ_j, and the student-level error, ε_{ij}.

Furthermore, an intraclass correlation-coefficient (ICC), ρ, is calculated to determine the percentage of the overall variance that is due to school-level effects. The ICC is defined as

$$\rho = \frac{\psi}{\psi + \theta}$$

where ψ is the variance of the school-level error term ζ_j, and θ is the variance of the student-level error term ε_{ij}.

The results of the ANOVA model are presented in Table 39. The table shows the number of observations for each model (between 252 and 378 students) and the number of clusters (between 60 and 75 schools). The fixed part of the model gives an estimate of the grand mean test score for each grade and area. These grand mean estimates are higher for each grade and testing area than the national estimates (see Table 11 on page 102), though the standard errors of the estimations are relatively high. The random part shows the estimated standard deviations at the school-level ($\sqrt{\psi}$) and at the student-level ($\sqrt{\theta}$). For each grade and testing area, except for 3rd grade mathematics, the student-level variance is larger than the school-level variance. This means that in all areas except that one, student-level characteristics are more important than school-level characteristics for the explanation of differences in exam scores. This is true particularly for the case of 5th grade language exam scores, where only 23% of overall variance is due to differences between schools. These results mirror the variance components results of the country-level study and again confirm results from previous studies (Casassus et al. 2000; Rangel and Lleras 2010; Baron 2012; Zambrano Jurado 2013).

The last row of the table presents the results of the likelihood-ratio tests which test the null hypothesis $H_0: \psi = 0$ against the hypothesis $H_a: \psi > 0$, i.e. which test whether the variance at the school-level is zero, and no multilevel-model is necessary. In all cases, the null hypothesis of zero variance can clearly be rejected. The conclusion thus is that a multilevel model is necessary.

Table 39 Two-level variance component models, department-level analysis

	Language, 3rd grade	Language, 5th grade	Math, 3rd grade	Math, 5th grade	Civics, 5th grade
n (students)	252	376	254	318	378
j (schools)	66	72	63	60	75
Fixed part:					
Grand mean	304.63 (7.54)	308.60 (6.78)	311.94 (10.09)	302.21 (7.85)	314.51 (6.44)
Random part (sd):					
School-level	42.72 (7.79)	36.88 (6.78)	66.73 (10.10)	47.23 (7.12)	39.11 (6.09)
Student-level	56.92 (3.39)	68.33 (2.87)	62.11 (3.84)	56.31 (3.33)	62.60 (2.74)
ICC (schools)	0.36	0.23	0.54	0.41	0.28
LR χ^2	36.93	34.48	93.18	72.53	69.47
p-value (LR χ^2)	0.00	0.00	0.00	0.00	0.00

Standard errors in parenthesis.

6.2.2 Three-level analysis

Schools are nested within municipalities, which in the country-level analysis turned out to be an important level of analysis. The question of interest is whether the same holds true for the department-level analysis. The ANOVA model is extended by a municipality-level to

$$score_{ijm} = \beta_{jm} + \varepsilon_{ijm}$$

Where: $\beta_{jm} = \beta_m + \zeta_{jm}$

and: $\beta_m = \beta + \zeta_m$

So that $\xi_{ijm} = \zeta_m + \zeta_{jm} + \varepsilon_{ijm}$.

In this model, β_{jm} is the school mean, which is composed of the municipality-level mean β_m and a school-level error term ζ_{jm}. The municipality-level mean is it composed of the grand mean β and a municipality-level error ζ_m. Together with the student-level error ε_{ijm}, the error terms add up to the total error ξ_{ijm}.

Table 40 shows the estimation results for this model. The observations for all grades and errors are now grouped into the 10 municipality-clusters. The decomposition of the error term shows that the standard deviation of the municipality-level error term, $\sqrt{\psi_{(3)}}$, is estimated as zero for three of the models, and as relatively small for the other two models (language grade 5 and mathematics grade 3). The overall share of variance explained by differences across schools and students, respectively, does thus not change much. Given these results, it is no surprise that the likelihood-ratio test of the null hypothesis that the municipality-level error variance is zero ($H_0\colon \psi_{(3)} = 0$) produces a small χ^2-statistic. The null hypothesis of no municipality-level random error term fails to be rejected, and the analysis will be continued based on a two-level framework.

Table 40 Three-level variance component models, department-level analysis

	Language, 3rd grade	**Language, 5th grade**	**Math, 3rd grade**	**Math, 5th grade**	**Civics, 5th grade**
n (students)	252	376	254	318	378
j (schools)	66	72	63	60	75
m (municipalities)	10	10	10	10	10
Fixed part:					
Grand mean	304.64 (7.61)	308.24 (8.42)	311.68 (10.60)	302.21 (7.85)	314.28 (7.65)
Random part (sd):					
Municipality-level	0.00 (0.03)	16.09 (11.00)	1.25 (10.41)	0.00 (.)	0.83 (.)
School-level	42.59 (8.04)	33.17 (7.66)	65.85 (10.73)	47.23 (8.59)	38.25 (6.39)
Student-level	56.92 (3.42)	68.34 (2.86)	62.11 (3.86)	56.31 (3.37)	62.52 (2.76)
ICC (schools)	0.36	0.18	0.53	0.41	0.27
ICC (municipalities)	0.00	0.04	0.00	0.00	0.00
ICC (j, m)	0.36	0.23	0.53	0.41	0.27
LR χ^2	0.00	0.54	0.03	0.00	0.03
p-value (LR χ^2)[20]	1.00	0.46	0.86	1.00	0.86

Standard errors in parenthesis.

[20] Halving the p-value is, as explained in the footnote 14 on page 117, does not change the conclusions.

6.3 The full multilevel model

As was the case for the country-level analysis, the full model is built using the "step-up" method of model construction (Ryoo 2011; cited in Kim, Anderson, and Keller 2013). Starting with a simple random-intercept model, fixed effects are successively added. Due to the lack of between-municipality variance, no random coefficients are modeled.

Two different sets of models are developed to analyze the effect of EN implementation in Quindío. The first set uses the total implementation index as the key explanatory variable to test the effect of overall model implementation. The second set replaces the overall implementation index with the five dimensions of the index (teacher training and support, classroom organization, school and community, learning guides, and roles of students) to assess which of these areas are more closely associated with improved learning outcomes. The development of both sets of models is described in detail in Annex B (sections 2 and 3). The following pages provide a summary.

6.3.1 Development of the implementation index model

The starting point is the two-level null model, given that between-municipality differences were found to be negligible in Quindío. Step by step, regressors are added, beginning with the EN implementation index and student-level variables, followed by core school-level variables necessary to test the research hypotheses, other school-level variables, and finally municipality-level variables. Only very few coefficients turn out to be individually significant, and the joint significance of regressors is different for the different testing areas. As shown extensively in Annex B, the best fit for grade 3 results is achieved with model QRI2, which is defined as:

Model QRI2: $$score_{ijm} = \beta_0 + \beta_1 ENI_{jm} + \beta_2 male_{ijm} + \beta_3 (male * ENI)_{ijm} + \beta_4 NSE_{jm} + \beta_5 (NSE * ENI)_{jm} + \xi_{ijm}$$

Model QRI3 provides the best fit for grade 5 results and is defined as:

Model QRI3: $$score_{ijm} = \beta_0 + \beta_1 ENI_{jm} + \beta_2 male_{ijm} + \beta_3 (male * ENI)_{ijm} + \beta_4 NSE_{jm} + \beta_5 (NSE * ENI)_{jm} + \beta_6 ethnic_{jm} + \beta_7 conflict_{jm} + \xi_{ijm}$$

In both cases, ξ_{ijm} is the composed error term consisting of the student-level error term ε_{ijm} and the school-level error term ζ_{jm}. Subscripts are added to all regressors to indicate the level on which the variables change (i for students, j for schools, and m for municipalities). Apart from the EN implementation index (ENI), the models include a predictor for gender ($male$) and a cross-level interaction term of the implementation index and gender ($male * ENI$). This interaction term is testing the hypothesis that the effect of the EN model differs by gender. As core school-level variables, both models include the average socioeconomic level of the school (NSE) and the interaction of socioeconomic level and the EN implementation index ($NSE * ENI$). Model QRI3 contains in addition some school-level control variables, namely the presence of students with ethnic background ($ethnic$) and who are victims of the conflict ($conflict$). The results for these models are summarized in Table 41 on page 197. The results are discussed in section 6.3.3, after the development of the random intercept model based on index areas in the next section.

6.3.2 Development of the implementation index-dimensions model

The EN model consists of a wide range of different elements. Despite Fundación Escuela Nueva's focus on the holistic nature of the model, it is conceivable that some aspects of the model are more strongly correlated with positive learning outcomes than others. Therefore, a second series of department-level hierarchical models are developed which use the five index dimensions instead of the overall index to identify EN implementation.

The starting point is again the two-level null model, and regressors are added step by step according to the level they belong to (student-level, core school-level, other school-level, and municipality-level). All models include the five implementation index dimensions ($D1$ to $D5$), each of them rescaled so that zero is the lowest value observed in the sample. The models and their respective fit is presented in detail in section 3 of Annex B.

As was the case for the models for the overall implementation index, the final model based on the best fit differs between grades and testing areas. For language as well as for grade 3 mathematics, model QRI6 provides the best fit and is defined as:

Model QRI6: $$score_{ijm} = \beta_0 + \beta_{1\_1} EN\ D1_{jm} + \beta_{1\_2} EN\ D2_{jm} + \beta_{1\_3} EN\ D3_{jm} + \beta_{1\_4} EN\ D4_{jm} + \beta_{1\_5} EN\ D5_{jm} + \beta_2 male_{ijm} + \beta_{3\_1}(male * D1)_{ijm} + \beta_{3\_2}(male * D2)_{ijm} + \beta_{3\_3}(male * D3)_{ijm} + \beta_{3\_4}(male * D4)_{ijm} + \beta_{3\_5}(male * D5)_{ijm} + \beta_4 NSE_{jm} + \beta_{5\_1}(NSE * D1)_{jm} + \beta_{5\_2}(NSE * D2)_{jm} + \beta_{5\_3}(NSE * D3)_{jm} + \beta_{5\_4}(NSE * D4)_{jm} + \beta_{5\_5}(NSE * D5)_{jm} + \xi_{ijm}$$

For grade 5 mathematics as well as for civic competencies, the best-fitting model is model QRI7, which is defined as:

Model QRI7: $$score_{ijm} = \beta_0 + \beta_{1\_1} EN\ D1_{jm} + \beta_{1\_2} EN\ D2_{jm} + \beta_{1\_3} EN\ D3_{jm} + \beta_{1\_4} EN\ D4_{jm} + \beta_{1\_5} EN\ D5_{jm} + \beta_2 male_{ijm} + \beta_{3\_1}(male * D1)_{ijm} + \beta_{3\_2}(male * D2)_{ijm} + \beta_{3\_3}(male * D3)_{ijm} + \beta_{3\_4}(male * D4)_{ijm} + \beta_{3\_5}(male * D5)_{ijm} + \beta_4 NSE_{jm} + \beta_6 ethnic_{jm} + \beta_7 conflict_{jm} + \xi_{ijm}$$

All fixed and random terms have the same interpretation as in the previous section, with the important difference that the overall index ENI was split up into its five components $D1$ to $D5$, which affects the main effects of the EN dimensions as well as the interaction terms with gender and socioeconomic status. The difference between models QRI6 and QRI7 is the inclusion of the regressors $ethnic$ and $conflict$ in the latter. The final models and results are summarized in Table 42.

6.3.3 Results

The estimation results for the final models are presented in Table 41 for estimations based on the overall index, and in Table 42 for estimations based on the index dimensions. As the results for both approaches largely mirror each other, they are discussed jointly.

The main question of interest concerns the effect of EN program implementation on learning outcomes. As was made clear in the previous two sections, the models provide no evidence for such an effect, neither based on the overall index, nor based on the implementation dimensions. For the former, the point estimates are positive for all cases expect for grade 5 language exams, yet the t-statistics of the coefficients are well below conventional significance thresholds. For the individual dimensions, the sign of the effect differs between dimensions, grades, and areas, and the coefficients also lack statistical significance. Thus, the null hypothesis of no effect of program implementation fails to be rejected.

Table 41 Overview: Results of the final random intercept models for the overall index, all grades and areas (department-level study)

	Language Grade 3		Language Grade 5		Mathematics Grade 3		Mathematics Grade 5		Civics Grade 5	
n (students)	252		369		254		312		371	
j (schools)	66		71		63		59		74	
Fixed part:										
EN Index	0.92	(1.55)	-1.60	(1.47)	0.79	(2.23)	0.41	(1.68)	1.23	(1.30)
Male	-20.28	(19.64)	-20.49	(20.57)	31.03	(22.32)	16.70	(18.56)	-19.81	(16.27)
ENI*Male	0.08	(0.50)	-0.28	(0.49)	-0.80	(0.58)	-0.34	(0.47)	-0.43	(0.41)
Socioeconomic level	14.20	(59.13)	-53.19	(52.18)	19.17	(83.69)	-24.29	(68.08)	67.52	(48.13)
ENI*Socioec. level	-0.16	(1.45)	1.70	(1.34)	0.05	(2.08)	0.23	(1.64)	-0.85	(1.21)
w/ ethnic students			-50.38*	(20.76)			-25.21	(21.79)	-38.03*	(18.79)
w/ conflict victims			-15.73	(13.09)			-38.20**	(14.51)	-13.21	(12.40)
Grand mean	271.11***	(63.35)	388.16***	(59.76)	262.58**	(89.68)	321.30***	(68.94)	265.37***	(53.06)
Random part (sd):										
School-level	41.45	(8.06)	32.25	(7.47)	65.49	(10.06)	37.44	(7.37)	31.93	(6.79)
Student-level	56.22	(3.33)	66.26	(2.79)	61.66	(3.80)	56.28	(3.60)	60.03	(2.58)
ICC (schools)	0.35		0.19		0.53		0.31		0.22	
Variance explained (c)	3.7%		9.9%		2.6%		15.4%		15.2%	

Standard errors in parenthesis. ***p≤0.001; ** p<0.01; * p<0.05

Table 42 Overview: Results of the final random intercept models for the index dimensions, all grades and areas (department-level study)

	Language Grade 3		Language Grade 5		Mathematics Grade 3		Mathematics Grade 5		Civics Grade 5	
n (students)	252		376		254		318		378	
j (schools)	66		72		63		60		75	
Fixed part:										
Dimension 1 (Training)	-1.05	(0.77)	-0.50	(0.63)	-0.31	(1.18)	0.05	(0.62)	-0.54	(0.57)
Dimension 2 (Classroom)	1.06	(1.32)	-0.56	(1.15)	0.46	(1.88)	1.31	(1.56)	0.24	(0.99)
Dimension 3 (Community)	1.26	(0.89)	0.23	(0.75)	0.12	(1.32)	1.26	(0.76)	0.62	(0.71)
Dimension 4 (Guides)	-0.02	(1.67)	-0.13	(1.72)	2.27	(2.31)	0.26	(1.78)	1.05	(1.45)
Dimension 5 (Roles)	-1.59	(2.34)	0.00	(2.05)	-2.74	(3.53)	-3.78	(2.45)	-1.28	(1.87)
Male	-2.00	(31.21)	5.59	(28.20)	49.05	(35.88)	21.62	(28.62)	-10.02	(25.87)
EN D1*Male	0.41	(0.47)	-0.75	(0.45)	-0.32	(0.58)	-0.52	(0.36)	0.14	(0.39)
EN D2*Male	0.72	(0.72)	1.01	(0.67)	0.37	(0.79)	0.16	(0.82)	-0.09	(0.74)
EN D3*Male	0.06	(0.63)	-0.56	(0.52)	-0.29	(0.76)	-0.64	(0.52)	0.04	(0.53)
EN D4*Male	-1.48	(0.84)	-0.38	(0.75)	-0.33	(0.82)	-0.03	(0.80)	-0.60	(0.66)
EN D5*Male	0.49	(0.98)	-0.28	(1.03)	-0.66	(1.06)	0.72	(0.96)	0.26	(0.95)
Socioeconomic level	-48.68	(100.13)	-110.05	(87.56)	60.38	(131.94)	8.96	(107.34)	63.41	(74.39)
EN D1*Socioec. level	1.23	(0.76)	0.40	(0.63)	1.10	(1.11)	0.73	(0.65)	0.45	(0.59)
EN D2* Socioec. level	-1.63	(1.14)	-0.86	(0.89)	-0.40	(1.49)	-1.01	(1.56)	-0.82	(0.75)
EN D3* Socioec. level	-0.87	(0.86)	0.61	(0.73)	-0.64	(1.23)	-0.55	(0.80)	0.11	(0.63)
EN D4* Socioec. level	0.90	(1.79)	1.43	(1.73)	-2.70	(2.39)	-0.75	(1.78)	-0.92	(1.49)
EN D5*v. level	1.78	(2.28)	0.85	(1.96)	4.04	(3.31)	2.69	(2.38)	1.64	(1.79)
w/ ethnic students							-45.74	(25.05)	-47.73*	(19.40)
w/ conflict victims							-39.77**	(14.63)	-16.66	(12.23)
Grand mean	307.42**	(102.15)	371.67***	(88.92)	207.34	(136.79)	309.82**	(111.24)	277.24***	(79.39)
Random part (sd):										
School-level	38.70	(8.59)	30.01	(7.25)	62.29	(9.80)	33.52	(6.78)	21.64	(13.22)
Student-level	54.88	(3.20)	66.13	(2.77)	61.54	(3.81)	55.75	(3.28)	60.84	(2.80)
ICC (schools)	0.33		0.17		0.51		0.27		0.11	
Variance explained	10.9%		12.5%		7.7%		21.7%		23.5%	

Standard errors in parenthesis. ***p≤0.001; ** p<0.01; * p<0.05

The conclusion is similar for the hypothesis that the model is particularly beneficial for students from disadvantaged backgrounds and helps to close gender gaps. None of the interaction terms is statistically significant in any of the model specifications.

Of the remaining coefficients in the model, only very few are statistically significant. For grade 5 language and civics, students in schools with students of ethnic background tend to score worse; in grade 5 mathematics, students in schools with victims of the conflict tend to score worse. The grand mean estimate is significant in most, though not all, models. Its interpretation differs depending on the variables included in the model. For grade 3 models (and grade 5 language in the case of the estimates based on indicator dimensions), it is the estimated exam score for girls in schools with the lowest program implementation level and the lowest average socioeconomic level. For the other models, it is the estimated exam score for girls in schools with the lowest program implementation level and the lowest average socioeconomic level, and with no students who are conflict victims or of ethnic background.

The estimates for the random part of the model are also congruent between the two modelling approaches. In the case of grade 3 mathematics, unexplained variance is divided more or less equally between the student- and school-levels, meaning that exam scores vary just as much between schools as within schools. In all other cases, the largest share of unexplained variance is at the student-level, with intraclass correlation coefficients between 0.11 (for civics in the model based on indicator dimensions) and 0.35 (for language grade 3 in the model based on the overall index).

Compared to the total variance found in the null model, unexplained variance is reduced in the final random intercept models, as indicated by the R^2 reported in the last row of each table (see footnote 15 on page 115 for an explanation of R^2). This is particularly the case for grade 5 mathematics and civic competencies. However, the R^2 statistic reported in the table does not take into account the number of variables used in the model, and thus tends to exaggerate the model fit in models with many regressors. This is well demonstrated by comparing the respective statics of Table 41 and Table 42 – even though the five index dimensions add up to the overall index, the share of explained variance is considerably higher in Table 42. It seems unreasonable that the index dimension-model explains over 20% of the model variance (in grade 5 mathematics and civics), when only one of the regressors is statistically significant.

Overall, the results are thus not encouraging. However, a lack of evidence for an effect of EN implementation is not the same as evidence for the lack of an effect of EN implementation. The sample with which calculations are performed is relatively small for a multilevel analysis, which results in a lot of uncertainty and in large standard errors. It is telling that variables that were clearly significant in the country-level analysis lack significance in the Quindío sample. The reason may be an actual lack of correlation in Quindío, but the more likely interpretation is a lack of statistical power. In this sense, the (scarce) significant effects found in the grade 5 models are likely not due to the fact that the respective variables only have a significant effect in these grades, but due to the fact that the grade 5 sample is considerably larger.[21]

Unfortunately, Stata13's multilevel commands do not allow for an integration with survey estimation techniques. This means that it is not possible to correct for the finite population of the analysis: While implementation data is available for around 50% of Quindío's rural primary schools, Stata treats the sample of schools as if they represented an infinite (or large) population

[21] This is because of the way in which the Pruebas SABER are performed: Each grade 5 student takes two out of three area exams, while each grade 3 student takes one out of two (see section 1)

of schools. Therefore, standard errors are overestimated, and the coefficients are less likely to be significant. This is addressed in the next section (section 6.4).

6.3.4 Random coefficients

In the country-level study, there was evidence for department- and municipality-level random coefficients of EN. The former is clearly out of question in the department-level analysis. The latter would be possible; yet, given that there is no evidence for municipality-level random intercepts, it would make little sense to include municipality-level random slopes (Rabe-Hesketh and Skrondal 2012, 213). Doing so would imply variance in test results between municipalities depending on program implementation, but without variance in the mean score between municipalities—which is quite unlikely. Hence, the only possible random slope given the available data is at the school-level. The exploratory data analysis did indeed suggest that the effect of gender may vary between schools; yet there is no theoretic reason to believe that this may be the case in this model, other than via an interaction with the EN school model, an effect that lacked statistical significance in all specifications presented. Hence, given the strong limitations of the available database, no random coefficients are added to the model.

6.4 Survey analysis model

Multilevel modeling is not the only way of dealing with hierarchical data. Another possibility is survey data analysis, that is, OLS estimation with clustered standard errors. Compared to hierarchical models, the main disadvantage of survey data analysis is that it does not give any insights in the variance between the clusters (and its determinants). However, an important practical advantage is that survey analysis allows for the use of a finite population correction factor, which decreases the standard errors when the sample is large in relation to the total population, as in the case of the study sample. Multilevel survey estimation is a relatively new field, and Stata13 does not provide that option.[22] Survey estimation is used as a second strategy to analyze the effect of program implementation in Quindío. The cluster variables (or primary sampling units) are the schools; the finite population correction factor is calculated as $\sqrt{\frac{N-n}{N-1}}$, where N is the population size (149) and n is the sample size (76).

6.4.1 Model specification

Two sets of models are estimated, one based on the overall implementation index, and the other one based on the index dimensions. In each case, there are two model specifications. The first one is a "basic" specification that includes only the variables necessary for testing the research hypotheses. The second one is the "full" specification that includes the available control variables.

The basic model for the full implementation index is defined as follows:

Model QSA1: $score_i = \beta_0 + \beta_1 ENI_i + \beta_2 male_i + \beta_3 (male * ENI)_i + \beta_4 NSE_i + \beta_5 (NSE * ENI)_i + \epsilon_i^{(j)}$

The corresponding model for the index dimensions is:

Model QSA2: $score_i = \beta_0 + \beta_{1\_1} EN\ D1_i + \beta_{1\_2} EN\ D2_i + \beta_{1\_3} EN\ D3_i + \beta_{1\_4} EN\ D4_i + \beta_{1\_5} EN\ D5_i + \beta_2 male_i + \beta_{3\_1} (male * D1)_i + \beta_{3\_2} (male * D2)_i + \beta_{3\_3} (male * D3)_i + \beta_{3\_4} (male * D4)_i + \beta_{3\_5} (male * D5)_i + \beta_4 NSE_i + \beta_{5\_1} (NSE * D1)_i + \beta_{5\_2} (NSE * D2)_i + \beta_{5\_3} (NSE * D3)_i + \beta_{5\_4} (NSE * D4)_i + \beta_{5\_5} (NSE * D5)_i + \epsilon_i^{(j)}$

[22] Survey commands for multilevel models were introduced in Stata14.

In order to mark the difference to the multilevel model, the subscript i now indicates that all variables are defined at the level of the student. Furthermore, the error term is now defined at the student-level as $\epsilon_i^{(j)}$; the superscript (j) is added to indicate school-level clustering. The variables are defined in the same way as before: ENI is the EN implementation index, rescaled so that zero denotes the smallest index value observed; $D1$ to $D5$ are the five index dimensions, rescaled in the same way; $male$ is a dummy for boys; and NSE is the socioeconomic level, 0 denoting NSE1. The models contain the same interactions between gender and EN implementation index (dimensions) and socioeconomic level and EN implementation index (dimensions) as the models estimated in the previous section.

The full specification of the models includes the dummy variables $ethnic$ (schools with students of ethnic background), $conflict$ (schools with students who are conflict victims), and $governance$ (the municipal governance index, centered at the mean of the municipalities):

Model QSA3: $score_i = \beta_0 + \beta_1 ENI_i + \beta_2 male_i + \beta_3 (male * ENI)_i + \beta_4 NSE_i + \beta_5 (NSE * ENI)_i + \beta_6 ethnic_i + \beta_7 conflict_i + \beta_8 governance_i + \epsilon_i^{(j)}$

Model QSA4: $score_i = \beta_0 + \beta_{1\_1} EN\ D1_i + \beta_{1\_2} EN\ D2_i + \beta_{1\_3} EN\ D3_i + \beta_{1\_4} EN\ D4_i + \beta_{1\_5} EN\ D5_i + \beta_2 male_i + \beta_{3\_1} (male * D1)_i + \beta_{3\_2} (male * D2)_i + \beta_{3\_3} (male * D3)_i + \beta_{3\_4} (male * D4)_i + \beta_{3\_5} (male * D5)_i + \beta_4 NSE_i + \beta_{5\_1} (NSE * D1)_i + \beta_{5\_2} (NSE * D2)_i + \beta_{5\_3} (NSE * D3)_i + \beta_{5\_4} (NSE * D4)_i + \beta_{5\_5} (NSE * D5)_i + \beta_6 ethnic_i + \beta_7 conflict_i + \beta_8 governance_i + \epsilon_i^{(j)}$

6.4.2 Results

The estimation results are presented in Table 43 through Table 46. The first table contains the results for model QSA1, the "basic" specification with the full index. The effect of EN implementation is positive and highly to moderately significant in all grades and areas. For girls of the lowest socioeconomic level, the expected improvement in the exam scores is between 2.0 and 5.9 points per additional percentage point of model implementation, depending on grade and area. Given that the observed index values have a range of about 60 percentage points and a standard deviation of 11.9, this is a large effect. The EN effect is larger for the grade 3 than for the grade 5 exams. Boys do better than girls in grade 3 math and worse in civic competencies. In grade 3 math, the EN model helps to diminish these gender differences. Finally, the socioeconomic level has a large positive and significant effect, and in all grades and areas, EN helps to decrease the discrepancy between the socioeconomic levels. However, the coefficient on the interaction term is small compared to the main effect of socioeconomic level. This indicates that a fairly high implementation level is necessary in order to make up for the differences. The last part of the table contains the number of observations, and R^2 as a measure for model fit.[23] The basic model explains only between 5.1% and 18.3% of the score variance, which is not much.

The results of adding control variables ($ethnic$, $conflict$, and $governance$) to create model QSA3 are presented in Table 44. EN implementation is significant only in two of the five cases: In grade 3 mathematics and in civic competencies, where students can expect to score 3.7 and 2.3 points, respectively, higher for each percentage point increase in the index. Gender, socioeconomic level, and the interaction between EN implementation and socioeconomic level are also only significant in these two models, with the expected signs (socioeconomic level alone is also significant for

[23] Because the estimations are performed through multiple imputation, the R^2 statistic is not provided as part of the standard output. The statistic is computed by applying Rubin's combination rules. Harel (2009) suggests to transform the R^2 of each imputed dataset through Fisher's z (or inverse hyperbolic tangent) transformation to improve its asymptotic normality. The thus transformed R^2 statistics are combined, and then transformed back into the original metric (following Cañette and Marchenko 2017). The R^2 estimate obtained may be biased upwards and should ideally be accompanied by adjusted R^2 estimates, which in this procedure tend to be biased downwards (UCLA Statistical Consulting Group 2017). However, this is not done here as reporting an adjusted R^2 for survey data is not straight forward.

grade 3 language). Students in schools with ethnic populations score lower in all cases; students in schools with conflict victims score lower in two cases. Governance is significant only for grade 5 mathematics, where the effect is unexpectedly negative (worse governance is associated with higher learning outcomes). The model fit improves compared to model QSA1; between 13% and 23% of the variance is explained.

Table 43 Results of the survey estimation models QSA1 (department-level study)

	Language Grade 3			Language Grade 5			Mathematics Grade 3			Mathematics Grade 5			Civic Competencies Grade 5		
	β		se	β		se	β		se	β		se	β		se
EN Index	3.40	***	0.47	1.96	**	0.67	5.93	***	0.94	3.22	**	1.03	4.7	***	0.60
male	-16.84		17.34	-21.85		13.48	38.71	*	12.44	10.66		12.02	-22.93	*	9.55
ENI*male	-0.11		0.43	-0.11		0.30	-1.00	*	0.38	-0.1		0.36	-0.35		0.24
Socioeconomic Level	115.22	***	18.38	53.79	*	20.95	189.59	***	28.42	86.28		48.87	176.72	***	21.84
ENI*Socioec. level	-2.73	***	0.43	-1.46	**	0.48	-4.68	***	0.68	-2.42	*	0.99	-4.05	***	0.47
Intercept	169.00	***	19.36	242.47	***	28.61	70.43	*	34.27	170.63	***	46.96	126.14	***	26.67
R^2	0.092			0.051			0.130			0.076			0.183		
n	252			376			254			318			378		

***p≤0.001; ** p<0.01; * p<0.05

Table 44 Results of the survey estimation models QSA3 (department-level study)

	Language Grade 3			Language Grade 5			Mathematics Grade 3			Mathematics Grade 5			Civic Competencies Grade 5		
	β		se	β		se	β		se	β		se	β		se
EN Index	1.34		0.75	-1.36		0.68	3.70	*	1.46	0.74		0.66	2.26	**	0.68
male	-12.34		16.11	-14.42		13.99	35.20	*	12.43	15.41		12.71	-17.60	*	8.37
ENI*male	-0.11		0.39	-0.31		0.31	-0.94		0.38	-0.28		0.35	-0.46	*	0.21
Socioeconomic Level	56.13	*	25.51	-32.40		22.20	122.07	**	38.99	33.30		27.89	113.43	***	24.72
ENI*Socioec. level	-1.11		0.61	1.22		0.58	-2.88	**	1.03	-0.59		0.59	-2.09	***	0.57
w/ ethnic students	-32.35	**	11.09	-58.97	***	10.24	-40.64	*	16.01	-19.15	*	8.43	-44.12	***	9.21
w/ conflict victims	-16.11		9.48	-17.15	*	7.80	-1.95		12.69	-52.6	***	7.96	-9.73		6.73
Governance	0.63		0.95	-0.67		0.95	1.62		1.32	-3.18	***	0.81	-0.55		0.87
Intercept	262.12	***	30.84	378.24	***	27.91	166.59	**	55.10	288.38	***	30.98	223.24	***	29.14
R^2	0.133			0.131			0.154			0.202			0.228		
n	252			376			254			318			378		

***p≤0.001; ** p<0.01; * p<0.05

Table 45 Results of the survey estimation models QSA2 (department-level study)

	Language Grade 3			Language Grade 5			Mathematics Grade 3			Mathematics Grade 5			Civic Competencies Grade 5		
	β		**se**	**β**		**se**	**β**		**se**	**β**		**se**	**β**		**se**
Dimension 1 (Training)	-0.82		0.50	-0.08		0.34	-0.07		0.87	-0.02		0.46	-0.43		0.35
Dimension 2 (Classroom)	1.34	*	0.63	-1.03		0.63	1.11		1.06	2.05		1.06	1.51	*	0.64
Dimension 3 (Community)	1.75	**	0.48	0.27		0.37	1.63	*	0.80	1.05	*	0.50	0.83		0.48
Dimension 4 (Guides)	0.10		0.82	1.16		1.07	3.03	*	1.33	0.98		1.06	2.16	*	0.82
Dimension 5 (Roles)	-1.22		1.43	-0.87		1.15	-3.70		2.41	-4.57	*	1.79	-2.80	**	1.02
male	-0.28		16.25	18.47		15.22	53.61	*	21.53	21.36		17.90	-11.18		13.04
D1*male	0.42		0.27	-1.03	**	0.31	-0.23		0.36	-0.44		0.23	0.16		0.28
D2*male	0.77		0.39	1.14	*	0.43	0.93	*	0.45	0.48		0.64	0.09		0.50
D3*male	-0.31		0.38	-0.47		0.31	-1.15	*	0.44	-0.79		0.38	-0.11		0.33
D4*male	-1.29	*	0.47	-0.79		0.40	-0.14		0.51	-0.43		0.48	-0.67	*	0.32
D5*male	0.30		0.57	0.32		0.63	-1.28	*	0.60	1.03		0.75	0.32		0.53
Socioeconomic level	16.43		45.98	-69.82		41.67	152.86	*	69.75	8.26		66.82	125.91	*	55.22
D1*Socioeconomic level	0.93	*	0.42	-0.06		0.35	0.47		0.66	0.11		0.46	0.04		0.37
D2*Socioeconomic level	-1.78	**	0.49	-0.56		0.43	-1.54	*	0.58	-1.97	*	0.97	-1.82	***	0.46
D3*Socioeconomic level	-1.43	**	0.49	0.55		0.34	-0.91		0.64	-0.14		0.46	-0.12		0.42
D4*Socioeconomic level	0.49		0.88	0.19		1.08	-3.75	**	1.34	-0.63		1.09	-1.67		0.90
D5*Socioeconomic level	1.69		1.38	1.85		1.10	5.98	**	2.05	3.98	*	1.61	3.24	**	1.00
Intercept	243.43	***	47.75	332.52	***	40.64	93.17		72.18	234.77	**	66.16	163.54	**	52.76
R^2	0.106			0.122			0.152			0.109			0.209		
n	252			376			254			318			378		

***p≤0.001; ** p<0.01; * p<0.05

Table 46 Results of the survey estimation models QSA4 (department-level study)

	Language Grade 3			Language Grade 5			Mathematics Grade 3			Mathematics Grade 5			Civic Competencies Grade 5		
	β		se	β		se	β		se	β		se	β		se
Dimension 1 (Training)	-0.77		0.54	-0.31		0.34	-0.26		0.88	0.14		0.43	-0.65		0.34
Dimension 2 (Classroom)	0.53		0.59	-2.07	**	0.68	1.04		1.08	0.24		0.93	0.20		0.64
Dimension 3 (Community)	1.48	**	0.45	0.27		0.34	1.45		0.83	0.89	*	0.39	0.72		0.38
Dimension 4 (Guides)	-0.04		0.75	0.16		1.04	2.70	*	1.29	0.75		1.01	1.09		0.79
Dimension 5 (Roles)	-1.91		1.13	0.25		1.20	-4.46		2.30	-2.15		1.66	-1.52		0.95
male	9.00		17.42	15.82		15.38	52.08	*	20.50	28.41		18.21	-5.80		13.10
D1*male	0.24		0.28	-1.00	**	0.31	-0.24		0.35	-0.67	**	0.23	0.07		0.28
D2*male	0.87	*	0.39	1.08	*	0.44	0.78		0.41	0.46		0.64	0.02		0.50
D3*male	-0.33		0.37	-0.55		0.31	-1.06	*	0.45	-0.51		0.38	-0.02		0.34
D4*male	-1.18	*	0.48	-0.49		0.43	-0.08		0.51	-0.33		0.48	-0.52		0.32
D5*male	0.06		0.59	-0.07		0.67	-1.20		0.61	0.66		0.76	0.01		0.54
Socioeconomic level	-36.32		44.57	-116.10	*	43.48	102.42		76.10	27.09		61.13	73.04		50.92
D1*Socioeconomic level	1.28	**	0.41	0.43		0.36	0.80		0.71	0.27		0.44	0.62		0.35
D2*Socioeconomic level	-1.15	*	0.43	0.21		0.48	-1.29		0.64	-0.84		0.84	-0.79		0.50
D3*Socioeconomic level	-1.02	*	0.42	0.64		0.33	-0.79		0.65	-0.06		0.34	0.03		0.30
D4*Socioeconomic level	0.21		0.77	0.59		1.07	-3.67	**	1.24	-0.63		1.05	-1.31		0.82
D5*Socioeconomic level	2.09		1.13	0.84		1.12	6.35	**	1.96	1.94		1.50	2.06	*	0.90
w/ ethnic students	-47.55	***	11.90	-46.16	***	10.63	-35.26		19.61	-23.77		12.45	-58.82	***	10.10
w/ conflict victims	-20.56		11.63	-20.46	*	8.36	4.98		13.42	-49.60	***	7.07	-21.72	*	8.10
Governance	1.30		1.06	-0.69		0.96	1.79		1.29	-2.77	*	1.03	-1.06		0.84
Intercept	344.16	***	48.83	436.83	***	46.00	163.45		85.30	284.49	***	60.03	283.76	***	50.07
R^2	0.202			0.187			0.210			0.236			0.277		
n	252			376			254			318			378		

***p≤0.001; ** p<0.01; * p<0.05

Table 45 and Table 46 contain the results for the models based on the index dimensions. In the basic specification, some index dimensions are statistically significant in some of the cases. Specifically, there are significant, positive main effects of the dimension classroom organization in grade 3 language and civics; of dimension three (community relations) in mathematics and in grade 3 language; and of dimension four (learning guides) in grade 3 mathematics and in civics. Additionally, the main effect of dimension five (student roles) is *negative* for grade 5 mathematics and for civics. The effect of EN implementation seems to vary by gender for dimension one in grade 5 language; for dimension two in grade 5 language and grade 3 mathematics; for dimensions 3 and 5 in grade 3 mathematics; and for dimension 4 in grade 3 language and civics. It also seems to vary by socioeconomic level for dimensions one, two, and three in grade 3 language; for dimensions two, four, and five in grade 3 mathematics; and for dimensions two and five in grade 5 mathematics and civics. That being said, the direction of the effect is not always clear. For instance, in the case of socioeconomic level, a higher implementation percentage in dimension five (roles of students) tends to help children in schools with a high average socioeconomic level more than other children, while a higher implementation percentage in dimension two (classroom organization) tends to benefit children in schools with lower average socioeconomic level more. The R^2 statistics show that the basic model can explain between 10% and 21% of total variance in exam scores.

Finally, the results for model QSA4 (index dimensions and control variables), are shown in Table 46. There are only a few instances of a significant effect of EN implementation. For dimension one (teacher training), the base effect is not significant, but the interaction with gender is in two cases (with a negative sign), and the interaction with socioeconomic level is in one case (with a positive sign). Dimension two (classroom organization) has a significant negative effect in grade 5 language, and that effect is significantly smaller for boys; classroom organization is also significant in its interaction with socioeconomic level for grade 3 language, where the model helps to decrease the gap between the socioeconomic levels. Dimension three (community relations) is positively associated with scores in grade 3 language and grade 5 math; it decreases gender gaps in grade 3 math, and differences between socioeconomic levels in grade 3 language. Dimension four (learning guides) has a positive main effect in grade 3 mathematics, decreases gender gaps in grade 3 language, and differences between socioeconomic levels in grade 3 mathematics. Finally, dimension five (student roles) is only significant in its interaction, where it benefits children from higher socioeconomic levels more in grade 3 mathematics and in civics. Of the control variables, the school characteristics are significant in three out of the five cases; students tend to score worse if their schools cater to students of ethnic backgrounds or to students who are conflict victims. The effect of better governance is negative for grade 5 math in this specification as well. The model fit could be improved slightly; between 19% and 28% of the variance in scores is explained.

6.5 Robustness

In order to check whether small changes to the implementation index change the conclusion, the models are estimated replacing the overall implementation index first with the student- and then with the teacher index. This is done for the final random intercept model in each area and for model QSA3 and QSA4 using survey analysis techniques.

The results of the final random intercept model for the overall index were presented in Table 41 on page 197. Table 47 and Table 48 (starting on page 213) show how the results change when using the student- and teacher index, respectively. The overall conclusion based on Table 41 was a lack of statistical significance and thus a failure to reject the null hypothesis of no effect of the EN model. This conclusion is only slightly adapted, namely, in the case of language grade 3: Here, the model based on the student index results in a significant positive effect of EN implementation, as well as a significant, negative interaction term with socioeconomic level (the latter indicating that EN is more beneficial for children from poorer families). All of the significant effects from the final model are still significant in the robustness check, with comparable estimation coefficients (though the intercept estimates vary).

Table 49 and Table 50 present the estimation results based on the student- and teacher index dimensions, which should be compared to Table 42 on page 198. Note that the table based on the student index dimensions lacks dimension one, as teacher training and support is not measured on the student index. The results are again comparable: Most of the effects continue to lack statistical significance, though the interaction of dimension four (learning guides) and socioeconomic level is significant and negative in mathematics and civics in the student index dimension model (indicating that EN guides help students in schools of low average socioeconomic status more), and the interaction of dimension two (classroom organization) and socioeconomic level is significant and negative in language and civics in the teacher index dimensions model (with the same conclusions).

For the survey analysis models using the overall index, the results based on the student and teacher index can be found in Table 51 and Table 52, respectively (the results of the combined index were presented in Table 44 on page 206). The robustness checks confirm the conclusion that there is some evidence for a positive effect of EN implementation: Based on the student index, grade 3 test scores as well as civics scores improve with increased program implementation; according to the teacher index, that is the case in civics. The effect size is comparable to the one found with the overall index. The models based on the student- and teacher index also indicate a negative interaction with socioeconomic level.

Finally, Table 53 and Table 54 show how the results from Table 46 (page 208) change if the effect of individual model components are estimated for the dimensions of the student- and teacher index separately. Table 46 showed some evidence for a negative effect of classroom organization in grade 5 language, for a positive effect of community relations in language grade 3 and mathematics grade 5, and for a positive effect of learning guides in grade 3 mathematics. These results change somewhat when using only a student- or teacher index. In the former case, the negative effect of classroom organization in grade 5 language persists in the student index; yet, the coefficient is significant and *positive* in grade 3 and for civics for the student index, and for civics for the teacher index. A positive effect of community relations is not confirmed in either of the two sub-indices. The model based on the student index confirms a positive effect of learning guides in all specifications except language grade 5. Furthermore, the student index model suggests a negative effect of the dimension “student roles” for grade 3 mathematics tests. These effects are modified by interaction terms, which are for the student index in most specifications negative for socioeconomic level with dimensions two and four, and positive for dimensions three and five for one specification each. There are two, opposite, interactions with gender in language grade 5, namely, in dimensions three and five. For the teacher index, there are positive socioeconomic level interactions with dimension one, three, and four in one or two specifications, and negative interactions with dimension two in two specifications. Hence, though there are some patterns, the overall effect is not entirely clear, as will be discussed in the next section.

All put together, the robustness analysis upholds the general findings presented in this chapter. At the same time, it suggests that the effects of the individual model dimensions are not always robust, and that it does matter whose perspective on program implementation is considered. This, in turn, strengthens the case for using a combined metric of students’ and teachers’ perspectives in order to obtain the fullest possible picture.

Table 47 Robustness: Results of the final random intercept models for the overall student index, all grades and areas (department-level study)

	Language Grade 3		**Language Grade 5**		**Mathematics Grade 3**		**Mathematics Grade 5**		**Civics Grade 5**	
n (students)	234		343		226		289		348	
j (schools)	59		64		55		53		66	
Fixed part:										
EN student Index (ENSI)	3.79**	(1.25)	-1.35	(1.24)	1.85	(1.92)	0.51	(1.54)	1.27	(1.13)
Male	-20.73	(19.65)	-27.40	(19.96)	26.37	(21.91)	14.94	(18.55)	-21.00	(16.80)
ENSI*Male	0.14	(0.44)	-0.10	(0.42)	-0.62	(0.51)	-0.31	(0.41)	-0.36	(0.37)
Socioeconomic level	121.35*	(51.79)	-34.80	(47.27)	73.12	(78.06)	18.32	(68.93)	88.72*	(44.28)
ENSI*Socioec. level	-2.88*	(1.17)	1.12	(1.08)	-1.34	(1.78)	-0.62	(1.47)	-1.23	(1.01)
w/ ethnic students			-58.26**	(21.22)			-33.41	(22.87)	-45.63*	(18.98)
w/ conflict victims			-11.41	(13.99)			-39.99*	(16.12)	-10.98	(13.23)
Grand mean	154.97**	(54.73)	384.55***	(57.39)	220.22**	(84.09)	314.15***	(73.04)	258.84***	(51.95)
Random part (sd):										
School-level	34.76	(9.75)	31.42	(7.90)	66.24	(11.17)	38.67	(7.95)	29.84	(7.75)
Student-level	56.84	(3.44)	66.99	(3.04)	61.71	(4.25)	56.94	(3.61)	60.74	(2.76)
ICC (schools)	0.27		0.18		0.54		0.32		0.19	

Standard errors in parenthesis. ***p≤0.001; ** p<0.01; * p<0.05

Table 48 Robustness: Results of the final random intercept models for the overall teacher index, all grades and areas (department-level study)

	Language Grade 3		**Language Grade 5**		**Mathematics Grade 3**		**Mathematics Grade 5**		**Civics Grade 5**	
n (students)	250		372		251		314		372	
j (schools)	65		70		61		58		73	
Fixed part:										
EN teacher Index (ENTI)	-0.65	(1.36)	-0.63	(1.24)	-0.02	(1.96)	0.39	(1.35)	0.81	(1.11)
Male	-15.72	(19.89)	-19.14	(21.82)	33.39	(23.46)	13.87	(19.22)	-16.89	(16.97)
ENTI*Male	-0.02	(0.48)	-0.26	(0.49)	-0.81	(0.59)	-0.22	(0.46)	-0.49	(0.41)
Socioeconomic level	-43.47	(59.80)	-28.81	(51.65)	-16.52	(83.49)	-25.69	(62.68)	39.03	(48.23)
ENTI*Socioec. level	1.19	(1.34)	1.07	(1.20)	0.86	(1.91)	0.50	(1.39)	-0.24	(1.09)
w/ ethnic students			-45.15*	(20.22)			-28.89	(22.66)	-41.14*	(19.57)
w/ conflict victims			-11.79	(13.13)			-39.75**	(14.61)	-11.03	(12.60)
Grand mean	335.52***	(61.36)	345.99***	(54.93)	294.15**	(86.13)	314.36***	(60.57)	281.83***	(50.14)
Random part (sd):										
School-level	42.68	(7.80)	31.55	(7.37)	64.67	(10.18)	37.19	(7.25)	32.60	(6.27)
Student-level	56.22	(3.31)	66.41	(2.74)	61.71	(3.81)	56.67	(3.45)	59.89	(2.55)
ICC (schools)	0.37		0.18		0.52		0.30		0.23	

Standard errors in parenthesis. ***p≤0.001; ** p<0.01; * p<0.05

Table 49 Robustness: Results of the final random intercept models for the student index dimensions, all grades and areas (department-level study)

	Language Grade 3		**Language Grade 5**		**Mathematics Grade 3**		**Mathematics Grade 5**		**Civics Grade 5**	
n (students)	234		343		226		289		348	
j (schools)	59		64		55		53		66	
Fixed part:										
S Dimension 2 (Classroom)	1.45	(1.11)	-0.99	(1.33)	2.84	(1.56)	-1.18	(1.57)	0.73	(0.88)
S Dimension 3 (Commun.)	0.48	(0.77)	0.39	(0.75)	-0.46	(1.19)	0.37	(0.71)	-0.27	(0.66)
S Dimension 4 (Guides)	2.74	(1.92)	1.40	(1.82)	4.32	(2.22)	1.86	(1.68)	3.32*	(1.31)
S Dimension 5 (Roles)	-0.40	(2.06)	-0.59	(2.41)	-4.31	(3.23)	-0.39	(2.52)	-0.87	(1.72)
Male	-12.77	(28.76)	-0.61	(26.86)	33.13	(36.38)	33.53	(25.02)	-6.38	(24.69)
ENS D2*Male	0.24	(0.52)	0.41	(0.54)	0.05	(0.66)	0.99	(0.83)	-0.02	(0.57)
ENS D3*Male	0.75	(0.55)	-0.59	(0.47)	-0.14	(0.72)	-0.69	(0.47)	-0.08	(0.45)
ENS D4*Male	-1.10	(0.62)	-0.80	(0.64)	-0.25	(0.69)	-0.49	(0.64)	-0.53	(0.51)
ENS D5*Male	0.05	(0.95)	1.16	(1.09)	-0.37	(1.08)	-0.22	(0.99)	0.30	(0.91)
Socioeconomic level	221.09	(126.94)	16.34	(116.65)	288.22	(152.26)	122.22	(119.25)	265.21**	(83.59)
ENS D2* Socioec. level	-1.46	(0.87)	0.20	(1.00)	-2.17	(1.22)	1.56	(1.63)	-1.02	(0.65)
ENS D3* Socioec. level	-0.41	(0.83)	0.27	(0.74)	-0.35	(1.00)	0.19	(0.76)	0.93	(0.58)
ENS D4* Socioec. level	-2.41	(1.92)	-0.80	(1.88)	-4.97*	(2.28)	-3.57*	(1.79)	-3.85**	(1.34)
ENS D5*v. level	1.07	(2.19)	0.28	(2.40)	5.92	(3.14)	0.29	(2.60)	0.64	(1.66)
w/ ethnic students							-72.03**	(26.07)	-70.60***	(19.94)
w/ conflict victims							-31.16	(16.65)	-15.34	(12.75)
Grand mean	48.30	(127.15)	278.47*	(118.54)	22.48	(149.01)	257.89*	(123.52)	104.23	(86.06)
Random part (sd):										
School-level	35.26	(10.26)	35.11	(7.46)	55.87	(11.27)	35.06	(7.12)	0.53	(230.61)
Student-level	55.37	(3.34)	66.46	(2.96)	61.88	(4.33)	56.08	(3.46)	61.08	(3.00)
ICC (schools)	0.29		0.22		0.45		0.28		0.00	

Standard errors in parenthesis. ***p≤0.001; ** p<0.01; * p<0.05

Table 50 Robustness: Results of the final random intercept models for the teacher index dimensions, all grades and areas (department-level study)

	Language Grade 3		Language Grade 5		Mathematics Grade 3		Mathematics Grade 5		Civics Grade 5	
n (students)	250		372		251		314		372	
j (schools)	65		70		61		58		73	
Fixed part:										
T Dimension 1 (Training)	-1.10	(0.74)	-0.38	(0.65)	0.62	(1.13)	0.15	(0.67)	-0.80	(0.56)
T Dimension 2 (Classroom)	2.18	(1.31)	1.09	(0.94)	0.38	(1.74)	1.76	(1.11)	1.29	(0.87)
T Dimension 3 (Commun.)	0.07	(1.17)	-0.50	(0.93)	-1.36	(1.62)	0.93	(0.93)	0.06	(0.79)
T Dimension 4 (Guides)	-0.82	(1.08)	-0.99	(0.98)	-0.38	(1.52)	-0.72	(1.17)	-0.29	(0.82)
T Dimension 5 (Roles)	-2.65	(1.74)	0.50	(1.35)	0.04	(2.38)	-2.08	(1.67)	-0.55	(1.31)
Male	-13.24	(27.09)	-2.38	(28.69)	45.77	(32.61)	13.19	(25.78)	-17.74	(22.70)
ENT D1*Male	0.52	(0.45)	-0.84	(0.44)	-0.25	(0.55)	-0.55	(0.35)	0.27	(0.39)
ENT D2*Male	0.43	(0.73)	0.30	(0.55)	0.02	(0.70)	-0.56	(0.52)	-0.31	(0.53)
ENT D3*Male	-0.60	(0.57)	-0.50	(0.50)	-0.33	(0.69)	-0.59	(0.53)	-0.09	(0.54)
ENT D4*Male	-0.56	(0.87)	0.53	(0.62)	-0.25	(0.81)	0.79	(0.66)	-0.23	(0.60)
ENT D5*Male	0.04	(0.76)	-0.60	(0.68)	-0.15	(0.83)	0.78	(0.63)	0.13	(0.65)
Socioeconomic level	-81.54	(86.13)	-53.28	(64.12)	-59.03	(118.58)	-8.95	(88.00)	22.55	(55.33)
ENT D1*Socioec. level	1.16	(0.68)	0.18	(0.60)	-0.14	(0.99)	0.30	(0.68)	0.52	(0.54)
ENT D2* Socioec. level	-2.61*	(1.31)	-1.88*	(0.86)	-0.42	(1.59)	-1.50	(1.10)	-1.78*	(0.78)
ENT D3* Socioec. level	0.31	(1.15)	1.21	(0.91)	1.08	(1.63)	-0.14	(0.97)	0.67	(0.77)
ENT D4* Socioec. level	1.68	(1.27)	1.85	(0.97)	0.79	(1.50)	0.94	(1.16)	0.78	(0.81)
ENT D5*v. level	2.57	(1.68)	0.38	(1.29)	0.36	(2.29)	1.13	(1.68)	0.98	(1.23)
w/ ethnic students							-24.61	(25.55)	-40.88*	(18.32)
w/ conflict victims							-38.59**	(14.62)	-11.01	(12.42)
Grand mean	354.33***	(85.87)	320.43***	(64.40)	321.35**	(123.32)	282.55**	(85.92)	293.06***	(58.58)
Random part (sd):										
School-level	38.20	(8.65)	25.63	(8.74)	62.85	(9.97)	31.90	(7.32)	20.95	(10.92)
Student-level	55.60	(3.35)	65.91	(2.74)	61.61	(3.81)	55.93	(3.33)	60.49	(2.63)
ICC (schools)	0.32		0.13		0.51		0.25		0.11	

Standard errors in parenthesis. ***p≤0.001; ** p<0.01; * p<0.05

Table 51 Robustness: Results of the survey estimation models QSA3 using student index (department-level study)

	Language Grade 3			Language Grade 5			Mathematics Grade 3			Mathematics Grade 5			Civics Grade 5		
	β		se	β		se	β		se	β		se	β		se
EN Student Index	3.51	***	0.72	-1.06		0.57	4.22	**	1.19	0.46		0.62	1.70	**	0.49
male	-14.80		16.58	-33.66	*	13.18	33.05	*	11.91	7.56		11.00	-21.52	*	9.44
ENSI*male	0.01		0.36	0.11		0.27	-0.84	*	0.30	-0.15		0.27	-0.35		0.20
Socioeconomic Level	114.38	**	24.94	-24.63		18.29	128.67	***	31.35	27.35		29.35	120.18	***	16.67
ENSI*Socioeconomic level	-2.70	***	0.59	0.78		0.44	-3.18	***	0.81	-0.44		0.52	-1.76	***	0.37
w/ ethnic students	-12.91		9.91	-57.59	***	10.15	-34.64		17.53	-22.81		11.01	-53.30	***	8.77
w/ conflict victims	0.10		10.52	-8.40		8.80	27.69		14.05	-45.46	***	8.11	-13.97		8.23
Governance	0.98		1.03	-0.39		1.15	2.75	*	1.33	-2.88	*	1.10	-1.34		1.01
Intercept	167.94	***	32.55	373.78	***	24.84	126.69	*	50.05	298.35	***	30.65	233.48	***	22.75
n	234			343			226			289			348		

***p≤0.001; ** p<0.01; * p<0.05

Table 52 Robustness: Results of the survey estimation models QSA3 using teacher's index (department-level study)

	Language Grade 3			Language Grade 5			Mathematics Grade 3			Mathematics Grade 5			Civics Grade 5		
	β		se	β		se	β		se	β		se	β		se
EN Teacher Index	-0.29		0.73	-0.59		0.54	1.39		1.31	0.75		0.62	1.78	**	0.64
male	-9.42		15.77	-7.89		15.59	35.73	*	14.48	15.40		12.87	-12.30		8.20
ENTI*male	-0.08		0.36	-0.44		0.33	-0.84		0.42	-0.27		0.34	-0.57	*	0.21
Socioeconomic Level	12.83		28.36	-19.33		21.42	66.40		42.11	29.39		29.44	100.79	***	27.60
ENTI*Socioeconomic level	0.07		0.62	0.80		0.51	-1.21		1.00	-0.49		0.59	-1.63	**	0.58
w/ ethnic students	-48.66	***	11.00	-52.72	***	9.19	-68.50	***	13.31	-20.33	*	8.62	-53.82	***	8.44
w/ conflict victims	-15.76		9.68	-13.99		8.04	-1.19		12.87	-53.95	***	8.18	-10.44		7.04
Governance	0.46		0.91	-0.81		0.93	1.37		1.38	-3.05	***	0.79	-0.41		0.88
Intercept	322.32	***	31.89	348.11	***	24.95	245.84	***	54.53	289.05	***	31.49	237.21	***	30.07
n	250			372			251			314			372		

***p≤0.001; ** p<0.01; * p<0.05

Table 53 Robustness: Results of the survey estimation models QSA4 using student index (department-level study)

	Language Grade 3			Language Grade 5			Mathematics Grade 3			Mathematics Grade 5			Civics Grade 5		
	β		se	β		se	β		se	β		se	β		se
S Dimension 2 (Classroom)	1.73	**	0.51	-2.29	*	0.82	3.51	***	0.89	-1.42		0.71	1.04	*	0.47
S Dimension 3 (Commun.)	0.45		0.39	0.27		0.41	0.40		0.72	0.27		0.38	-0.37		0.46
S Dimension 4 (Guides)	2.39	**	0.55	1.14		0.89	4.65	***	0.60	1.57	*	0.60	3.65	***	0.66
S Dimension 5 (Roles)	-0.45		1.08	-0.05		1.21	-4.91	*	2.04	-0.26		1.58	-0.89		0.84
male	-10.42		14.91	-6.14		14.74	47.89		22.54	29.63		15.24	-4.61		12.58
ENS D2*male	0.27		0.25	0.24		0.33	0.35		0.39	1.07		0.69	0.02		0.38
ENS D3*male	0.72		0.34	-0.67	*	0.29	-0.60		0.43	-0.58		0.32	-0.07		0.29
ENS D4*male	-0.85		0.36	-0.73		0.37	-0.19		0.46	-0.69		0.39	-0.50		0.25
ENS D5*male	-0.42		0.61	1.65	*	0.62	-0.50		0.67	-0.09		0.75	0.13		0.55
Socioeconomic level	227.40	***	51.26	-11.70		57.06	335.48	***	67.48	63.19		38.24	301.69	***	41.61
ENS D2*Socioecon. level	-1.42	***	0.30	0.97		0.55	-3.06	***	0.54	0.50		0.75	-1.28	***	0.28
ENS D3*Socioecon. level	-0.33		0.40	0.67		0.41	-0.30		0.52	0.15		0.33	0.95	*	0.34
ENS D4*Socioecon. level	-2.44	***	0.49	-0.97		0.92	-5.60	***	0.68	-2.00	*	0.70	-4.21	***	0.74
ENS D5*Socioecon. level	0.94		1.16	-0.25		1.15	6.62	**	1.89	1.27		1.61	0.83		0.70
w/ ethnic students	-26.32	*	12.91	-74.17	***	12.66	-33.64		17.41	-41.42	**	13.60	-73.00	***	11.81
w/ conflict victims	3.25		10.23	-20.27	*	8.37	29.73	*	13.20	-44.01	***	8.98	-16.92		8.68
Governance	0.75		1.07	-0.15		1.18	1.51		1.29	-3.17	**	1.18	-1.55		0.95
Intercept	54.51		55.80	367.26	***	59.85	-74.36		71.24	276.48	***	38.55	66.15		40.62
n	234			343			226			289			348		

***p≤0.001; ** p<0.01; * p<0.05

Table 54 Robustness: Results of the survey estimation models QSA4 using teacher index (department-level study)

	Language Grade 3			Language Grade 5			Mathematics Grade 3			Mathematics Grade 5			Civics Grade 5		
	β		**se**	**β**		**se**	**β**		**se**	**β**		**se**	**β**		**se**
T Dimension 1 (Training)	-0.68		0.50	-0.37		0.32	1.36		0.79	0.35		0.42	-0.89	**	0.29
T Dimension 2 (Classroom)	0.99		0.85	0.40		0.59	1.27		1.05	1.44		0.70	1.26	*	0.49
T Dimension 3 (Commun.)	0.55		0.57	-0.44		0.44	-0.81		0.99	0.52		0.44	0.25		0.34
T Dimension 4 (Guides)	-0.65		0.54	-0.95		0.55	-1.05		0.75	-0.84		0.81	-0.39		0.45
T Dimension 5 (Roles)	-1.93		1.04	0.63		0.80	-1.28		1.64	-1.51		1.09	-0.94		0.71
male	-2.47		15.86	11.65		20.14	56.75	*	20.02	16.63		17.19	-14.33		11.10
ENT D1*male	0.30		0.24	-1.00	*	0.31	-0.65		0.34	-0.63	**	0.22	0.20		0.28
ENT D2*male	0.38		0.51	0.51		0.38	-0.03		0.45	-0.47		0.34	-0.30		0.33
ENT D3*male	-0.61		0.35	-0.68		0.35	-0.54		0.40	-0.61		0.38	-0.17		0.36
ENT D4*male	-0.51		0.60	0.46		0.38	0.35		0.52	0.78		0.44	-0.10		0.39
ENT D5*male	0.15		0.39	-0.77		0.39	-0.62		0.50	0.65		0.43	0.00		0.37
Socioeconomic level	-40.42		36.50	-73.00	*	28.21	-13.63		68.78	-4.09		50.83	20.86		25.22
ENT D1*Socioecon. level	1.26	**	0.36	0.28		0.29	-0.46		0.60	-0.08		0.44	0.74	*	0.27
ENT D2* Socioecon. level	-1.57		0.76	-1.44	*	0.52	-1.15		0.81	-1.32		0.65	-1.72	***	0.44
ENT D3* Socioecon. level	-0.42		0.55	1.16	*	0.45	1.09		0.99	0.30		0.46	0.53		0.35
ENT D4* Socioecon. level	0.88		0.63	1.74	**	0.55	0.42		0.60	0.92		0.79	0.53		0.46
ENT D5* Socioecon. level	1.84		1.04	0.22		0.75	1.65		1.46	0.78		1.08	1.34		0.67
w/ ethnic students	-60.47	***	14.96	-29.96	**	9.02	-62.00	**	21.55	-11.06		11.36	-49.66	***	9.82
w/ conflict victims	-26.29		12.15	-7.56		9.16	-2.85		14.00	-47.43	***	8.70	-17.45	*	7.94
Governance	0.24		0.99	-0.84		0.94	1.35		1.36	-2.24	*	0.85	-0.79		0.76
Intercept	371.57	***	32.95	366.62	***	25.40	299.77	***	72.89	295.29	***	51.27	311.42	***	24.58
n	250			272			251			314			372		

***p≤0.001; ** p<0.01; * p<0.05

6.6 Discussion

The results of the department-level analysis may lack the clarity of the country-level study, but they do provide some insights into the effects of the EN model on learning, and some limited support for the research hypotheses.

The first research hypothesis related to the program’s outcomes states that EN schools are more effective than conventional schools in improving students’ numeracy, literacy, and civic competencies. Table 55 compiles the evidence from this chapter: It shows the *ceteris paribus* main estimation coefficients on EN implementation for the full index and its dimensions, for the multilevel and survey estimation models. Note that this does not include the effects of the interaction terms discussed below; *ceteris paribus* thus means that these are the estimated effects for girls in a school with the average socioeconomic level NSE1.

The first half of the table shows what was discussed in section 6.3.3: EN implementation has no statistically significant effect on learning outcomes in the multilevel models, neither as measured by the implementation index in its entirety, nor by its individual dimensions. As was argued, this should not be seen as evidence for a lack of effectiveness of the EN model, given that the sample is small for a multilevel model in *absolute* terms; the largest part of the variance lies at the student-level where no control variables are available; and the estimation procedure used does not take into account the fact that data is available for as much as 50% of the reference population.

The second half of Table 55 shows the estimated main effect of EN implementation based on survey estimation techniques, which do take into account the large *relative* sample size. While not as unambiguous as the results from the country-level study, these results largely support the research hypothesis. At least for the cases of civic competencies and grade 3 mathematics, as well as for some individual dimensions, the estimated effect of EN implementation is positive and statistically significant. The coefficient is the estimated effect for a change of one percentage point of the implementation index. As shown in Table 56 below, this small-looking coefficient adds up to a large effect of implementation. The observed range of the index is about 60 points, with a standard deviation of 12. A girl in a school with the lowest empirically observed value of the implementation index can expect to score 167 points on the grade 3 mathematics exam, while a girl in a school with an average implementation score can expect to score 312 points, and a girl in a school with the highest observed score can expect to score 391 points, all other things being equal. In the case of civic competencies, the *ceteris paribus* prediction for a girl in a school with the lowest implementation index is 223 points, in a school with average implementation, 312 points, and in a school with the highest observed score, 360 points. The standard errors and confidence intervals for these estimates are relatively large, but the low-end and high-end estimates clearly differ from each other. The estimated c.p. difference between a very conventional school and a school with a high level of EN implementation is thus 224 points on the mathematics grade 3 exam, and 137 points on the civics exam. As per the discussion of ICFES-defined achievement levels (see section 4.4), this difference is enough to bridge up to two of the four achievement levels. In this sense, for the case of grade 3 mathematics and grade 5 civics, the effect of EN is even stronger in the department-level study than in the country-level study, and of clear practical significance. These results thus support the hypothesis that EN implementation matters, and that the model is a powerful tool to improve learning outcomes. The large caveat is, of course, that no statistically significant effect could be found for language exams, or for grade 5 mathematics—and that the effect on grade 3 mathematics and on civics exam scores is significant only at the 5% level (and not at all in the multilevel model). Whether this is due to the sample size or to a lack of a significant effect in the population remains an open question.

Table 55 Overview of ceteris paribus effects of EN implementation scores in department-level study models

	Language Grade 3	Language Grade 5	Mathematics Grade 3	Mathematics Grade 5	Civics Grade 5
RI model					
Full index	0.92	-1.60	0.79	0.41	1.23
D1 (Training)	-1.05	-0.50	-0.31	0.05	-0.54
D2 (Classroom)	1.06	-0.56	0.46	1.31	0.24
D3 (Community)	1.26	0.23	0.12	1.26	0.62
D4 (Guides)	-0.02	-0.13	2.27	0.26	1.05
D5 (Roles)	-1.59	0.00	-2.74	-3.78	-1.28
Survey model					
Full index	1.34	-1.36	3.70 *	0.74	2.26 **
D1 (Training)	-0.77	-0.31	-0.26	0.14	-0.65
D2 (Classroom)	0.53	-2.07 **	1.04	0.24	0.20
D3 (Community)	1.48 **	0.27	1.45	0.89 *	0.72
D4 (Guides)	-0.04	0.16	2.70 *	0.75	1.09
D5 (Roles)	-1.91	0.25	-4.46	-2.15	-1.52

***p≤0.001; ** p<0.01; * p<0.05

Table 56 Predicted test scores for girls in schools with minimum, mean, and maximum observed implementation index scores (all other variables set at zero), based on survey analysis model

	Mathematics Grade 3				**Civics Grade 5**			
	Margin	Std.Err	95% Conf. Interval		Margin	Std.Err	95% Conf. Interval	
Minimum	166.6	55.1	54.3	278.9	223.2	29.1	165.0	281.5
Mean	312.1	18.0	274.9	349.4	312.2	7.7	296.7	327.7
Maximum	390.7	40.0	307.4	474.0	360.3	14.9	330.3	390.2

Table 55 also helps to assess the effect of EN implementation in the individual model dimensions. No effect is found in the multilevel models. In the survey estimation models, three out of the five dimensions have a statistically significant effect in at least one area or grade. The results suggest that implementing the community-relations elements of the model improves scores in language grade 3 and mathematics grade 5, that implementing the learning guides improves mathematics grade 3 scores, and that implementing classroom organization elements *decreases* grade 5 language scores. This variation will be discussed below in the context of interaction terms.

The second research hypothesis posited that the effect of EN is stronger for children from disadvantaged socioeconomic backgrounds (or, in other words, that the effect of socioeconomic status is smaller in EN schools than in conventional schools). A significant and negative coefficient on the interaction term of EN implementation and socioeconomic status is required to reject the null hypothesis of no difference. The estimation coefficients on the interaction term for the different models are combined in Table 57 (together with estimated main effect of socioeconomic level). Again, the overview shows that the null hypothesis fails to be rejected based on the multilevel models, but can be rejected for some specifications of the survey estimation models, namely, again for grade 3 mathematics and civics (in the case of the full implementation index).

The estimated exam score differences between students in schools of different average socio-economic levels and different levels of EN implementation are summarized in Table 58. The baseline is a student of the lowest socioeconomic level and implementation score; the differences in the index refer to a change in the implementation index of one or two standard deviations (i.e., 11.8 and 23.6 points). The table shows that students in schools with the lower two socioeconomic levels tend to do better with a higher level of EN implementation, while the opposite is the case for students in schools of higher socioeconomic levels. The difference is particularly large for the most disadvantaged students. As already discussed in the country-level analysis, it is not possible based on the available data to distinguish the effect of an individual student's socioeconomic status from the effect of a school's average. However, the strong effect that EN implementation has in schools with the lowest average socioeconomic level supports the second research hypothesis: For the case of mathematics grade 3 and civic competencies, the null hypothesis of no effect of socioeconomic status can be rejected.

Table 57 Overview of ceteris Paribus effects of EN implementation-socioeconomic level interaction terms in department-level study models

	Language Grade 3	Language Grade 5	Mathematics Grade 3	Mathematics Grade 5	Civics Grade 5
Main effect of socioeconomic level					
RI models					
Full index	14.20	-53.19	19.17	-24.29	67.52
Dimensions	-48.68	-110.05	60.38	8.96	63.41
Survey models					
Full index	56.13 *	-32.40	122.07 **	33.30	113.43 ***
Dimensions	-36.32	-116.10 *	102.42	27.09	73.04
Interactions of implementation and socioeconomic level					
RI models					
Full index	-0.16	1.70	0.05	0.23	-0.85
D1 (Training)	1.23	0.40	1.10	0.73	0.45
D2 (Classroom)	-1.63	-0.86	-0.40	-1.01	-0.82
D3 (Community)	-0.87	0.61	-0.64	-0.55	0.11
D4 (Guides)	0.90	1.43	-2.70	-0.75	-0.92
D5 (Roles)	1.78	0.85	4.04	2.69	1.64
Survey models					
Full index	-1.11	1.22	-2.88 **	-0.59	-2.09 ***
D1 (Training)	1.28 **	0.43	0.80	0.27	0.62
D2 (Classroom)	-1.15 *	0.21	-1.29	-0.84	-0.79
D3 (Community)	-1.02 *	0.64	-0.79	-0.06	0.03
D4 (Guides)	0.21	0.59	-3.67 **	-0.63	-1.31
D5 (Roles)	2.09	0.84	6.35 **	1.94	2.06 *

***p≤0.001; ** p<0.01; * p<0.05

Table 58 Estimated joint marginal effect of socioeconomic status and a one- and two-standard deviation change in the EN implementation index, based on survey analysis model (department-level study)

	Mathematics Grade 3			**Civic Competencies Grade 5**		
	Low index	**ENI +1SD**	**ENI +2SD**	**Low index**	**ENI +1SD**	**ENI +2SD**
NSE1	(baseline)	43.7	87.3	(baseline)	26.7	53.3
NSE2	122.1	131.7	141.4	113.4	115.4	117.4
NSE3	244.1	219.8	195.5	226.9	204.2	181.5
NSE4	366.2	307.9	249.6	340.3	293.0	245.7

The joint effect of EN implementation and socioeconomic level differs by index dimension. In the case of grade 3 mathematics, where the interaction term of the overall index was negative and statistically significant, the same effect can be confirmed for dimension four (learning guides)—implementing this element is particularly beneficial in schools with low average socioeconomic status. However, the interaction term is positive for dimension five (roles of students), which is also the case for civics exams. This particular dimension of the model may thus be beneficial only for students in schools with higher average socioeconomic status. Interestingly, the interaction between teacher training and socioeconomic status is also positive for grade 3 language, suggesting that here, too, teacher training has a strong positive effect (only) in schools with higher average socioeconomic levels. Two other dimensions (classroom organization and community relations) have the expected negative effect for grade 3 language.

Finally, the third research hypothesis posits that the EN model helps to diminish the differences between genders. Table 59 shows that no relationship between gender and implementation can be established based on the multilevel model, and that the survey estimation results provide some evidence that the EN model is particularly beneficial for girls. The estimations, based both on the full index and on the index dimensions, show statistically significant better results for boys in mathematics grade 3, and the estimations based on the full index shows better results for girls in civic competencies. At the same time, the interaction term based on the full index has a negative sign only for civics. In this specification, the EN model thus benefits girls but *increases* gender differences. When disaggregating the EN effect into its five components, the significant interaction terms are negative for dimension one (training) in grade 5 language and mathematics, for dimension three (community relations) in mathematics grade 3, and for dimension four (learning guides) in language grade 3. Only in one dimension do the results favor boys, namely, in the case of classroom organization for both language exams.

Table 59 Overview of ceteris Paribus effects of EN implementation-gender interaction terms in department-level study models

	Language Grade 3	**Language Grade 5**	**Mathematics Grade 3**	**Mathematics Grade 5**	**Civics Grade 5**
Main effect of male					
RI models					
Full index	-20.28	-20.49	31.03	16.70	-19.81
Dimensions	-2.00	5.59	49.05	21.62	-10.02
Survey models					
Full index	-12.34	-14.42	35.20 *	15.41	-17.60 *
Dimensions	9.00	15.82	52.08 *	28.41	-5.80
Interactions of implementation and male					
RI models					
Full index	0.08	-0.28	-0.80	-0.34	-0.43
D1 (Training)	0.41	-0.75	-0.32	-0.52	0.14
D2 (Classroom)	0.72	1.01	0.37	0.16	-0.09
D3 (Community)	0.06	-0.56	-0.29	-0.64	0.04
D4 (Guides)	-1.48	-0.38	-0.33	-0.03	-0.60
D5 (Roles)	0.49	-0.28	-0.66	0.72	0.26
Survey models					
Full index	-0.11	-0.31	-0.94	-0.28	-0.46 *
D1 (Training)	0.24	-1.00 **	-0.24	-0.67 **	0.07
D2 (Classroom)	0.87 *	1.08 *	0.78	0.46	0.02
D3 (Community)	-0.33	-0.55	-1.06 *	-0.51	-0.02
D4 (Guides)	-1.18 *	-0.49	-0.08	-0.33	-0.52
D5 (Roles)	0.06	-0.07	-1.20	0.66	0.01

***p≤0.001; ** p<0.01; * p<0.05

In short, evidence for the third hypothesis is mixed, though there is some indication that EN implementation does favor girls. Given that the main effect is not significant in many of the specifications, it remains unclear whether the model has thus an equalizing or an un-equalizing effect for differences in learning outcomes between boys and girls.

There is a potential sample selection bias arising from using only data of schools that correctly reported testing results, which means that this analysis is based on data from generally higher-quality schools. As already discussed on several occasions in this dissertation, this probably leads to an overestimation of the effect of the EN model as well as of its potential to reduce the socioeconomic achievement gap. The respective true effects in the population might thus be smaller than suggested by this analysis, though given the large size of the coefficients it seems unlikely that overall school quality alone can explain all of the variation.

All put together, the department-level analysis shows that the EN model, if properly implemented, can indeed improve learning outcomes, at least in grade 3 mathematics and civic competencies. It is important to remember that for language and grade 5 mathematics, there is no evidence that the model does *not* improve learning outcomes, or that learning outcomes are better in conventional schools – the data simply does not allow one to draw conclusions about differences between the models for these areas. The implications of these findings, as well as the need for further research, will be the subject of the next, and final, chapter.

7 Concluding Discussion

This final chapter provides a concluding discussion of the research project. First, section 7.1 offers a synthesis of the study's findings, bringing together the main conclusions from each of the study sections. Then, section 7.2 discusses the limitations of the research by describing threats to internal and external validity. Section 7.3 lines out areas for future research. Finally, section 7.4 gives policy recommendations based on the findings.

7.1 Synthesis of findings

The starting point for this study was a puzzle about the Colombian primary school system: How can it be that learning outcomes, as measured by international standardized tests, are so poor, while at the same time the country prides itself with the development of a progressive, student-centered school model, Escuela Nueva, that has been in place for decades and is used in half of the country's primary schools? There are only two possible explanations: Either the EN model is not as effective as it is portrayed to be, or it is not being used as extensively as suggested.

Two sets of questions thus needed to be answered. The first one is related to the extent of EN program implementation: How many schools actually use the model, compared to the number of schools that report using it? How faithfully are the individual elements implemented, and which elements are adapted or left out? And how different are EN schools in practice from conventional schools? The second set of questions asks if learning outcomes are better in EN schools than in conventional schools—in general terms, and specifically, given the persistent achievement gaps in the country, for children from disadvantaged socioeconomic backgrounds and for girls. The study employed a mixed methods design to answer these questions, with a strong focus on quantitative multilevel modeling. The analysis was based on secondary data on learning outcomes (the standardized test Pruebas SABER, administered to all 3rd and 5th grade students in 2013); administrative data, which includes an official Escuela-Nueva classifier; and primary data on program implementation that was collected from a representative sample of rural schools in one Colombian department, Quindío.

7.1.1 Program implementation

The results of the implementation evaluation show that there are indeed large differences in classroom practices across the department that was studied. Even though most of Quindío's rural schools are officially classified as EN schools, they implement, on average across schools, just over 60% of the model's elements. Implementation varies between the different aspects of the model, too. While the model dimensions "classroom organization" and "roles of students" are implemented comparatively faithfully (72% and 73% of the respective elements are being used on average across schools), in other elements of the model there is a larger departure from the ideal-typical EN. Most prominently, departures are seen in the dimension "teacher training and support", where only 43% of elements are in place across schools, and in the dimension "community relations", where 57% of the elements are being used. In the remaining fifth dimension, "learning guides", schools typically use 66% of the elements.

Schools that are officially classified as EN do, in fact, implement a larger share of the model's elements than conventional schools, but variation within each group is very large and the difference in means is small (62% versus 51% of the elements). The difference between the groups is weakly statistically significant. In the individual model dimensions, the difference between EN schools and conventional schools is particularly pronounced in the dimension "classroom organization” and “community relations” (with a difference of 21 and 13 percentage points, respectively)—these are the only dimensions where the school types are distinguishable statistically. These findings support the evidence presented by earlier studies (McEwan 1998; Forero-Pineda, Escobar-Rodriguéz, and Molina 2006) that found that EN schools are more likely than conventional schools to use EN methodologies, but that the difference between the school types is not clear-cut.

The qualitative evidence backs up the finding that model implementation is very heterogeneous. Additionally, it shows that the ways in which the same elements are being used in different schools, and the reasons for why they are being used (or not used), vary widely. EN sets out to be a flexible toolbox for teachers, and invites adaptations to the model to accommodate local necessities. However, flexibility in the implementation does not mean that “anything goes”. Not every adaptation by a teacher can automatically be assumed to be based on local necessities, or to be coherent with the rest of the EN model. The wide range in which some of the EN instruments are being used even within the small qualitative sample raises the question of what exactly constitutes the use of the tools in coherence with the EN model, i.e., where the line is to be drawn between an appropriate local adaptation and the failure to use the model correctly. A clear and detailed description of what the model entails, together with explanations about how each of the model’s elements fit into the overall model and what their specific roles, is lacking.

It is telling in this context that the dimension “teacher training and support” is the one with the largest deficit in proper implementation. Arguably, adequate teacher training and in-service support has to be the cornerstone of the program: How else can a teacher know how to adequately use the program elements, what their purposes and key features are, and how they can be adapted without giving up on the role they play within the EN package? This is not to say that teachers need to be drilled to be mechanical executors of the EN model. Still, they need a functional training and support system that allows them to effectively take advantage of the EN model to deal with the wide range of challenges that come with teaching in rural schools.

Put together, the empirical evidence clearly supports the hypothesis that the model is not being properly implemented, and backs up the assertion that the official EN classifier is not a precise way of identifying EN schools. It also reveals that variation in program implementation is even larger than suggested by previous studies, given that for virtually every element of the model, there is variance in the way in which it is being used.

7.1.2 Learning outcomes

The second set of questions asked how the EN model affects learning outcomes. This study adopted several strategies to answering these questions. On the one hand, learning outcomes were analyzed on the national level based on the exam scores of a total of over 810,000 students in 21,235 schools, using the official EN classifier to identify EN schools. On the other hand, learning outcomes were also analyzed for the department of Quindío based on a sample of 76 schools (around half of the department's rural primary schools) and 1,068 students, identifying EN schools with the EN implementation index constructed from the primary data on program implementation.

Results from the two levels of analysis are largely coherent: the EN model is indeed associated with improved learning outcomes, controlling for a range of other factors of influence. For the country-level analysis, a statistically significant positive effect was found for all grade levels and subject areas under study (language grades 3 and 5; mathematics grade 3 and 5; and civic competencies grade 5). Country-wide, the expected *ceteris paribus* difference in exam scores between a student in an EN school and a student in a conventional school is between 10 and 23 points, depending on the exam (the mean score in the sample is close to 300 points for all exams, the standard deviation is around 75 points). This effect is comparable in size with the effect of the difference of one socioeconomic level (out of four official socioeconomic levels), and up to a third of the size of the distance between two achievement levels as defined by ICFES. When using the EN implementation index instead of the official EN classifier, the effect is even stronger. The expected *ceteris paribus* difference in exam scores between a girl in a school with a very low implementation index and one in a school with a very high implementation index is between 140 and 220 points, which is an enormous difference that is enough to bridge two achievement levels.

That being said, in the department-level study based on the implementation index, the effect of overall EN implementation was only significant in grade 3 mathematics and grade 5 civics. This is not to say that no effect exists in the other areas; there is also no evidence that students in EN schools perform *worse*. The lack of a significant effect in the other grades and areas may just be due to limitations of the dataset—a hypothesis that is supported by the fact that other variables generally known to have an effect, such as gender or socioeconomic level, also lack significance in the other grades and areas. However, it cannot be ruled out that the significance of the EN effect in other grades or areas at the country-level is due to the poor identification of EN schools, and disappears once actual program implementation is taken into account.

Because of some of the limitations of the database (see discussion on limitations below), the estimated effect of the EN model on learning outcomes is likely to be biased upwards: Data is available only from schools that managed to report them properly, and it seems likely that EN schools are of higher quality even before implementing the EN model. That implies that the estimated effect of the EN model may be confounded with the effect of higher-quality schools, and the isolated effect of the EN model may be smaller than suggested by the analysis.

The finding that the EN model helps to improve learning outcomes confirms previous studies on the model that came to similar conclusions (for academic achievements: Rojas and Castillo 1988; Velez 1991; Psacharopoulos, Rojas, and Velez 1992; McEwan 1998; additionally for civic competencies: Pitt 2002; Forero-Pineda, Escobar-Rodriguéz, and Molina 2006). Apart from updating these dated studies, adding evidence about the effect of actual program implementation, and extending the analysis to the level of the whole country, this study provides several important insights about differences in the effect of EN across Colombia. Specifically, the country-wide multilevel analysis revealed significant differences in the effect of the EN model on learning across municipalities and across departments. The estimated differences in the slopes are considerable: Holding everything else constant, the effect of EN schools varies across municipalities with a standard deviation of 18 points. In practice that means that in 95% of municipalities the effect of EN on grade 3 language scores lies between -21 and 51 points. In other words, there are municipalities where students in EN schools perform worse than students in conventional schools (other things being equal), but there are also municipalities where the score difference for students in the different school types is very large in favor of the EN model. A similar effect is found for departments, at which level the EN effect varies with a standard deviation of 9 points: in two thirds of the departments, the effect of being in an EN school amounts to a difference in language grade 3 exam scores of between 6 and 24 points. These are considerable variations across departments and municipalities, which suggest that local policies and support for the EN model may matter a lot for the success of the model. Unfortunately, no data was available to further investigate what determines the size of the municipality- or department-level effect of the EN model. A preliminary analysis suggests that the EN effect tends to be stronger in departments with a longer history of program implementation, and weaker in departments without support for the model, though the correlation is far from clear.

Colombia's education system is marked by large achievement gaps between students from different social backgrounds. Improving the equity of the system is not just an important objective from the perspective of global development, but also one of the pronounced goals of the Ministry of Education (Ministerio de Educación Nacional de Colombia 2016). The second part of the research question related to program outcomes addresses this issue and asks whether the EN model helps to diminish the differences between children from different socioeconomic backgrounds, or, put differently, whether the effect of the EN model is particularly strong for children of lower socioeconomic levels. This indeed seems to be the case: Both in the country-level study and in the department-level study, the predicted exam score for students in schools of low average socioeconomic level is higher if the school implements the EN model, but the respective expected score is lower for students in schools with high average socioeconomic level. The important caveat here is that the data only allows for the identification of socioeconomic status at the school-level; the effect at the level of individual students need not necessarily be the same (this would be a potential ecological fallacy). That being said, if a school's average socioeconomic level is very low, it implies that the majority of students in that school need to be of that low socioeconomic level. Also note that at the department-level the effect is only significant for grade 3 mathematics and civic competencies, with the same implications discussed above.

As was the case for the main effect of the EN model, the finding that EN helps reduce socioeconomic achievement gaps may be partly the result of cofounding with the effect of higher-quality schools. The true effect of the interaction of EN and socioeconomic status may thus be smaller than suggested by the estimation results presented in this study.

Socioeconomic background is not the only source of achievement gaps. As was depicted in Figure 2 on page 9, there are also considerable differences between genders. The question is thus whether the EN model helps to diminish these gender gaps. The answer is affirmative only in the case of the country-level study, where the model benefits boys more in areas where girls typically score better (language and civic competencies), and it benefits girls more in areas where boys tend to do better (mathematics). This is in line with the model's mission to address the specific needs of every child. A similar effect could not be confirmed for the department-level study; here, the effect of the model differs between boys and girls only in civic competencies, where it *increases* the advantage of girls. As before, this lack of a significant, equalizing effect in comparison with the country-level study might be due to the limited statistical power of the department-level study, to a different population effect for the case of Quindío, or to the fact that the effect found in the country-level study is actually an artefact of improper EN-identification.

Some additional insights can be gained from the department-level study by disaggregating the overall implementation index into its five dimensions (teacher training, classroom organization, community relations, learning guides, and roles of students). Practically, this endeavor is limited by the relatively small amount of available observations and the fact that estimating the effect of individual dimensions while controlling for implementation of the other EN elements requires many degrees of freedom, especially when interactions with socioeconomic level and gender are introduced. With that in mind, there is some evidence that the effect of individual dimensions might vary. For instance, the main effect of the dimensions "community relations", "classroom organization", and "learning guides" seems to be positive in at least one grade or area. However, the main effect of "roles of students" is estimated to be negative in one case (grade 5 mathematics). The main effects of the individual dimensions are in some cases modified by interactions with socioeconomic level and gender. The dimensions "Classroom organization" and "learning guides" seem to be particularly beneficial for students in schools of lower average socioeconomic levels, while "teacher training" and "student roles" seem to be particularly beneficial for students in schools of high average socioeconomic levels. Similarly, higher implementation of the elements of “teacher training” and “student roles” seem to benefit girls more, while higher implementation in the dimension “classroom organization” benefits boys more. All of the discussed dimension effects are, however, only based on one or two significant effects per dimension (for instance, in grade 5 mathematics and language, but not in other areas or grades). Additionally, robustness tests which used only information provided by students or only information provided by teachers confirmed only some of these effects, while modifying others. Therefore, the dimension-based results need to be taken with sufficient caution.

In conclusion, the study found support for all of its hypotheses relating to the effect of the EN model on learning outcomes: The model does indeed have a positive effect on test scores; the effect is stronger when considering actual implementation levels instead of the imprecise official classifier; the model tends to benefit students in schools of lower socioeconomic level more than students in schools of higher socioeconomic level; and the model tends to decrease differences in achievement scores between genders, though this last effect could not be confirmed in the department-level study.

There is a list of caveats to consider which may limit the validity of these results. A discussion of these limitations is the goal of the next section.

7.2 Limitations

Some of the limitations of previous research on the topic could be addressed, but a number of issues remain. These can be grouped into threats to internal validity and threats to external validity, and are discussed in the following pages.

7.2.1 Threats to internal validity

Internal validity refers to the extent to which causal conclusions can be drawn based on a study design. For the quantitative part of the study, major limitations in this sense relate to possible endogeneity, the reliance on a mix of data sources, sample selection, omitted variables, and the implementation measurement instrument / identification of program implementation.

First, the research design does not preclude the possibility that the choice of schools is endogenous, i.e. that students with specific characteristics self-select into certain school models, or perhaps more relevant, that schools or teachers with certain characteristics are more likely to adopt the EN model. If school choice on part of the students is not random, not being able to control for a students' ability upon entry and other unobserved characteristics may overestimate the effect of the school. However, the actual school choice for students is limited in many areas (especially in rural areas where schools are small and distances are large, students do not have the practical choice between going to an EN or to a conventional school). Therefore, endogeneity may not be a major issue on the student-level at least for rural areas. It is, however, likely to be an issue on the school-level: Whether or not a school implements the EN method, and to what extent, is not randomly determined, but a result of the decision of the teachers, local authorities, regional policy makers, and the pressure from other stakeholders. Hence, it is likely that EN (and EN teachers) are different in observed and unobserved ways from conventional schools even before the program is being implemented. This implies that the statistical effects that were found may not actually be an indication of the causal effect of the EN model in improving learning outcomes. Rather, they may reflect differences that were already in place.

The available data does not allow for drawing firm conclusions about the effect of the potential endogeneity. However, it is possible to make some statements about the likely direction of biases in estimates that may arise from it. This discussion parallels the discussion of a potential sample selection bias resulting from the fact that a large number of schools had to be excluded from the analysis due to reporting errors (see Annex A, section 1.2 for details). Essentially, it seems likely that schools implementing the EN model are higher-quality schools to begin with (counting with better motivated teachers, for instance). This would indicate that the effect of EN implementation might be overestimated (due to confounding with high school ex-ante school quality). Additionally, schools of higher average socioeconomic level are likely to be schools of better unobserved quality. As socioeconomic status is also correlated with better learning outcomes, the effects of socioeconomic status and of its interaction with EN implementation are probably overestimated, again due to confounding the effect of EN with the effect of school quality.

Second, a major limitation arises from the fact that this project relies on the combination of primary and secondary data sources. In particular, this is an issue for the department-level analysis which takes the *Pruebas SABER* test results as a starting point for the collection of additional data on program implementation. The result is time inconsistency in the data: The Pruebas SABER results are from 2013, while the data on program implementation was collected in 2016. On the one hand, this means that most of the students who took the exams in 2013 had already graduated by the time the implementation data was collected. On the other hand, the school model and the teaching methods could have potentially changed over this period. This concern is particularly important in the more remote rural schools, where teacher turnover seems to be high (according to anecdotal evidence from the field work).

Third, there are some issues of potential selection bias – a first set arising from missing school-level identifiers in the ICFES datasets, and a second set arising from the primary data collection. As discussed in detail in Annex A, a significant share of observations in the ICFES database cannot be allocated to a specific school (branch) because of irregularities in the reported exam scores.

These observations needed to be dropped from the study (because no EN identification is possible). The analysis in the annex suggests that the respective observations are not simply a random sample of all observations, but that information from school branches of educational institutions with certain characteristics (rural, public, lower socioeconomic level, lower average test scores, etc.) is more likely to be missing. This, in turn, means that the estimation coefficients are likely biased, because the unobserved factors responsible for the missing data ("administrative capacity") are most likely correlated with learning outcomes and explanatory variables such as the EN identifier. As discussed in detail in the annex as well as above in the context of possibility endogeneity, the direction of the bias depends on whether EN schools are more or less likely to be missing from the sample than conventional schools (i.e., whether administrative capacity and likelihood of being an EN are positively or negatively correlated). It seems likely that administrative capacity and EN implementation are positively correlated, leading to a positive bias in the estimated effect of the school model. In addition, selection bias in a multilevel setting means that fixed and random effects at different levels may be affected differently, which makes it even harder to understand the direction of a possible bias. The second source of selection bias arises from the data collection process. In some cases, data could not be collected in sampled schools, either because the school could not be located, or because the principal or the respondents did not consent to participate in the data collection. This may add an additional layer of bias if non-participation is not random. Schools with a higher implementation level may be more likely to participate in a research project about EN than conventional schools.

Fourth, apart from a potential sample selection bias, the estimates may also be affected by an omitted variable bias. The share of overall variance in test scores that can be explained by the model is small in all specifications. In particular, this is the case for between-student variance (as opposed to between-school, between-municipality, or between-department variance), which is responsible for most of the differences in test scores across students. If student characteristics that affect learning outcomes were equally distributed across schools, municipalities, and departments, this would not be a problem—but unfortunately, that is unlikely. As it is, omitting student characteristics from the model most likely produces biases.

Fifth, there is a range of issues with this study's approach to measuring EN implementation. The survey instrument that was used was developed by FEN. Using this instrument is justified by an attempt to base the research on an "official" definition of the EN methodology. However, while the survey instrument is intended to capture the degree of implementation and thus differences between EN and conventional schools, the resulting focus on EN methods may make the questionnaire less relevant to non-EN schools that use alternative, progressive teaching styles, which may go unnoticed. Thus, the resulting group of "conventional schools" may seem more homogenous that what it actually is, and thus a distinction between EN and conventional schools may be misleading.

Another issue related to the measurement of EN implementation is that the field work suggested that teaching practices are not necessarily stable over time in the same school, not even over relatively short periods of time. For one thing, reported teacher turnover is high, as most young teachers start out in the most distant schools but try to secure a teaching assignment in a village or city as soon as possible. At the same time, "student turnover" is high: It is common for students to change school several times over the course of their primary school career, moving around with parents who are seasonal farm workers. Furthermore, the school system itself may change often and rapidly. An anecdote from the quantitative data collection illustrates this point: One Thursday, a circular was sent out to all teachers (including the field work team of this project, who all are teachers) informing them that starting Friday, i.e. the next day, the school day would end two hours earlier because the budget for school lunch had run out. It is telling that none of the field workers were surprised or impressed. Against the background of a high level of fluctuations in schools, a spotlight measure of EN program implementation can thus only give a limited impression of the classroom experiences of students. A related concern is that there is no way to measure the time period over which students were exposed to EN methods, as the extent of implementation may change over time (or students may change schools). Clearly, the period of program exposure can be expected to have an influence on the effect size.

Sixth, there may be a bias arising from teachers (or, probably to a lesser extent, students) providing what they perceive as acceptable or desirable answers. For instance, EN teachers may know that the EN pedagogy stipulates that students work in pairs or groups for most of the time, but for some reason have their students work individually more often. They may still over-report group work time in order to comply with EN's guidelines, which would skew results. Two strategies were used to minimize this problem. First, teachers were assured that the results are anonymous and teachers are not evaluated individually based on their answers. Second, the results were triangulated through student surveys, as students may have fewer incentives to misreport classroom practices. Still, reports from the field workers indicate that the problem may persist. For instance, a field worker noted that certain EN elements were not posted on the wall of a classroom even though the respective teacher said they were.

There are also some limitations related to the qualitative part of the research. First, there are some limitations in the sampling process. Qualitative sampling methods typically call for iterative sampling where subsequent schools should be added to the sample until theoretical saturation is reached. This strategy is logistically demanding, though, as it requires a high degree of spatio-temporal flexibility on behalf of both the researcher and the research subjects. For these reasons, the sample size for the qualitative component of the project was determined ex-ante. The results of the interviews and observations suggested a high level of variation between schools, with new insights gained in each additional school that was visited. This means that qualitative saturation was not reached.

Second, the research design contained an element of participatory observations. Due to resource constraints, little time was spent at each school to conduct these observations (typically 2-3 hours). This time period may be too small to observe the *actual* every-day school life for various reasons. For instance, teachers and principals may want to present themselves and their schools from their best side; or children may be either excited about or intimidated by the visit of a stranger and thus not be “their true selves”. In fact, in at least one of the schools the students had never before received a visitor from outside of Colombia, which made the data collection resemble an event more than a regular school day.

7.2.2 Threats to external validity

In quantitative research designs, external validity refers to the extent to which findings of a study are generalizable to other situations, people, or institutions. In qualitative designs, the term transferability refers to a similar concept, namely, to the extent to which the findings can be transferred to similar situations, people, or institutions. For a research project this limited in size and budget, these goals are hard to achieve. In order to assure the highest possible level of *internal* validity, certain steps are taken to limit outside influences that may interfere with causal interpretations. Unfortunately, this implies reduced *external* validity and transferability. Specifically, there are three major limitations.

First, a small sample size is obviously related to a decreased degree of generalizability. This is of particular concern to the department-level analysis. While the sample accurately represents

Quindío's rural primary schools, it is unclear how the results can be transferred to other Colombian departments, let alone to other countries. The country-level study suggested large differences between departments with regard to influences on learning outcomes in general, and with regard to the effect of the EN model in particular. Thus, the conclusions from the department-level study may not be fully applicable to other departments.

Second, in an attempt to improve the practical feasibility of the project, the sample for the department-level study was limited to defined sub-sets of the primary school population: to specific departments, to rural schools, to a region with a long history of EN-implementation, to an area with comparatively good access and low levels of violence, etc. Generalizability to other schools outside of these subgroups is necessarily limited, as many intervening factors (such as cultural, political, or economic differences) may completely change the program's effect.

Third, there are some shortcomings related to construct validity. Construct validity refers to the extent to which a test or instrument measures what it claims to be measuring. As the research design relies on outcome data from the *Pruebas SABER* database, schooling outcomes are measured only by scores in a set of standardized tests. This approach to capturing quality of education is superior to simple attendance and input measures, but it does not necessarily gauge the acquisition of skills and knowledge that are relevant to a student's life, or any other long-term benefits to the student. In the end, what is truly of interest is not whether students in EN have higher test scores than their peers in conventional schools; it is whether EN is better in preparing students for the challenges they will face in life. Comparing how EN graduates do later in their lives (relative to graduates from conventional schools) may thus be an interesting alternative way to evaluate the school model, yet collecting appropriate data was not part of this research project. In fact, FEN needed to dismiss the plan for such a study because of the impossibility to track down an adequate sample of graduates. Standardized test scores are unfortunately the best widely available quality-of-school measure at the moment. Generalizing from EN's impact on average test scores to the model's impact on learning outcomes in general may be problematic.

7.3 Future research

Although this project answers important questions, some issues remain unresolved, and others have emerged over the course of the project.

A first area of future research is a more in-depth investigation of teaching practices in EN (and conventional) schools. In order to better understand how the model is being used in practice and where adaptations might be necessary, a more detailed study on the use of the individual instruments is necessary. In particular, it is important to understand how teachers use the model's elements, and why they choose to use, not use, or adapt certain features. This understanding will help improve the efficiency of the model. In this context, a larger, more focused study on the effects of the different parts of the model would help to obtain a better understanding of which of the model's elements to focus on and of the room for adaptations.

Relatedly, the role of teacher training for the accurate implementation of the model needs to be investigated more closely. While it seems logical that teacher training and support are the basis for implementing the other parts of the model, the results of the department-level analysis of test scores did not point to an outstanding effect of this dimension of the implementation index. Nevertheless, the qualitative analysis indicated that teachers may feel left alone with the model and its challenges, and thus not be able to benefit from the methodology for their work as much as possible.

A second area of future research is the role and use of technology in EN classrooms. The qualitative data collection revealed that most schools are equipped with tablet computers. They seem to be used in a variety of ways: as research tools, as a substitute for missing printed learning guides (by simply working with scanned digital versions of the guides), to play educational games, and more. EN, as a student-focused school model, may lend itself even more than conventional instruction to the effective use of digital learning content. With the large variety of free or low cost educational content available on the Internet, integrating technology into a multigrade EN classroom where every student works at their own rhythm seems like a logical step. Research is necessary in order to understand how technology is currently being used in schools, and in order to assess the potential of a "digital Escuela Nueva".

There are not only open questions about implementation at the school-level, but also at the municipality- and department-levels. The multilevel analysis has clearly shown that there are differences across the country with regard to the effect that the EN model has on learning outcomes. More specifically, there seems to be some correlation between political support and a stronger effect on learning outcomes. However, there are departments with low levels of political support and above-average effects of the model, and vice versa. A better understanding of what drives these differences can help improve the program's efficacy.

One of the largest limitations of the dataset used for this study is the lack of any student-level explanatory variables besides gender. Still, the analysis of variance showed that differences between students are responsible for at least half—in most cases even more—of the total variance in exam scores. There is some research about student-level factors that influence learning outcomes in Colombia, but little is known about the student-level factors in the context of the EN model. This is a particular shortcoming for a model that promotes student-centered and personalized learning.

Outcomes of the EN model beyond test scores (or school behaviors) are another topic where much more research is necessary. What happens with EN graduates later in life, compared to students of conventional schools? Do the skills that they (supposedly) learn at EN, such as self-responsibility, learning to learn, comprehension and problem solving skills, etc., help at higher education levels and later in life? Are EN students more likely to graduate high school, or go to university? Are they less likely to live in poverty, and more likely to have stable employment? And are they more likely to be active and informed citizens? These questions pose many challenges for a research design, yet the answers to them are crucial to evaluate the true benefits of the EN model.

Finally, this study did not address the issue of costs. The model was found to be more effective than conventional schools in improving learning outcomes, but apart from a rough estimate from almost three decades ago and an indirect study of efficiency through modeling production possibility frontiers from two decades ago, there is no information on the efficiency of the model, i.e., on how costs compare between the school types. Clearly, this type of information is crucial for policy decisions.

7.4 Policy recommendations

The results show that the EN model helps to improve learning outcomes, especially for children from disadvantaged backgrounds. The main policy recommendation is thus clear: the proper use of the model should be promoted and politically supported. In more specific terms, that means:

- Clearly spell out the key elements of the model and their roles and purpose within the overall EN approach. The field work revealed different interpretations of what EN entails. While the model needs to maintain a certain flexibility to be able to address specific local conditions, a clear definition of the EN package is crucial for effective program scale-up. Focusing on specifying the purpose of each EN element and its relationship to other parts of the model (as opposed to focusing on the operational details of each element) can help to replicate the model up in a more effective manner while maintaining adaptability.

- Provide funding for the necessary learning resources. The department-level case study showed a lack of sufficient learning guides for each student and subject area, or a lack of up-to-date learning materials. This is despite the fact that learning guides are continuously updated by Fundación Escuela Nueva in cooperation with different departments.
- Provide funding for teacher training workshops. EN pre-service and in-service workshops are designed to complement formal teacher training and to help teachers understand and appreciate the EN methodology in a hands-on, learning-through-experience setting. There are only a handful of workshops that teachers should complete, each a week long or less, which is not much time compared to the resources already spent on formal teacher training. It would be an investment that pays off if it helps to better implement the EN model.
- Provide support and resources for in-service teacher support systems. Ongoing support for teachers is a key feature of the EN model. At least in the department under study, monthly teacher gatherings (micro centers) and mentoring visits to school are not always taking place. While these are not highly resource-intense features, support from the Secretaries of Education and/or municipalities for the organization and logistics of these features is necessary.
- Create systems to monitor program implementation. The department-level analysis confirms preliminary findings from other studies indicating that the model is not always being properly implemented, which limits its potential success. This is not a unique challenge of EN, or of the Colombian education sector. Rather, implementation failures

are a common phenomenon in many policy areas and projects. It is important to design a monitoring system that helps to keep track of what is happening in the schools, and where schools or teachers may need additional support.

In addition to these EN-specific issues, there are three recommendations to improve Colombian information systems and provide a more effective basis for evidence-based policy making:

- Provide basic information on student characteristics. Given that learning outcomes vary more between students than between schools or other higher levels, information on student characteristics is crucial for the analysis of learning outcomes—not just in order to better understand these student-level characteristics, but also to be able to correctly control for them when estimating the effect of higher-level factors. While there is an obvious concern about the privacy of the test takers, providing anonymized data on key student characteristics significantly increases the quality and range of types of analysis than can be done with the standardized test scores.
- Create a clearing house for education data. Colombia lacks a centralized education statistics system. Different types of education-related statistics are collected and managed by different institutions, and there is no institution that combines, or even catalogues, the different data bases. Since the data is generally available but not effectively integrated, this constitutes an area of large potential efficiency gains.
- Provide municipality-level data on key indicators. Most indicators published by the Statistics Bureau DANE are available only aggregated at the country- or at best department-level, or sometimes as micro data. Providing municipality-level data would facilitate policy research and analysis.

ANNEX A: Description of the Data

1 The Pruebas SABER dataset

1.1 Description of the data

The main dataset for this study is the dataset containing the results of the 2013 round of the Pruebas SABER. This section describes the Pruebas SABER as well as the dataset.

The Pruebas SABER 3º, 5º Y 9º (henceforward, Pruebas SABER) is a standardized evaluation of learning outcomes that has been carried out in Colombia since 1991. Until 1999, the evaluation was based on a representative sample of students in grades 3, 5, 7, and 9 who were tested in mathematics and language. In 2001, Ley 715 de 2001 established that participation in the tests is obligatory, and that the test was to be carried out every three years. Based on that law, starting 2001 all 5th and 9th grade students in both private and public schools participated in the Pruebas SABER. Since in 2012, 3rd grade students also participate in the evaluation, and tests are carried out on a yearly basis (ICFES 2015b).

Starting in 2012, every evaluation round contains a mathematics and a language assessment for grades 3, 5, and 9. Additionally, alternating each year there is also an evaluation in natural sciences or civic competencies for students in grades 5 and 9. Each student in grade 3 receives a test either in mathematics or in language, the assignment being random. By contrast, each student in grade 5 and 9 receives tests for two of the three test areas (also randomly assigned). The testing time in grade 3 is 2 hours and 50 minutes, the testing time in grades 5 and 9 is 4 hours and 35 minutes. The Pruebas SABER are complemented by a socioeconomic survey in two versions, one for 3rd grade students and one for 5th and 9th grade students (ICFES 2016b, 6–7). This study only uses data for primary grades levels, that is, for grades 3 and 5. Test results are published as plausible values (see section 3.2.1.8.2).

According to ICFES (2016b), 772,394 3rd grade students and 759,150 5th grade students (1,531,544 in total) participated in the 2013 round of the Pruebas SABER. These were distributed across 18,255 educational institutions and 52,004 branches. However, results are only available for 704,697 3rd graders and for 706,204 5th graders (1,410,901 in total), as well as for 17,073 educational institutions and 31,050 uniquely identifiable branches. When only looking at students in uniquely identifiable branches, data is only available for 567,939 3rd graders and for 574,948 5th graders (1,142,887 in total).

The considerable discrepancy in the numbers is due to reporting errors. For some educational institutions with different branches and/or sessions, identification of results at the school branch- or session-level is not possible "either because the educational institution has not submitted the test materials at session-branch-level, or because inconsistencies in the information about student enrollment in each session or branch were detected".[24] In these cases, unique session- or branch identifiers are not available, and test results are only reported at the highest level (the educational institution).

[24] Translation by the author. Original text: "[Estas] sedes-jornadas no tienen reportes debido a que el establecimiento educativo no entregó el material identificado por sede-jornada, o porque se detectaron inconsistencias en la información de estudiantes matriculados en cada sede-jornada." (ICFES 2016a)

Table 60: Loss of data in the Pruebas SABER dataset due to reporting errors

	Participated in exam (according to ICFES (2016b))	Results available (in database)	Branch-level identifiable results available (= sample)	Sample size reduction due to missing branch-level identifiers
Students (grade 3)	772,394	704,697	567,939	19.41%
Students (grade 5)	759,150	706,204	574,948	18.59%
Sessions	n/a	32,942*	32,942	n/a*
Branches	52,004	31,050*	31,050	n/a*
Institutions	18,255	17,073	14,729	13.7%

* It is not possible to determine the number of branches and sessions that the unidentifiable results belong to.

Because of these problems, the respective observations have to be removed from the dataset. The sample is reduced as follows: 13.7% of educational institutions (2,344 out of 17,073) have to be dropped, corresponding to 268,014 of the 1,142,887 student-level observations (19%). While this represents a considerable share of the sample, the reduction is necessary because identification of the school model (EN or not) happens at the branch-level, and is thus not possible for institutions where data is not disaggregated by branches. The loss of data is summarized in Table 60.

1.2 Assessment of a possible sample selection bias

Dropping so many observations does not matter for unbiasedness of results[25] if the branch- and session identifiers are missing randomly, that is, if the institutions with missing data do not differ in observable or unobservable ways from the institutions without missing data. For observable factors, this can easily be tested for variables available in the dataset. As it turns out, there are

[25] It does, of course, lead to a smaller sample size and thus to larger standard errors. Hence, while estimations may remain unbiased, they become less efficient and may fail to achieve statistical significance even if the research hypothesis is true.

considerable and significant differences between educational institutions with missing lower-level identifiers and those with complete information. Institutions are more likely to have missing branch-level identifiers if the they are public and of a lower socioeconomic level (these and the following differences are significant at the 0.1% level (all p-values <0.001)). For each component of the SABER exam (mathematics, language, and civic competencies), institutions with missing branch-level identifiers are more likely to have cases of reported cheating, and have significantly lower average test scores (between a third and two fifth of a standard deviation lower). Finally, these institutions are typically larger; this, however, could also be because larger institutions are more likely to have more branches and sessions, and thus a larger potential for errors in any sub-unit. Institutions with missing branch-level identifiers are no more or less likely to be rural than other institutions.

One very important characteristic is still missing from this analysis: whether EN schools are more likely to be missing from the sample. Unfortunately, it is impossible to determine that conclusively given that the branch-level identifiers are missing, but it is possible to check whether institutions that have missing branch-level identifiers are more likely to also have a branch that offers EN education. DANE, in its EDUC C-600 database (described below in section 2), provides information on EN implementation both at the branch-level and at the institution-level. Adding this information to the Pruebas SABER dataset allows one to check how the school model and the lack of branch-level identifiers are correlated. Fortunately, it turns out that at the institution-level there is no significant correlation between having a branch with a missing identifier and having a branch that offers the EN school model (in both groups, around 40% of institutions have branches that offer the EN model). However, this is only an approximation, as it is not possible to determine whether EN *branches* are more likely to be missing.

In any case, given the important observable differences described above, it is clear that sub-unit identifiers are not just randomly missing, which will lead to biases in the estimation results if the factors correlated with missing sub-unit identifiers are also correlated with the outcome and explanatory variables. This is clearly the case for observable factors: For instance, students in private schools have average scores of around 1.25 standard deviations above students in public schools. Excluding a disproportionate number of public schools will lead to an upward bias in the estimation of the overall skill level.

More importantly, it seems reasonable to assume that failure to report the test results correctly is strongly correlated with a latent variable (which is part of the error term), which may be called “administrative capacity”. Administrative capacity could include the resource endowment of the school and the motivation and skill level of teachers and administrators, among other things. The direction of a potential bias arising from omitting this latent variable from the analysis depends on how one thinks that “administrate capacity” and “likelihood to be an EN” are correlated.

On the one hand, it is possible that the omitted latent variable “administrative capacity” is negatively correlated with the probability of a school to be an EN school. This would be the case if more poorly funded, more remote, more understaffed schools are more likely to adapt the EN school model – or, maybe, to be assigned to use that model. In this case, if the hypothesis holds that EN schools have better learning outcomes, estimations for the effect of the EN model based on the dataset that includes only institutions without missing sub-unit identifiers (i.e., with higher administrative capacity) will have a negative bias. In other words, the estimated effect of EN based on that sample is probably weaker than the “true effect”, because the EN effect is confounded with the effect of low capacity.

On the other hand, however, this negative bias depends on the assertion that EN schools are more likely to be low capacity-schools, where capacity is also negatively correlated with learning outcomes. If instead it was the case that it requires high administrative capacity to become an EN—for instance, because it takes above-average teacher engagement and resources to start the model—the picture changes. If the correlation between being an EN and administrative capacity is positive, EN are more likely to be in the sample, and the bias in the estimator for the effect of the EN model becomes positive. In that scenario, the “true” effect of EN would be weaker than suggested by the estimations.

Finally, it is possible that there is indeed no correlation between the likelihood to be in the sample and the likelihood to be an EN. This would be the ideal case, as the estimates would be unbiased (with regard to this specific factor). However, this scenario seems unlikely.

While the data shows that institutions that have at least one branch where EN is implemented are no more likely to have missing branch-level identifiers, this does not necessarily mean that there a no correlation between *branch* capacity and EN implementation. The main problem here of course is that it is not possible to establish with certainty whether or not EN schools are more likely to be missing due to reporting errors caused by the respective branch. It is only possible to determine whether educational institutions that have at least one EN branch are more likely to have branch-level reporting errors, which may or may not be due to errors in the actual EN branches. Furthermore, it is not possible to actually test whether EN and non-EN schools truly differ in the unmeasured variable “administrative capacity”. The fact that data was misreported can only be seen as a proxy.

The available evidence is not conclusive about the direction of the correlation between EN and “administrative capacity”. Based on the literature and the evidence gathered for this research, the second scenario seems slightly more likely: While one might think that “low capacity” schools may be more likely to be officially labelled as EN (which would result in a negative bias in the estimated effect of EN classification), the evidence gathered for this research suggests that better or higher-motivated teachers are more likely to actually adopt more elements of the school model, resulting in a positive correlation between EN implementation and likelihood of being part of the sample. Thus, the estimated effect of EN may have a positive bias, i.e., the true effect may be smaller than suggested.

Assessing a potential bias on other estimators adds to the complexity. Of special interest is to know whether the estimated effects of socioeconomic status and gender and their interactions with EN may be biased, and if so, in what direction. The case of gender is relatively straightforward: There is little *ex ante* reason to believe that a students’ gender is correlated with unobserved school quality (administrative capacity); thus, no bias is expected on the main effect, and the bias of the interaction term follows the bias of the main effect of EN implementation.

For socioeconomic status, the situation is more complicated. It is likely that average socioeconomic level and unobserved administrative capacity/school quality are positively correlated, introducing a positive bias in the estimate of the effect of socioeconomic status. The direction of the bias of the interaction term with EN implementation again depends on whether the correlation between school quality and likelihood of using the EN model is positive or negative. Again assuming a positive correlation, the effect of the interaction term is likely to be overestimated (i.e., the equalizing effect of the EN model may be smaller than the estimation results suggest).

This discussion shows that understanding a potential bias is complicated, as any intuition on its effect will depend on assumptions. In addition to these challenges, however, in a multilevel model the selection bias may also affect fixed and random coefficients at the different levels, maybe in different ways. Understanding the effect of selection bias in multilevel settings is thus even more complicated. Grilli and Rampichini (2012) discuss the problem for a simple two-level model, showing that sample selection affects not only estimation coefficients (fixed and random), but also the covariance structure. Unfortunately, theoretic research on understanding the direction and adjusting for sample selection bias in multilevel models is not very advanced. This project can only acknowledge the existence of this bias as a shortcoming.

1.3 The Pruebas SABER dataset

By removing observations where the branch-level identifier is missing, the sample is reduced to 1,142,887 students in 31,050 branches and 14,729 educational institutions. This is henceforward referred to as "Pruebas SABER dataset" or "sample". Table 61 summarizes how many observations are available for the different institutional levels and grades, as well as the percentage of observations in rural areas. Only about half of the education institutions are in rural areas, while around two thirds of school branches are. Not surprisingly, these rural schools tend to be smaller than urban ones: only around a quarter of student-level observations comes from rural areas. Finally, Table 62 summarizes the number of observations available in the dataset.

Table 61: Overview of schools in Pruebas SABER 2013 dataset

Based on SABER 2013	Total	In grade 3 / results for grade 3		In grade 5 / results for grade 5		Located in rural zone	
		N	% of total	N	% of total	N	% of total
Institutions	14,729	14,209	96.47%	13,573	92.15%	7,034	47.76%
Branches	31,050	28,658	92.30%	27,047	87.11%	21,229	68,37%
Sessions	32,942	29,989	91.04%	28,182	85.55%	21,365	64.86%
Students	1,142,887	567,939	49.69%	574,948	50.31%	277,397	24.27%

Table 62 Overview of Pruebas SABER 2013 dataset: missing observations

Variable	Missing observations	Non-missing observations	Unique values
ID Student	0	1,142,887	1,142,887
ID Institution	0	1,142,887	14,729
ID Branch	0	1,142,887	31,050
ID Session	0	1,142,887	32,942
Code Department	0	1,142,887	33
Code Municipality	0	1,142,887	1,084
Area	0	1,142,887	2
Sector	0	1,142,887	2
Session type	0	1,142,887	3
Socioeconomic level of school	51,463	1,091,424	4
School calendar	0	1,142,887	2
Grade	0	1,142,887	2
Sex	0	1,142,887	2
Plausible values Language	476,935	665,952	
Plausible values Mathematics	482,003	660,884	
Plausible values Civic competencies	762,005	380,882	

2 The administrative dataset EDUC (C-600)

Official administrative data, including data on the educational model used in each school and school branch, is provided by DANE in the EDUC dataset. All Colombian public and private school are required by law to respond to the "C-600 survey" every year, which is why the dataset covers most schools. Educational institutions and school branches are identified by the same unique codes that are also used in the Pruebas SABER dataset (and for other education-related purposes). This study uses data from 2013, which is the year of the Pruebas SABER results.

The following variables come from the C-600 dataset:

- School identifiers at institutional, branch, and session level
- Geographic information: department, municipality, urban or rural area
- Sector: private or public

- Educational offers: education level and school model
- Student population: Number of students in current and previous year, presence of students from ethnic minorities or students who are victims of the armed conflict

Table 63 gives an overview of the C-600 data used for this study. Note that this data contains information on all Colombian schools, not only those offering primary education, hence the share of missing observations is different in the final sample used. Key variables are available for a large share of the dataset. The final sample is described at the end of this section.

Table 63 Overview of C-600 dataset: missing observations

Variable	Missing	Non-missing	Unique values	Min	Max
ID Institution	4	69,975	>500	-	-
ID Branch	0	69,979	>500	-	-
Code Session	6	69,973	5	1	5
ID Department	36	69,943	33	5	99
ID Municipality	36	69,943	>500	1	980
Area	36	69,943	2	1	2
Sector	36	69,943	2	1	2
Primary in institution	42	69,937	2	0	1
Primary in branch	36	69,943	2	0	1
Model: Traditional	6,687	63,292	2	0	1
Model: Escuela Nueva	6,687	63,292	2	0	1
Students of ethnic minorities	6,687	63,292	2	0	1
Students who are conflict victims	6,687	63,292	2	0	1
Students primary, previous year	21,120	48,859	>500	1	2307
Students primary, current year	18,820	51,159	>500	1	2833

According to the EDUC dataset, there are 25,133 educational institutions in Colombia, 21,351 (84.95%) of which offer education at the primary level. There are furthermore 57,354 school branches, 49,932 (87.06%) of which offer primary education. Finally, there are 69,979 sessions, out of which 51,159 (73.11%) reported having primary level students in 2013 (see Table 64).

11,358 institutions (45.19% of all institutions) are located in rural areas, as are 37,633 branches (65.62% of the total number) and 40,589 sessions (58.03% of the total). The focus of this study are schools located in rural areas and offering primary education, which are 10,997 institutions (43.76% of the total), 36,360 branches (63.40% of the total) and 34,996 sessions (50.04% of the total). Not surprisingly, the percentage of schools that offer primary education is higher in rural areas than for the country average, for all three institutional levels.

Table 64 Overview of Colombian schools. Source: DANE, EDUC 2013 dataset

Based on EDUC	**Total**	**Offering primary education**		**Located in rural zone**		**Offering primary education and located in rural zone**		
		N	**% of total**	**N**	**% of total**	**N**	**% of rural**	**% of total**
Institutions	25,133	21,351	84.95%	11,358[1]	45.19%	10,997[1]	96.82%	43.76%
Branches	57,354	49,932	87.06%	37,633	65.62%	36,360	96.63%	63.40%
Sessions	69,980	51,159	73.11%	40,589	58.03%	34,996	86.22%	50.04%

[1] Includes 956 institutions (3.80%) that are classified as "urban and rural"

3 Other data

Some additional variables were added from other data sources in order to complement the analysis. Unfortunately, the official statistics system in Colombia is relatively scattered, and municipality-level statistics are rarely available. Data had thus to be chosen based on availability, even if different data sources mean that the data may not always be perfectly comparable (due to different base years or data collection methods, for instance).

Homicides: Based on data from the Instituto National de Medicina Legal y Ciencias Forenceses (MED. LEGAL), the Federación Colombiana de Municipios (2016) provides municipality-level data on homicide rates (homicides per 100,000 inhabitants). The most recent data available is from the year 2009. The dataset contains information for 734 of Colombia's 1103 municipalities. The mean homicide rate is 40.7, the range goes from 1 to 417, and the standard deviation is 42.7.

Gross Domestic Product by Department: The National Statistics Bureau (DANE 2016) provides data on per capita GDP by department (but unfortunately not by municipality). The data used are the definite statistics for 2013 (the year of the Pruebas SABER results). In order to facilitate interpretation, the unit of the data is changed (1 million pesos), and the data is rescaled (centered at the mean of all departments). The centering changes the range from between 4.41 and 44.90 to between -8.66 and 31.83, and the mean from 13.1 to close to 0. The standard deviation is 9.62.

Public education expenditure: Public expenditure on education is a control variable of high theoretic merit, but data below the country-level is not publicly available. The only available data that could be identified are from as far back as 2004 and were presented in a paper by Iregui, Melo, and Ramos (2006). There are department-level estimates for the 32 departments and Bogota, and municipality-level estimates for 6 departments (including Quindío) and four major cities. The per-student expenditure (in 1000 pesos) ranges from 160.5 to 1,988.0, with a mean of 623.6 and a standard deviation of 217.6. In order to facilitate interpretation, the data is then rescaled (centered at mean of all municipalities), so that the centered variable ranges is from -463.1 to 1,364.3, with a mean close to 0 (the standard deviation remains of course unchanged).

Municipal governance: The National Planning Department has developed an index to evaluate the overall performance of municipalities, according to the Leyes 152 de 1994, 617 de 2000 and 715 de 2001 (DNP 2014). The index for the year 2013 is used as a proxy variable for municipal governance. It consists of four equally weighted components: Efficacy (progress towards development goals, compliance with provision goals); efficiency (comparison between expenditures and outcomes in the provision of education, health, and drinking water, definition of potentials, productivity analysis); legal aspects (compliance with budgeting and reporting requirements for different sectors); and administration (administrative capacity and fiscal performance). Data is available for 1,101 municipalities. The index value ranges from 9.7 to 92.4, with mean of 68 and a standard deviation of 13.6. In order to facilitate interpretation, the data is rescaled (centered at mean of all municipalities), so that the new range is -58.4 to 24.4, and the mean is 3.

4 The merging process

The final dataset is a combination of all the datasets described above. Unfortunately, as is common when combining datasets like the ones used in this study, the datasets do not match up perfectly. Even though both the Pruebas SABER dataset and the DANE EDUC C-600 dataset use the same unique identifiers for schools, a number of educational institutions and branches are only contained in one of the two datasets.

The two data sets were merged using branch-level information from the DANE EDUC C-600 dataset. In other words, before merging the two datasets, all session-level information from the

C-600 dataset was aggregated onto the branch-level.[26] This means that some level of precision is lost: information such as the number of students is now only available as an average of the school branch, not for each session (i.e., morning, afternoon, evening, or other session). However, using branch-level information increases the success rate in the matching process dramatically: session-level data could only be matched between the two datasets for 69.7% of student-level observations, while branch-level data could be matched for 96.3% of student-level observations. Given that the main variable of interest, the school model, is defined at the branch-level, the loss of precision resulting from using branch-level C-600 information is justified by the large expansion of the available data.

Table 65 summarizes how the two datasets overlap. Four points should be noted to make better sense of the numbers. First, the Pruebas SABER datasets only refers to the part of the dataset where the branch was identifiable (see page 251). Second, the same institutions may have branches that could be matched and branches that could not be matched. Third, the number of "DANE EDUC C-600-only" branches can be partly explained by the large number of schools for which Pruebas SABER results could not be assigned to a specific branch (see page 251). The failure to match these branches is thus not due to a mismatch in the branch identifiers per se, but to the fact that ICFES purposefully omitted branch-level identifiers where there were irregularities in the reported results. In these cases, the corresponding branches of the DANE EDUC C-600 database remain thus without a match. A sample of "DANE EDUC C-600-only" branches was checked via

[26] Two different methods were used to calculate the branch-level data, depending on the data type. First, where appropriate, the relevant numbers were added up. For instance, the number of students in each session of the same branch was added up to the total number of students in that branch. Second, where appropriate, the session-level information was transferred onto the branch-level. For instance, if one session within a branch reported having students who are conflict victims, per definition the entire branch was marked as having students who are conflict victims.

ICFES's online query form (ICFES 2016a), and in fact in many cases the results page indicated that branch-level results are not available due to reporting errors.

Forth, the DANE EDUC C-600 dataset includes data for all school branches, including those that—according to this dataset—don't offer primary education. The decision to include all branches in the merging process was taken because there are a few cases where Pruebas SABER results from grade 3 or 5 are available for branches that, according to the DANE EDUC dataset, don't offer primary education. Around 5,700 (or 21%) of the branches that appear only in the DANE EDUC C-600 dataset report not offering primary education. Of the branches that could be matched to the primary test results from the Pruebas SABER dataset, 89 (or 0.3%) report not offering primary education.

Table 65 Overlap of the DANE EDUC C-600 dataset and the ICFES Pruebas SABER dataset

	Institution-level: (Inst. with *at least one* branch in the following categories)	**Branch-level**	**Student-level**
Pruebas SABER only	798	951	42,490
DANE EDUC only	13,667	27,245	--
Matched	14,371	30,099	1,100,397
Total	25,632*	58,295	1,142,887

* Numbers do does not add up to this total because institutions may have branches in different categories.

There are some variables in the dataset—specifically, the area where the school is located and the sector of the school (public versus private)—where the information provided by ICFES and DANE is conflicting. The case of the rural-urban classification is presented in Table 66. Out of the

30,099 branches in the sample, ICFES classifies 20,732 as rural while DANE classifies 20,985 as rural. The overlap between the two classification—the number of branches that are classified as rural in both datasets—is 20,614 branches. That means that there are 118 branches that are classified as rural by ICFES and as urban by DANE, as well as 371 branches that are classified as rural by DANE and as urban by ICFES. The number of apparent misclassifications is smaller for the public-private distinction, but as Table 67 shows there are still 39 cases in which the two datasets do not agree in their classification.

Based on the available information, it is impossible to know which of the datasets is "right". Given that the main data source for this analysis is ICFES, and DANE's database is used only to provide background information, ICFES' classification is used for the study. However, the robustness of the results will be checked by re-running the estimates based on DANE's classification.

Table 66: Comparison of urban-rural classification between ICFES Pruebas SABER dataset and DANE EDUC C-600 dataset

		DANE classification		
		Urban	**Rural**	**Total**
ICFES classification	**Urban**	8,996	371	9,367
	Rural	118	20,614	20,732
	Total	9,114	20,985	30,099

Table 67 Comparison of private-public classification between ICFES Pruebas SABER dataset and DANE EDUC C-600 dataset

		DANE classification		
		Public	**Private**	**Total**
ICFES classification	**Public**	25,309	7	25,316
	Private	32	4,751	4,783
	Total	25,341	4,758	30,099

Adding the supplementary datasets (homicides, expenditures on education, governance, and per capita GDP) does not cause any issues; the municipality and department identifiers match up almost perfectly between the datasets. Whenever data for a municipality or department was missing in the original dataset, the data for the respective municipality or department were still kept in the study dataset, with a missing entry for the supplementary variable. This is the case for a large number of observations for homicide rates.

The merging of the primary data on program implementation was expected to be straightforward, as the secondary dataset was used as the sampling frame. However, for two of the schools in the sample, no data from the C-600 database was available. Data from the other 76 schools in 10 municipalities could be successfully added to the study dataset.

ANNEX B: Methodological Annex

1 Development of the country-level random-intercept model

The goal of this annex is the development of a random-intercept multilevel model. The four-level null model developed in section 4.2 is the starting point; control variables will be added step by step according to the level they belong to. This level-by-level procedure helps to create a more stable model, and it also helps to better understand which part of the respective level's remaining variance the predictors can explain. The main variable of interest (the dummy for EN) is included from the beginning. The first model to be tested is the RI1 model. It is defined as follows:

Model RI1: $score_{ijmd} = \beta_0 + \beta_1 EN_{jmd} + \beta_2 male_{ijmd} + \beta_3 (male * EN)_{ijmd} + \xi_{ijmd}$

, where ξ_{ijmd} is the composed error term consisting, as discussed in the previous section, of the student-level error term ε_{ijmd}, the school-level error term ζ_{jmd}, the municipality-level error term ζ_{md}, and the department-level error term ζ_{d}. Subscripts are added to all regressors to indicate the level on which the variables change. Apart from the EN dummy, the model includes a predictor for gender ($male$) and a cross-level interaction term of EN and gender ($male * EN$). This interaction term is testing the hypothesis that the effect of the EN model differs by gender.

The estimation results for this model are presented in the second column of Table 68 through Table 72 for the different testing areas and grades. In this simple model, the effect of EN is significant only in some of the models (not for language grade 3 and mathematics grade 5). The effect of the other explanatory variables will be discussed further below, once the final model is developed. What is interesting in this context is above all the result of the likelihood test presented in the last row; it shows, for all grades and testing areas, that the model fit has improved compared to the null model. This despite the fact that the errors presented in the lower part of the tables have barely changed, which indicates that gender only explains a small part of the between-student differences. The student-level share in the unexplained variance remains the largest of all error terms: between 72% (in the case of grade 5 civics) and 51% (in the case of grade 3 mathematics) of the total unexplained variance is due to student-level factors, the share being larger for grade five than for grade three in all areas. Unfortunately, the dataset does not provide any more student-level variables that could be used to analyze this variance.

The next step is the inclusion of school-level explanatory variables, as shown in Model RI2:

Model RI2: $$score_{ijmd} = \beta_0 + \beta_1 EN_{jmd} + \beta_2 male_{ijmd} + \beta_3 (male * EN)_{ijmd} + \beta_4 rural_{jmd} + \beta_5 private_{jmd} + \beta_6 NSE_{jmd} + \beta_7 (NSE * EN)_{jmd} + \beta_8 ethnic_{jmd} + \beta_9 conflict_{jmd} + \beta_{10} morning_{jmd} + \beta_{11} afternoon_{jmd} + \xi_{ijmd}$$

Added to RI1 are the variables $rural$ (a dummy for whether the school is in a rural area); $private$ (a dummy for whether the school is private); NSE (the socioeconomic level of the school, defined as the average official socioeconomic level of the children; level 1 is coded as zero, levels 2, 3, and 4 are coded as 1, 2, and 3); the interaction of NSE and EN, which tests the hypothesis that the EN model is particularly beneficial for children from disadvantaged backgrounds; $ethnic$ (a dummy for whether there are students of ethnic background in the school); $conflict$ (a dummy for whether there are children in the school who are victims of the conflict; and $morning$ and $afternoon$ to indicate the type of the session (a full school day being the base category).

The results are presented in the third columns of Table 68 through Table 72. The effect of EN is now strongly positive and clearly significant for all grades and testing areas (for a discussion of the other variables, see below). The likelihood tests reveal that the model fit has improved, and the new model is preferable over model RI1. Additionally, the unexplained school-level variance could be decreased, as becomes clear when comparing the size of the standard deviation of the school-level random part between models RI1 and RI2. Interestingly, the extent of the department-level unexplained variance could also be reduced by including school-level regressors, which indicates that school-level regressors differ considerably between departments.

Model RI3 includes the municipality-level variables, and is defined as:

Model RI3: $score_{ijmd} = \beta_0 + \beta_1 EN_{jmd} + \beta_2 male_{ijmd} + \beta_3 (male * EN)_{ijmd} + \beta_4 rural_{jmd} +$
$\beta_5 private_{jmd} + \beta_6 NSE_{jmd} + \beta_7 (NSE * EN)_{jmd} + \beta_8 ethnic_{jmd} +$
$\beta_9 conflict_{jmd} + \beta_{10} morning_{jmd} + \beta_{11} afternoon_{jmd} + \beta_{12} governance_{md} +$
ξ_{ijmd}

The only variable that was added is $governance$, the governance index of the municipality (see section 3 for an explanation of the index). The effect of EN remains strongly and significantly positive, and the likelihood-ratio tests reveal that the RI3 models are superior to the RI2 specifications for all testing areas and grades. While inclusion of this new variable reduces the difference in the error terms across municipalities only slightly, larger changes are, again, apparent for the department-level error terms, which again suggests differences in this variable across departments.

Finally, department-level indicators are introduced into the model. Model RI4 is defined as:

Model RI4: $score_{ijmd} = \beta_0 + \beta_1 EN_{jmd} + \beta_2 male_{ijmd} + \beta_3 (male * EN)_{ijmd} + \beta_4 rural_{jmd} +$
$\beta_5 private_{jmd} + \beta_6 NSE_{jmd} + \beta_7 (NSE * EN)_{jmd} + \beta_8 ethnic_{jmd} +$

$$\beta_9 conflict_{jmd} + \beta_{10} morning_{jmd} + \beta_{11} afternoon_{jmd} + \beta_{12} governance_{md} + \beta_{13} pcGDP_d + \beta_{12} educ\_expenditure_d + \xi_{ijmd}$$

Two new variables were added: $pcGDP$, which is the per capita departmental GDP (in million pesos, centered at the average across departments); and $duc\_expenditure$, which is the public education expenditure per student in 2004 (in 1000 pesos, centered at the average across departments). The effect of EN remains unchanged, yet the LR-test statistics turn out insignificant for all grades and testing areas, which indicates that Model RI4 does not improve the data fit, compared with model RI3. This is insofar not surprising as the effect of the two department-level explanatory variables is insignificant in all cases (except for education expenditure, which is borderline significant in the case of grade 5 civic competencies). Given that overfitting should be particularly avoided in multilevel models (Snijders and Bosker 2011), the department-level explanatory variables are discarded from the analysis[27], and model RI3 is chosen as the best-fitting model.

[27] Because of the theoretical importance of education expenditures, the variable will be re-introduced for the robustness analysis.

Table 68 Results of the random-intercept models, language grade 3 (country-level study)

Language Grade 3	**Model 0**			**Model RI1**			**Model RI2**			**Model RI3**			**Model RI4**		
n (students)	197,234			197,234			197,234			197,234			197,234		
j (schools)	17,652			17,652			17,652			17,652			17,652		
m (municipalities)	1,007			1,007			1,007			1,007			1,007		
d (departments)	33			33			33			33			33		
Fixed part:															
Escuela Nueva				-0.29		(1.28)	11.63	***	(1.59)	11.68	***	(1.59)	11.64	***	(1.59)
Male				-10.25	***	(0.32)	-10.24	***	(0.32)	-10.25	***	(0.32)	-10.25	***	(0.32)
EN*Male				1.46		(0.93)	1.459		(0.93)	1.46		(0.93)	1.46		(0.93)
Rural							1.730		(1.51)	1.70		(1.51)	1.71		(1.51)
Private							40.27	***	(2.00)	40.20	***	(2.00)	40.21	***	(1.99)
Socioeconomic level							13.94	***	(0.79)	13.77	***	(0.79)	13.76	***	(0.79)
EN*Socioeconomic level							-7.68	***	(1.52)	-7.67	***	(1.52)	-7.69	***	(1.52)
w/ ethnic students							-4.27	***	(1.28)	-4.30	***	(1.28)	-4.37	***	(1.28)
w/ conflict victims							-6.26	***	(1.07)	-6.32	***	(1.07)	-6.33	***	(1.07)
Morning session							-1.09		(1.67)	-1.24		(1.67)	-1.23		(1.67)
Afternoon session							-7.00	***	(1.77)	-7.17	***	(1.76)	-7.15	***	(1.76)
Governance Index										0.27	***	(0.08)	0.27	***	(0.08)
pc. GDP (mio. pesos)													0.07		(0.25)
Educ. expenditure													0.02		(0.01)
Grand mean	285.80	***	(3.40)	290.83	***	(3.43)	279.68	***	(3.53)	280.40	***	(3.31)	282.16	***	(3.31)
Random part (sd):															
Department-level	17.28		(2.56)	17.25		(2.56)	14.20		(2.14)	12.56		(2.00)	11.40		(1.93)
Municipality-level	19.46		(0.79)	19.48		(0.79)	17.66		(0.77)	17.53		(0.78)	17.54		(0.78)
School-level	46.96		(0.43)	46.93		(0.43)	43.09		(0.43)	43.09		(0.43)	43.09		(0.43)
Student-level	59.59		(0.12)	59.38		(0.12)	59.46		(0.12)	59.46		(0.12)	59.46		(0.12)
ICC (schools)	0.34			0.34			0.31			0.32			0.32		
ICC (municipalities)	0.06			0.06			0.05			0.05			0.05		
ICC (departments)	0.05			0.05			0.03			0.03			0.02		
LR χ^2	267.16 ***						1545.77 ***			13.45 ***			2.82		

Standard errors in parenthesis. ***: p≤0.001, **: p≤0.01, *: p≤0.05. LR-test statistics reported for test against model to the respective left. Calculations based on plausible value 1 (results of other plausible values not qualitatively different).

Table 69 Results of the random-intercept models, language grade 5 (country-level study)

Language Grade 5	**Model 0**			**Model RI1**			**Model RI2**			**Model RI3**			**Model RI4**		
n (students)	277,179			277,179			277,179			277,179			277,179		
j (schools)	17,586			17,586			17,586			17,586			17,586		
m (municipalities)	1,011			1,011			1,011			1,011			1,011		
d (departments)	33			33			33			33			33		
Fixed part:															
Escuela Nueva				-4.57	***	(1.18)	9.15	***	(1.43)	9.23	***	(1.43)	9.22	***	(1.42)
Male				-13.54	***	(0.29)	-13.55	***	(0.29)	-13.55	***	(0.29)	-13.55	***	(0.29)
EN*Male				-0.28		(1.02)	-0.32		(1.00)	-0.31		(1.00)	-0.31		(1.00)
Rural							4.25	***	(1.22)	4.22	***	(1.22)	4.23	***	(1.22)
Private							26.77	***	(1.57)	26.71	***	(1.57)	26.69	***	(1.57)
Socioeconomic level							17.80	***	(0.63)	17.67	***	(0.63)	17.66	***	(0.63)
EN*Socioeconomic level							-9.40	***	(1.28)	-9.42	***	(1.27)	-9.43	***	(1.27)
w/ ethnic students							-4.28	***	(1.05)	-4.33	***	(1.05)	-4.38	***	(1.05)
w/ conflict victims							-3.36	***	(0.91)	-3.40	***	(0.91)	-3.40	***	(0.91)
Morning session							-3.78	**	(1.37)	-3.91	**	(1.37)	-3.96	**	(1.36)
Afternoon session							-10.69	***	(1.44)	-10.85	***	(1.44)	-10.89	***	(1.44)
Governance Index										0.23	***	(0.06)	0.24	***	(0.06)
pc. GDP (mio. pesos)													0.37		(0.26)
Educ. expenditure													0.02		(0.01)
Grand mean	283.91	***	(3.94)	292.74	***	(4.07)	279.31	***	(3.52)	280.16	***	(3.28)	281.68	***	(3.16)
Random part (sd):															
Department-level	20.78		(2.93)	21.40		(3.00)	16.11		(2.28)	14.49		(2.13)	12.81		(2.03)
Municipality-level	17.00		(0.69)	16.98		(0.69)	15.17		(0.65)	15.07		(0.65)	15.07		(0.65)
School-level	37.59		(0.38)	37.35		(0.38)	32.78		(0.38)	32.79		(0.37)	32.79		(0.37)
Student-level	67.61		(0.16)	67.29		(0.16)	67.37		(0.16)	67.37		(0.16)	67.37		(0.16)
ICC (schools)	0.21			0.21			0.18			0.18			0.18		
ICC (municipalities)	0.04			0.04			0.04			0.04			0.04		
ICC (departments)	0.06			0.07			0.04			0.03			0.03		
LR χ^2	466.85	***					2071.33	***		14.78	***		4.89		

Standard errors in parenthesis. ***: p≤0.001, **: p≤0.01, *: p≤0.05. LR-test statistics reported for test against model to the respective left. Calculations based on plausible value 1 (results of other plausible values not qualitatively different).

Table 70 Results of the random-intercept models, mathematics grade 3 (country-level study)

Mathematics Grade 3	**Model 0**	**Model RI1**	**Model RI2**	**Model RI3**	**Model RI4**
n (students)	195,978	195,978	195,978	195,978	195,978
j (schools)	17,475	17,475	17,475	17,475	17,475
m (municipalities)	1,009	1,009	1,009	1,009	1,009
d (departments)	33	33	33	33	33
Fixed part:					
Escuela Nueva		11.66 *** (1.46)	19.35 *** (1.89)	19.44 *** (1.89)	19.39 *** (1.89)
Male		0.58 (0.32)	0.59 (0.32)	0.59 (0.32)	0.59 (0.32)
EN*Male		-1.83 (1.11)	-1.81 (1.11)	-1.80 (1.11)	-1.80 (1.11)
Rural			4.58 ** (1.70)	4.55 ** (1.69)	4.55 ** (1.69)
Private			42.27 *** (2.19)	42.20 *** (2.19)	42.20 *** (2.19)
Socioeconomic level			11.41 *** (0.90)	11.26 *** (0.90)	11.23 *** (0.90)
EN*Socioeconomic level			-7.13 *** (1.68)	-7.16 *** (1.68)	-7.19 *** (1.68)
w/ ethnic students			-3.36 * (1.46)	-3.39 ** (1.45)	-3.41 ** (1.45)
w/ conflict victims			-8.35 *** (1.23)	-8.41 *** (1.23)	-8.44 *** (1.23)
Morning session			1.14 (1.86)	0.98 (1.85)	0.96 (1.85)
Afternoon session			-4.96 ** (1.92)	-5.14 ** (1.92)	-5.15 ** (1.91)
Governance Index				0.29 ** (0.09)	0.29 ** (0.09)
pc. GDP (mio. pesos)					0.32 (0.28)
Educ. expenditure					0.01 (0.01)
Grand mean	292.72 *** (3.89)	287.87 *** (3.67)	277.03 *** (4.01)	278.00 *** (3.74)	279.41 *** (3.79)
Random part (sd):					
Department-level	19.71 (3.00)	18.17 (2.80)	16.24 (2.51)	14.23 (2.32)	13.02 (2.32)
Municipality-level	23.76 (0.92)	23.66 (0.92)	22.86 (0.90)	22.80 (0.90)	22.82 (0.90)
School-level	52.43 (0.44)	52.31 (0.44)	49.14 (0.44)	49.15 (0.44)	49.15 (0.44)
Student-level	61.53 (0.11)	61.53 (0.12)	61.58 (0.12)	61.58 (0.12)	61.58 (0.12)
ICC (schools)	0.37	0.37	0.35	0.35	0.35
ICC (municipalities)	0.08	0.08	0.07	0.07	0.08
ICC (departments)	0.05	0.04	0.04	0.03	0.02
LR χ^2	258.92 ***		1159.69 ***	10.60 **	2.64

Standard errors in parenthesis. ***: p≤0.001, **: p≤0.01, *: p≤0.05. LR-test statistics reported for test against model to the respective left. Calculations based on plausible value 1 (results of other plausible values not qualitatively different).

Table 71 Results of the random-intercept models, mathematics grade 5 (country-level study)

Mathematics Grade 5	**Model 0**			**Model RI1**			**Model RI2**			**Model RI3**			**Model RI4**		
n (students)			274,404			274,404			274,404			274,404			274,404
j (schools)			17,200			17,200			17,200			17,200			17,200
m (municipalities)			1,009			1,009			1,009			1,009			1,009
d (departments)			33			33			33			33			33
Fixed part:															
Escuela Nueva				0.40		(1.15)	12.31	***	(1.43)	12.41	***	(1.43)	12.39	***	(1.43)
Male				7.21	***	(0.30)	7.22	***	(0.30)	7.22	***	(0.30)	7.22	***	(0.30)
EN*Male				-3.08	***	(0.88)	-3.06	***	(0.88)	-3.06	***	(0.88)	-3.07	***	(0.88)
Rural							4.57	***	(1.33)	4.54	***	(1.33)	4.53	***	(1.33)
Private							27.41	***	(1.74)	27.36	***	(1.74)	27.34	***	(1.74)
Socioeconomic level							15.89	***	(0.71)	15.76	***	(0.72)	15.75	***	(0.72)
EN*Socioeconomic level							-9.06	***	(1.39)	-9.08	***	(1.39)	-9.10	***	(1.38)
w/ ethnic students							-4.45	***	(1.15)	-4.50	***	(1.15)	-4.53	***	(1.15)
w/ conflict victims							-2.73	**	(0.96)	-2.76	**	(0.96)	-2.77	**	(0.96)
Morning session							-2.68		(1.49)	-2.83		(1.49)	-2.89		(1.49)
Afternoon session							-9.67	***	(1.52)	-9.83	***	(1.52)	-9.90	***	(1.52)
Governance Index										0.27	***	(0.08)	0.27	***	(0.08)
pc. GDP (mio. pesos)													0.56		(0.29)
Educ. expenditure													0.02		(0.01)
Grand mean	284.70	***	(4.26)	281.45	***	(4.32)	268.78	***	(3.98)	269.80	***	(3.69)	271.15	***	(3.55)
Random part (sd):															
Department-level	22.44		(3.22)	22.61		(3.25)	18.22		(2.65)	16.20		(2.44)	14.24		(2.36)
Municipality-level	20.28		(0.78)	20.28		(0.78)	19.24		(0.77)	19.15		(0.77)	19.18		(0.77)
School-level	40.28		(0.38)	40.34		(0.38)	36.77		(0.37)	36.78		(0.37)	36.78		(0.37)
Student-level	63.97		(0.11)	63.88		(0.11)	63.93		(0.11)	63.93		(0.11)	63.93		(0.11)
ICC (schools)	0.24			0.25			0.22			0.22			0.23		
ICC (municipalities)	0.06			0.06			0.06			0.06			0.06		
ICC (departments)	0.08			0.08			0.05			0.04			0.03		
LR χ^2	435.40 ***						1618.85 ***			11.12 ***			4.93		

Standard errors in parenthesis. ***: p≤0.001, **: p≤0.01, *: p≤0.05. LR-test statistics reported for test against model to the respective left. Calculations based on plausible value 1 (results of other plausible values not qualitatively different).

Table 72 Results of the random-intercept models, civic competencies grade 5 (country-level study)

Civics, Grade 5	**Model 0**			**Model RI1**			**Model RI2**			**Model RI3**			**Model RI4**		
n (students)	276,169			276,169			276,169			276,169			276,169		
j (schools)	17,533			17,533			17,533			17,533			17,533		
m (municipalities)	1,010			1,010			1,010			1,010			1,010		
d (departments)	33			33			33			33			33		
Fixed part:															
Escuela Nueva				-3.57	***	(1.03)	8.36	***	(1.23)	8.41	***	(1.23)	8.40	***	(1.23)
Male				-18.59	***	(0.30)	-18.61	***	(0.30)	-18.61	***	(0.30)	-18.61	***	(0.30)
EN*Male				2.70	**	(0.87)	2.66	**	(0.86)	2.66	**	(0.86)	2.66	**	(0.86)
Rural							3.71	***	(1.06)	3.69	***	(1.06)	3.70	***	(1.06)
Private							26.01	***	(1.43)	25.97	***	(1.43)	25.96	***	(1.43)
Socioeconomic level							15.16	***	(0.56)	15.07	***	(0.56)	15.07	***	(0.56)
EN*Socioeconomic level							-8.10	***	(1.17)	-8.11	***	(1.17)	-8.12	***	(1.17)
w/ ethnic students							-3.31	***	(0.91)	-3.33	***	(0.91)	-3.39	***	(0.91)
w/ conflict victims							-3.02	***	(0.81)	-3.05	***	(0.81)	-3.05	***	(0.81)
Morning session							-2.99	**	(1.22)	-3.07	*	(1.22)	-3.11	*	(1.22)
Afternoon session							-7.55	***	(1.26)	-7.64	***	(1.26)	-7.68	***	(1.26)
Governance Index										0.16	**	(0.06)	0.17	**	(0.06)
pc. GDP (mio. pesos)													0.32		(0.25)
Educ. expenditure													0.02	*	(0.01)
Grand mean	281.25	***	(3.59)	291.66	***	(3.65)	279.46	***	(3.25)	280.00	***	(3.09)	281.50	***	(2.94)
Random part (sd):															
Department-level	18.98		(2.65)	19.26		(2.69)	15.09		(2.13)	14.00		(2.03)	12.30		(1.93)
Municipality-level	15.84		(0.63)	15.84		(0.63)	14.28		(0.61)	14.23		(0.61)	14.24		(0.61)
School-level	32.21		(0.32)	31.92		(0.32)	27.63		(0.31)	27.64		(0.31)	27.64		(0.31)
Student-level	65.05		(0.12)	64.45		(0.12)	64.54		(0.12)	64.54		(0.12)	64.54		(0.12)
ICC (schools)	0.18			0.18			0.14			0.14			0.14		
ICC (municipalities)	0.04			0.04			0.04			0.04			0.04		
ICC (departments)	0.06			0.06			0.04			0.04			0.03		
LR χ^2	460.25 ***						1985.12***			7.29 **			4.80		

Standard errors in parenthesis. ***: p≤0.001, **: p≤0.01, *: p≤0.05. LR-test statistics reported for test against model to the respective left. Calculations based on plausible value 1 (results of other plausible values not qualitatively different).

2 Development of the implementation index model (department-level)

The starting point for the development of the full multilevel model is the two-level null model, given that between-municipality differences in Quindío were found to be negligible. As a first step, the main variable of interest and level one predictors are added to the model. The Quindío Random Intercept Model 1 is defined as:

Model QRI1: $score_{ijm} = \beta_0 + \beta_1 ENI_{jm} + \beta_2 male_{ijm} + \beta_3 (male * ENI)_{ijm} + \xi_{ijm}$

, where ξ_{ijm} is the composed error term consisting of the student-level error term ε_{ijm} and the school-level error term ζ_{jm}. Subscripts are added to all regressors to indicate the level on which the variables change (i for students, j for schools, and m for municipalities). Apart from the EN implementation index, the model includes a predictor for gender ($male$) and a cross-level interaction term of the implementation index and gender ($male * ENI$). This interaction term is testing the hypothesis that the effect of the EN model differs by gender.

Estimation results for model QRI1 for all grades and areas are presented in Table 73 through Table 77. In this section, only the main variables of interest and their individual and joint significance are discussed; a more thorough description of the results follows in section 6.3.3 based on the final multilevel model.

The results for model QRI1 show that even though the coefficient on the EN implementation index has the expected positive sign in each model, the effect is not statistically significant in any of the grades or areas. The same is true for the effect of gender: as expected, the estimation coefficients for $male$ are negative for language and civics and positive for mathematics, yet the effect is not statistically significant; nor is the interaction between gender and implementation index.

The last rows of the table help to evaluate whether the newly introduced coefficients are jointly significant – in this first case, whether introducing them is an improvement compared to the null model. For the country-level analysis (chapter 4), this was done using the likelihood-ratio statistic. That test, and its alternative, the Wald test, is based on asymptotic $\chi^2(q)$ null distributions (q denoting the degrees of freedom as determined by the number of restrictions). This means that the two tests are asymptotically equivalent – yet in in a relatively small sample like the one available for the department-level study, the tests may produce conflicting conclusions (Rabe-Hesketh and Skrondal 2012, 138). While only the likelihood-ratio test can be used to test random coefficients, the Wald test has a practical advantage for fixed parameters: It can be performed after Stata's multiple imputation command based on all five plausible values, which is not possible with the likelihood ratio test (StataCorp 2013). Given the large sample in the country-level study, performing the test on each plausible value individually generated very similar results – yet in the case of the small sample used for the department-level analysis, the conclusions based on running the likelihood ratio tests for the five different plausible values separately differ. Therefore, this section of the study generally uses the Wald test for hypothesis testing in order to be able to make use of the information provided in all plausible values.

Coming back to the results presented in in Table 73 through Table 77, the conclusions based on the Wald test differ between the models. For the case of grade 5 language and civic competencies, the test shows that the three coefficients are jointly significant, even though none of them is individually. For mathematics and grade 3 language scores, the null hypotheses of joint significance are rejected. However, given that these coefficients are central to the research hypotheses, all coefficients are retained in the model.

The next step is the addition of core school-level regressors, that is, school-level regressors that are necessary to test the research hypotheses. These are the average socioeconomic level of the school (NSE) and the interaction of socioeconomic level and the EN implementation index ($NSE * ENI$). The second Quindío Random Intercept Model is thus:

Model QRI2: $$score_{ijm} = \beta_0 + \beta_1 ENI_{jm} + \beta_2 male_{ijm} + \beta_3 (male * ENI)_{ijm} + \beta_4 NSE_{jm} + \beta_5 (NSE * ENI)_{jm} + \xi_{ijm}$$

The estimation results can be found in the second columns of Table 73 through Table 77. As for model QRI1, even though most coefficients have the expected sign, none of them are statistically significant. Based on a Wald test for the joint significance of the newly introduced variables, the null hypothesis of a lack of a joint effect cannot be rejected, except in the case of civic competencies (F(2, 222.6) = 3.24, p=0.041). Despite the non-significant results, all of the regressors are retained in the model because of their importance to the research hypothesis.

Apart from the core school-level regressors, there are a few variables that have been shown to have a significant effect on learning outcomes in the country-level study. At the school-level, these are the presence of students with ethnic background ($ethnic$) and who are victims of the conflict ($conflict$), as well as the session type of the school. As there are no full session primary schools in Quindío, morning session schools are treated as base level and the dummy variable $afternoon$ is added to the model where appropriate (it is omitted in grade 5 estimates as there are no schools in the sample that have a grade 5 afternoon session). Model QRI3 is thus defined as follows:

Model QRI3: $$score_{ijm} = \beta_0 + \beta_1 ENI_{jm} + \beta_2 male_{ijm} + \beta_3 (male * ENI)_{ijm} + \beta_4 NSE_{jm} + \beta_5 (NSE * ENI)_{jm} + \beta_6 ethnic_{jm} + \beta_7 conflict_{jm} + \beta_8 afternoon_{jm} + \xi_{ijm}$$

As shown in the third columns of Table 73 through Table 77, $ethnic$ is significant for grade 5 language and civics, and $conflict$ is significant for grade 5 mathematics. No other regressor is statistically significant. Hence, even after controlling for other school factors the null hypothesis of no effect of EN implementation fails to be rejected. Wald tests of the joint significant of the school-level control variables are significant for grade 5 language and mathematics; in these cases, the control variables are retained in the model. For the case of civic competencies, only $ethnic$ is retained.

The last step for the department-level random intercept model is the inclusion of the municipality-level control variable $governance$, which is the municipality-level governance index centered at the mean of all ten municipalities in the sample. Formally:

Model QRI4: $$score_{ijm} = \beta_0 + \beta_1 ENI_{jm} + \beta_2 male_{ijm} + \beta_3 (male * ENI)_{ijm} + \beta_4 NSE_{jm} + \beta_5 (NSE * ENI)_{jm} + \beta_6 ethnic_{jm} + \beta_7 conflict_{jm} + \beta_8 afternoon_{jm} + \beta_9 governance_m + \xi_{ijm}$$

, where $ethnic$ only appears in the grade 5 models, and $conflict$ only appears in the mathematics and language grade 5 models.

The last column of Table 73 through Table 77 shows the estimation results. Controlling for municipality governance does not help to make the effect of EN implementation, or any other regressor, significant (with the exception of $ethnic$ for language grade 5 and $conflict$ for mathematics grade 5, as was already the case in model QRI3). Governance itself is not significant in any of the models.

The final random intercept model based on the total implementation index is thus model QRI2 for grade 3 estimations, and model QRI3 for grade 5 estimations. The models are summarized in Table 41 on page 197.

Table 73 Results of the random intercept models for the overall index, language grade 3 (department-level study)

Language Grade 3	**Model QRI1**		**Model QRI2**		**Model QRI3**		**Model QRI4**	
n (students)	252		252		245		252	
j (schools)	66		66		65		66	
Fixed part:								
EN Index	0.75	(0.75)	0.92	(1.55)	0.42	(1.62)	0.92	(1.55)
Male	-19.70	(-19.70)	-20.28	(19.64)	-16.08	(21.21)	-20.30	(19.63)
ENI*Male	0.07	(0.07)	0.08	(0.50)	0.05	(0.53)	0.08	(0.50)
Socioec. level			14.20	(59.13)	7.02	(59.51)	13.66	(59.51)
ENI*Socioec. level			-0.16	(1.45)	0.00	(1.48)	-0.16	(1.45)
w/ ethnic students					-18.56	(24.49)		
w/ conflict victims					-21.39	(15.70)		
Afternoon session					-15.62	(31.07)		
Governance							0.12	(1.53)
Grand mean	285.18***	(26.84)	271.11***	(63.35)	304.69***	(66.14)	271.72***	(63.66)
Random part (sd):								
School-level	41.90	(7.93)	41.45	(8.06)	40.09	(8.53)	41.44	(8.06)
Student-level	56.19	(3.32)	56.22	(3.33)	56.15	(3.27)	56.22	(3.33)
ICC (schools)	0.36		0.35		0.34		0.35	
Wald test statistic	F(3, 376.1) = 1.94		F(2,1000.3) = 0.21		F(3, 200.1) = 0.92		F(1,9003.2) = 0.01	
Wald test p-value	0.123		0.814		0.434		0.940	

Standard errors in parenthesis. ***p≤0.001; ** p<0.01; * p<0.05

Table 74 Results of the random intercept models for the overall index, language grade 5 (department-level study)

Language Grade 5	**Model QRI1**		**Model QRI2**		**Model QRI3**		**Model QRI4**	
n (students)	376		376		369		376	
j (schools)	72		72		71		72	
Fixed part:								
EN Index	0.51	(0.56)	-0.54	(1.49)	-1.60	(1.47)	-1.67	(1.44)
Male	-21.64	(20.60)	-21.60	(20.63)	-20.49	(20.57)	-19.25	(20.83)
ENI*Male	-0.19	(0.49)	-0.19	(0.49)	-0.28	(0.49)	-0.26	(0.49)
Socioec. level			-28.72	(55.72)	-53.19	(52.18)	-49.10	(51.70)
ENI*Socioec. level			1.06	(1.40)	1.70	(1.34)	1.76	(1.30)
w/ ethnic students					-50.38*	(20.76)	-50.58*	(20.69)
w/ conflict victims					-15.73	(13.09)	-17.80	(13.25)
Afternoon session					*omitted*			
Governance							-0.59	(1.44)
Grand mean	302.47***	(22.61)	331.47***	(59.87)	388.16***	(59.76)	385.50***	(59.28)
Random part (sd):								
School-level	38.14	(7.00)	38.13	(6.79)	32.25	(7.47)	31.21	(7.39)
Student-level	66.46	(2.75)	66.28	(2.73)	66.26	(2.79)	66.62	(2.79)
ICC (schools)	0.25		0.25		0.19		0.18	
Wald test statistic	F(3, 271.1) = 4.80		F(2, 315.5) = 0.83		F(2,1138.5) = 3.83		F(1, 696.4) = 0.17	
Wald test p-value	0.003		0.437		0.022		0.682	

Standard errors in parenthesis. ***p≤0.001; ** p<0.01; * p<0.05

Table 75 Results of the random intercept models for the overall index, math grade 3 (department-level study)

Mathematics Grade 3	**Model QRI1**		**Model QRI2**		**Model QRI3**		**Model QRI4**	
n (students)	254		254		246		254	
j (schools)	63		63		62		63	
Fixed part:								
EN Index	0.79	(0.90)	0.79	(2.23)	0.97	(2.39)	0.79	(2.23)
Male	30.65	(22.34)	31.03	(22.32)	34.36	(22.66)	30.71	(22.31)
ENI*Male	-0.81	(0.58)	-0.80	(0.58)	-0.86	(0.59)	-0.79	(0.58)
Socioec. level			19.17	(83.69)	21.42	(86.19)	11.62	(84.33)
ENI*Socioec. level			0.05	(2.08)	-0.05	(2.18)	0.06	(2.08)
w/ ethnic students					8.59	(37.27)		
w/ conflict victims					10.90	(22.11)		
Afternoon session					-17.90	(28.40)		
Governance							1.40	(2.22)
Grand mean	281.92***	(35.50)	262.58**	(89.68)	248.17*	(97.25)	270.34**	(90.34)
Random part (sd):								
School-level	66.38	(10.10)	65.49	(10.06)	66.53	(10.23)	65.45	(10.01)
Student-level	61.71	(3.81)	61.66	(3.80)	61.08	(3.84)	61.61	(3.79)
ICC (schools)	0.54		0.53		0.54		0.53	
Wald test statistic	F(3, 310.5) = 0.83		F(2,1585.1) = 0.70		F(3,1543.4) = 0.23		F(1,8916.1) = 0.39	
Wald test p-value	0.479		0.495		0.878		0.530	

Standard errors in parenthesis. ***p≤0.001; ** p<0.01; * p<0.05

Table 76 Results of the random intercept models for the overall index, math grade 5 (department-level study)

Mathematics Grade 5	**Model QRI1**		**Model QRI2**		**Model QRI3**		**Model QRI4**	
n (students)	318		318		312		318	
j (schools)	60		60		59		60	
Fixed part:								
EN Index	1.12	(0.64)	1.37	(1.86)	0.41	(1.68)	0.69	(1.64)
Male	15.38	(18.37)	15.52	(18.43)	16.70	(18.56)	17.59	(18.62)
ENI*Male	-0.26	(0.46)	-0.27	(0.47)	-0.34	(0.47)	-0.33	(0.47)
Socioec. level			2.67	(76.78)	-24.29	(68.08)	11.96	(67.63)
ENI*Socioec. level			-0.24	(1.86)	0.23	(1.64)	-0.17	(1.60)
w/ ethnic students					-25.21	(21.79)	-22.43	(21.24)
w/ conflict victims					-38.20**	(14.51)	-45.98**	(14.39)
Afternoon session					*omitted*			
Governance							-2.05	(1.48)
Grand mean	256.50***	(25.76)	252.77***	(74.97)	321.30***	(68.94)	295.11***	(68.55)
Random part (sd):								
School-level	45.45	(7.01)	45.15	(7.05)	37.44	(7.37)	35.13	(7.67)
Student-level	56.19	(3.34)	56.20	(3.35)	56.28	(3.60)	56.70	(3.45)
ICC (schools)	0.40		0.39		0.31		0.28	
Wald test statistic	F(3, 179.0) = 1.03		F(2,1265.6) = 0.14		F(2,1078.1) = 4.35		F(1,28756.2) = 1.92	
Wald test p-value	0.378		0.868		0.0131		0.166	

Standard errors in parenthesis. ***p≤0.001; ** p<0.01; * p<0.05

Table 77 Results of the random intercept models for the overall index, civics grade 5 (department-level study)

Civics Grade 5	**Model QRI1**		**Model QRI2**		**Model QRI3**		**Model QRI4**	
n (students)	378		378		371		378	
j (schools)	75		75		74		75	
Fixed part:								
EN Index	0.67	(0.54)	2.13	(1.30)	1.23	(1.30)	1.51	(1.30)
Male	-18.00	(16.22)	-20.20	(16.27)	-19.81	(16.27)	-19.68	(16.35)
ENI*Male	-0.43	(0.41)	-0.39	(0.41)	-0.43	(0.41)	-0.41	(0.41)
Socioec. level			82.29	(50.11)	67.52	(48.13)	69.36	(48.92)
ENI*Socioec. level			-1.46	(1.23)	-0.85	(1.21)	-1.03	(1.20)
w/ ethnic students					-38.03*	(18.79)	-35.49	(19.17)
w/ conflict victims					-13.21	(12.40)		
Afternoon session					*omitted*			
Governance							-0.05	(1.34)
Grand mean	305.11***	(21.87)	224.76***	(52.26)	265.37***	(53.06)	249.79***	(53.32)
Random part (sd):								
School-level	39.32	(5.98)	35.35	(6.18)	31.93	(6.79)	32.15	(6.55)
Student-level	59.91	(2.53)	60.02	(2.53)	60.03	(2.58)	60.29	(2.56)
ICC (schools)	0.30		0.26		0.22		0.22	
Wald test statistic	F(3, 487.6) = 9.01		F(2, 222.6) = 3.24		F(2, 954.0) = 2.66		F(1, 455.3) = 0.00	
Wald test p-value	<0.000		0.041		0.070		0.972	

Standard errors in parenthesis. ***p≤0.001; ** p<0.01; * p<0.05

3 Development of the implementation index-dimensions model (department-level)

The EN model consists of a wide range of different elements. Despite Fundación Escuela Nueva's focus on the holistic nature of the model, it is conceivable that some aspects of the model are more strongly correlated with positive learning outcomes than others. Therefore, a second series of department-level hierarchical models are developed which use the five index dimensions instead of the overall index to identify EN implementation.

Starting again from the two-level null model, in a first step the five implementation index dimensions ($D1$ to $D5$) are added to the model, each of them rescaled so that zero is the lowest value observed in the sample. Additionally, student-level regressors are added to the model, namely, gender ($male$) and interaction terms of gender and the area indices ($male * D1$ to $male * D5$). The Quindío random intercept model 5 thus is defined as:

Model QRI5:
$$\begin{aligned} score_{ijm} = {} & \beta_0 + \beta_{1\_1} EN\ D1_{jm} + \beta_{1\_2} EN\ D2_{jm} + \beta_{1\_3} EN\ D3_{jm} + \\ & \beta_{1\_4} EN\ D4_{jm} + \beta_{1\_5} EN\ D5_{jm} + \beta_2 male_{ijm} + \beta_{3\_1}(male * D1)_{ijm} + \\ & \beta_{3\_2}(male * D2)_{ijm} + \beta_{3\_3}(male * D3)_{ijm} + \beta_{3\_4}(male * D4)_{ijm} + \\ & \beta_{3\_5}(male * D5)_{ijm} + \xi_{ijm} \end{aligned}$$

As before, ξ_{ijm} is the composed error term consisting of the student-level error term ε_{ijm} and the school-level error term ζ_{jm}, and subscripts indicate the level on which the variables vary (i for students, j for schools, and m for municipalities).

The estimation results for the model are presented in the first columns of Table 78 to Table 82. The only index dimension with a consistent sign is dimension 4 (learning guides), which always has a positive sign—yet the coefficient is statistically significant only in the case of language grade

5. The only other instance of a significant coefficient is dimension 2 (classroom organization), equally for language grade 5; yet the effect of implementation is estimated to be negative (i.e. the more elements of classroom organization are implemented, the lower the language exam score). The coefficients for dimension 1 (teacher training), 3 (school and community) and 5 (roles of students) are not significant for any area or grade.

The last part of the table shows results for the Wald tests of joint significance of regressor groups. The first test statistic is for the joint significance of the five index dimensions. The test suggests that the five dimensions are not jointly significant in any of the areas or grades, which was expected given that the overall index was not significant in model QRI1, either. The second Wald test checks whether the other variables (gender and interaction terms with gender) are jointly significant. This is only the case for language grade 5 ($F(6, 572.3) = 3.03$; $p = 0.006$) and for civic competencies ($F(6, 387.5) = 4.08, p = 0.001$). In both cases, the Wald test of joint significance of only the interaction terms—i.e., without the coefficient on $male$—is *not* significant (not reported in the table). Given that $male$ alone is not significant, this suggests some statistical interplay between gender and EN implementation, yet the effect is weak.

Despite the interaction terms' importance for the hypotheses, they are removed from the model temporarily except in the two cases with joint significance. This is done because five additional coefficients can make an important difference in a sample of only between 252 and 376 observations, as they reduce the likelihood of detecting statistically significant effects.

The next step is the inclusion of school-level core indicators, namely, socioeconomic level (NSE) and the interaction between socioeconomic level and EN implementation ($NSE * D1$ to $NSE * D5$). Hence, model QRI6 is defined as:

Model QRI6: $score_{ijm} = \beta_0 + \beta_{1\_1}EN\ D1_{jm} + \beta_{1\_2}EN\ D2_{jm} + \beta_{1\_3}EN\ D3_{jm} + \beta_{1\_4}EN\ D4_{jm} + \beta_{1\_5}EN\ D5_{jm} + \beta_2 male_{ijm} + \beta_{3\_1}(male * D1)_{ijm} + \beta_{3\_2}(male * D2)_{ijm} + \beta_{3\_3}(male * D3)_{ijm} + \beta_{3\_4}(male * D4)_{ijm} + \beta_{3\_5}(male * D5)_{ijm} + \beta_4 NSE_{jm} + \beta_{5\_1}(NSE * D1)_{jm} + \beta_{5\_2}(NSE * D2)_{jm} + \beta_{5\_3}(NSE * D3)_{jm} + \beta_{5\_4}(NSE * D4)_{jm} + \beta_{5\_5}(NSE * D5)_{jm} + \xi_{ijm}$

The coefficients β_{3_1} to β_{3_5} are omitted for the grade 3 models and the grade 5 mathematics model. The estimation results for the model are presented in the second columns of Table 78 to Table 82. Neither socioeconomic level nor its interaction with the index dimensions are statistically significant in any of the models; nor does the inclusion of these control variables help to obtain significant results for the index dimensions. Based on the respective Wald tests at the bottom of the table, neither the Index dimensions nor socioeconomic level and its interactions are jointly significant. The conclusion is to remove the interaction terms from the model even though they are testing a research hypothesis, again in order to conserve degrees of freedom of the model. Socioeconomic level is retained in the model in order to be able to detect a possible effect.

It is possible that school-level control variables help to remove noise from the model and to thus uncover a statistically significant effect of EN implementation. At the school-level, these control variables are $ethnic$, $conflict$, and $afternoon$ – dummies for the presence of ethnic students, or students who are conflict victims, and for the session type, respectively. They are thus added to model QRI7:

Model QRI7: $score_{ijm} = \beta_0 + \beta_{1\_1}EN\ D1_{jm} + \beta_{1\_2}EN\ D2_{jm} + \beta_{1\_3}EN\ D3_{jm} + \beta_{1\_4}EN\ D4_{jm} + \beta_{1\_5}EN\ D5_{jm} + \beta_2 male_{ijm} + \beta_{3\_1}(male * D1)_{ijm} +$

$$\beta_{3\_2}(male * D2)_{ijm} + \beta_{3\_3}(male * D3)_{ijm} + \beta_{3\_4}(male * D4)_{ijm} + \beta_{3\_5}(male * D5)_{ijm} + \beta_4 NSE_{jm} + \beta_6 ethnic_{jm} + \beta_7 conflict_{jm} + \beta_8 afternoon_{jm} + \xi_{ijm}$$

Again, coefficients β_{3_1} to β_{3_5} are omitted for the grade 3 models and the grade 5 mathematics model. The third columns of Table 78 to Table 82 contain the estimation results. Significant effects can only be detected for $conflict$ for grade 5 mathematics and $ethnic$ for civic competencies; additionally, the effect of dimension 2 (classroom organization) is negative and significant for grade 5 language. The Wald tests suggest a lack of joint significance of the indicator dimensions for all grades and areas, and a joint significance of the school-level control variables for grade 5 mathematics and civic competencies. Only in these two areas are the control variables thus retained.

Finally, the municipality-level control variable $governance$ is added to the model. Model QRI8 is defined as:

Model QRI8: $$score_{ijm} = \beta_0 + \beta_{1\_1}EN\ D1_{jm} + \beta_{1\_2}EN\ D2_{jm} + \beta_{1\_3}EN\ D3_{jm} + \beta_{1\_4}EN\ D4_{jm} + \beta_{1\_5}EN\ D5_{jm} + \beta_2 male_{ijm} + \beta_{3\_1}(male * D1)_{ijm} + \beta_{3\_2}(male * D2)_{ijm} + \beta_{3\_3}(male * D3)_{ijm} + \beta_{3\_4}(male * D4)_{ijm} + \beta_{3\_5}(male * D5)_{ijm} + \beta_4 NSE_{jm} + \beta_6 ethnic_{jm} + \beta_7 conflict_{jm} + \beta_9 governance_m + \xi_{ijm}$$

, where coefficients β_{3_1} to β_{3_5} are omitted in the grade 3 models and the grade 5 mathematics model, and β_6 and β_7 are omitted in the grade 3 models and the grade 5 language model. Estimation results are presented in the last column of Table 78 to Table 82. Governance turns out to be not significant in this sample and model, as are the dimensions of the implementation index, both individually and collectively.

The final model thus differs between grades and areas: It includes school-level control variables for grade 5 mathematics and civic competences, but only the hypotheses-testing variables for the other grades and areas. The final models and results are summarized in Table 42 on page 198.

Table 78 Results of the random intercept models for the index dimensions, language grade 3 (department-level study)

Language Grade 3	**Model QRI5**		**Model QRI6**		**Model QRI7**		**Model QRI8**	
n (students)	252		252		245		252	
j (schools)	66		66		65		66	
Fixed part:								
Dim. 1 (Training)	-0.22	(0.46)	-0.96	(0.73)	0.26	(0.39)	-0.03	(0.37)
Dim. 2 (Classroom)	-0.67	(0.70)	1.14	(1.29)	-0.53	(0.59)	-0.41	(0.62)
Dim. 3 (Community)	0.72	(0.53)	1.24	(0.81)	0.78	(0.41)	0.79	(0.42)
Dim. 4 (Guides)	0.81	(0.79)	-0.20	(1.64)	-0.28	(0.75)	0.31	(0.73)
Dim. 5 (Roles)	0.32	(0.89)	-1.78	(2.23)	0.23	(0.70)	0.37	(0.74)
Male	-8.29	(31.37)	-16.81*	(8.06)	-13.12	(8.39)	-16.66*	(8.06)
EN D1*Male	0.48	(0.47)						
EN D2*Male	0.55	(0.70)						
EN D3*Male	0.07	(0.63)						
EN D4*Male	-1.28	(0.83)						
EN D5*Male	0.46	(0.99)						
Socioec. level			-46.50	(98.20)	7.47	(14.78)	7.51	(14.75)
EN D1*Socec. level			1.27	(0.75)				
EN D2*Socec. level			-1.39	(1.09)				
EN D3*Socec. level			-0.81	(0.85)				
EN D4*Socec. level			0.38	(1.74)				
EN D5*Socec. level			2.23	(2.24)				
w/ ethnic students					-31.43	(24.28)		
w/ conflict victims					-26.51	(17.13)		
Afternoon session					-19.78	(29.60)		
Governance							0.27	(1.60)
Grand mean	273.83 ***	(41.06)	315.76***	(99.02)	327.02***	(41.84)	271.29***	(36.55)
Random part (sd):								
School-level	42.06	(8.30)	37.13	(8.30)	36.37	(9.05)	39.95	(7.96)
Student-level	55.06	(3.19)	56.01	(3.32)	56.28	(3.26)	56.16	(3.29)
ICC (schools)	0.37		0.31		0.29		0.34	
Wald test (D1-D5)	F(5, 710.9) = 0.63		F(5,1547.7) = 1.22		F(5, 826.0) = 1.01		F(1,2069.6) = 0.03	
Wald test p-value	0.678		0.295		0.409		0.866	
Wald test (new var.)	F(6, 735.0) = 1.36		F(6,4209.7) = 0.98		F(3, 203.2) = 1.47		F(1,2069.6) = 0.03	
Wald test p-value	0.229		0.435		0.224		0.866	

Standard errors in parenthesis. ***p≤0.001; ** p<0.01; * p<0.05

Table 79 Results of the random intercept models for the index dimensions, language grade 5 (department-level study)

Language Grade 5	**Model QRI5**		**Model QRI6**		**Model QRI7**		**Model QRI8**	
n (students)	376		376		369		376	
j (schools)	72		72		71		72	
Fixed part:								
Dim. 1 (Training)	-0.13	(0.33)	-0.50	(0.63)	0.06	(0.34)	-0.14	(0.33)
Dim. 2 (Classroom)	-1.32*	(0.63)	-0.56	(1.15)	-1.32*	(0.64)	-1.25	(0.64)
Dim. 3 (Community)	0.74	(0.41)	0.23	(0.75)	0.74	(0.40)	0.78	(0.41)
Dim. 4 (Guides)	1.26*	(0.64)	-0.13	(1.72)	0.75	(0.69)	1.17	(0.66)
Dim. 5 (Roles)	0.81	(0.81)	0.00	(2.05)	0.74	(0.80)	0.78	(0.85)
Male	5.26	(28.19)	5.59	(28.20)	4.15	(28.12)	5.44	(28.21)
EN D1*Male	-0.76	(0.45)	-0.75	(0.45)	-0.62	(0.44)	-0.75	(0.45)
EN D2*Male	1.00	(0.67)	1.01	(0.67)	0.97	(0.67)	1.01	(0.67)
EN D3*Male	-0.55	(0.52)	-0.56	(0.52)	-0.41	(0.53)	-0.54	(0.52)
EN D4*Male	-0.39	(0.74)	-0.38	(0.75)	-0.40	(0.74)	-0.42	(0.74)
EN D5*Male	-0.21	(1.04)	-0.28	(1.03)	-0.56	(1.05)	-0.23	(1.03)
Socioec. level			-110.05	(87.56)	7.97	(13.03)	10.73	(12.73)
EN D1*Socec. level			0.40	(0.63)				
EN D2*Socec. level			-0.86	(0.89)				
EN D3*Socec. level			0.61	(0.73)				
EN D4*Socec. level			1.43	(1.73)				
EN D5*Socec. level			0.85	(1.96)				
w/ ethnic students					-34.91	(20.65)		
w/ conflict victims					-16.83	(13.48)		
Afternoon session					(omitted)			
Governance							-0.64	(1.43)
Grand mean	264.88***	(28.47)	371.67***	(88.92)	301.71***	(34.57)	256.47***	(30.44)
Random part (sd):								
School-level	30.90	(7.43)	30.01	(7.25)	28.41	(8.04)	30.72	(7.26)
Student-level	66.34	(2.79)	66.13	(2.77)	66.21	(2.84)	66.28	(2.78)
ICC (schools)	0.18		0.17		0.16		0.18	
Wald test (D1-D5)	F(5,5814.5) = 1.66		F(5,2307.7) = 0.26		F(5,6542.9) = 1.26		F(5,5145.1) = 1.57	
Wald test p-value	0.140		0.933		0.279		0.166	
Wald test (new var.)	F(6, 572.3) = 3.03		F(6,2264.2) = 0.50		F(2,3415.1) = 2.25		F(1, 582.5) = 0.20	
Wald test p-value	0.006		0.806		0.105		0.653	

Standard errors in parenthesis. ***p≤0.001; ** p<0.01; * p<0.05

Table 80 Results of the random intercept models for the index dimensions, math grade 3 (department-level study)

Mathematics Grade 3	**Model QRI5**		**Model QRI6**		**Model QRI7**		**Model QRI8**	
n (students)	254		254		246		254	
j (schools)	63		63		62		63	
Fixed part:								
Dim. 1 (Training)	0.58	(0.63)	-0.61	(1.03)	0.32	(0.50)	0.35	(0.47)
Dim. 2 (Classroom)	-0.37	(1.04)	0.82	(1.76)	0.11	(0.96)	0.01	(0.92)
Dim. 3 (Community)	-0.50	(0.86)	-0.08	(1.18)	-0.61	(0.62)	-0.60	(0.61)
Dim. 4 (Guides)	0.13	(1.04)	2.12	(2.19)	-0.02	(1.06)	0.12	(0.97)
Dim. 5 (Roles)	1.49	(1.28)	-3.39	(3.35)	0.71	(1.07)	0.57	(1.08)
Male	48.31	(35.50)	2.14	(9.12)	2.92	(9.30)	1.67	(9.09)
EN D1*Male	-0.32	(0.57)						
EN D2*Male	0.33	(0.79)						
EN D3*Male	-0.26	(0.75)						
EN D4*Male	-0.25	(0.81)						
EN D5*Male	-0.77	(1.06)						
Socioec. level			56.39	(131.53)	15.50	(22.49)	10.09	(21.63)
EN D1*Socec. level			1.18	(1.10)				
EN D2*Socec. level			-0.48	(1.47)				
EN D3*Socec. level			-0.63	(1.24)				
EN D4*Socec. level			-2.74	(2.38)				
EN D5*Socec. level			4.26	(3.28)				
w/ ethnic students					10.13	(38.73)		
w/ conflict victims					8.56	(23.49)		
Afternoon session					-9.09	(28.32)		
Governance							1.31	(2.37)
Grand mean	261.27***	(50.61)	238.93***	(134.4)	263.80***	(59.66)	277.77***	(48.01)
Random part (sd):								
School-level	64.63	(9.85)	62.38	(9.75)	65.71	(9.94)	64.46	(9.73)
Student-level	61.54	(3.80)	62.03	(3.85)	61.42	(3.88)	61.94	(3.82)
ICC (schools)	0.52		0.50		0.53		0.52	
Wald test (D1-D5)	F(5,2175.5) = 0.56		F(5,1972.2) = 0.30		F(5,4576.1) = 0.41		F 5,4265.1) = 0.39	
Wald test p-value	0.728		0.912		0.845		0.856	
Wald test (new var.)	F(6,1333.1) = 0.48		F(6,6344.7) = 0.50		F(3,2343.5) = 0.11		F(1,22330.4) = 0.31	
Wald test p-value	0.826		0.808		0.957		0.580	

Standard errors in parenthesis. ***p≤0.001; ** p<0.01; * p<0.05

Table 81 Results of the random intercept models for the index dimensions, math grade 5 (department-level study)

Mathematics Grade 5	**Model QRI5**		**Model QRI6**		**Model QRI7**		**Model QRI8**	
n (students)	318		318		312		318	
j (schools)	60		60		59		60	
Fixed part:								
Dim. 1 (Training)	0.20	(0.38)	-0.37	(0.65)	0.42	(0.35)	0.33	(0.34)
Dim. 2 (Classroom)	0.76	(0.77)	2.78	(1.61)	0.37	(0.61)	0.20	(0.61)
Dim. 3 (Community)	0.77	(0.49)	0.98	(0.79)	0.56	(0.41)	0.46	(0.40)
Dim. 4 (Guides)	0.29	(0.74)	0.98	(1.88)	-0.48	(0.67)	-0.33	(0.64)
Dim. 5 (Roles)	-1.36	(0.99)	-4.52	(2.61)	-0.87	(0.74)	-0.65	(0.75)
Male	20.21	(28.92)	5.73	(7.59)	3.99	(7.56)	5.39	(7.55)
EN D1*Male	-0.42	(0.36)						
EN D2*Male	0.14	(0.83)						
EN D3*Male	-0.65	(0.53)						
EN D4*Male	-0.25	(0.82)						
EN D5*Male	1.10	(0.98)						
Socioec. level			47.01	(118.87)	-8.94	(15.53)	13.18	(14.25)
EN D1*Socec. level			0.65	(0.70)				
EN D2*Socec. level			-2.16	(1.68)				
EN D3*Socec. level			-0.63	(0.89)				
EN D4*Socec. level			-0.97	(1.93)				
EN D5*Socec. level			4.04	(2.57)				
w/ ethnic students					-45.16	(24.01)	-41.54	(23.95)
w/ conflict victims					-40.11***	(14.57)	-48.56**	(14.62)
Afternoon session					(omitted)			
Governance							-1.90	(1.46)
Grand mean	246.46***	(34.82)	204.49	(118.41)	338.84***	(35.72)	322.40***	(34.34)
Random part (sd):								
School-level	43.25	(6.91)	39.74	(7.28)	34.65	(7.40)	33.31	(7.92)
Student-level	55.58	(3.24)	56.41	(3.44)	56.33	(3.66)	56.74	(3.52)
ICC (schools)	0.38		0.33		0.27		0.26	
Wald test (D1-D5)	F(5,2187.8) = 1.52		F(5,1854.5) = 1.37		F(5,1239.6) = 1.04		F(5,1229.0) = 0.69	
Wald test p-value	0.181		0.233		0.393		0.633	
Wald test (new var.)	F(6, 420.1) = 1.08		F(6,3042.4) = 0.89		F(2, 928.3) = 5.66		F(1,649087.1) =1.70	
Wald test p-value	0.372		0.503		0.004		0.192	

Standard errors in parenthesis. ***p≤0.001; ** p<0.01; * p<0.05

Table 82 Results of the random intercept models for the index dimensions, civics grade 5 (department-level study)

Civics Grade 5	**Model QRI5**		**Model QRI6**		**Model QRI7**		**Model QRI8**	
n (students)	378		378		371		378	
j (schools)	75		75		74		75	
Fixed part:								
Dim. 1 (Training)	-0.35	(0.35)	-0.49	(0.63)	-0.18	(0.34)	-0.09	(0.33)
Dim. 2 (Classroom)	-0.73	(0.66)	0.80	(1.07)	-0.69	(0.62)	-0.71	(0.61)
Dim. 3 (Community)	0.67	(0.47)	0.70	(0.78)	0.70	(0.45)	0.78	(0.44)
Dim. 4 (Guides)	0.91	(0.63)	1.40	(1.53)	0.33	(0.64)	0.15	(0.63)
Dim. 5 (Roles)	0.48	(0.85)	-1.27	(2.05)	0.33	(0.79)	0.38	(0.80)
Male	-11.30	(25.70)	-11.60	(25.82)	-10.99	(25.45)	-11.64	(25.73)
EN D1*Male	0.18	(0.40)	0.20	(0.40)	0.22	(0.41)	0.15	(0.40)
EN D2*Male	-0.26	(0.73)	-0.11	(0.74)	-0.21	(0.72)	-0.23	(0.72)
EN D3*Male	0.02	(0.54)	-0.01	(0.54)	0.12	(0.55)	0.05	(0.53)
EN D4*Male	-0.63	(0.65)	-0.75	(0.65)	-0.63	(0.65)	-0.51	(0.65)
EN D5*Male	0.67	(0.93)	0.62	(0.93)	0.34	(0.93)	0.37	(0.94)
Socioec. level			86.04	(81.67)	34.69**	(12.20)	31.83**	(11.42)
EN D1*Socec. level			0.12	(0.64)				
EN D2*Socec. level			-1.20	(0.84)				
EN D3*Socec. level			0.05	(0.71)				
EN D4*Socec. level			-0.62	(1.61)				
EN D5*Socec. level			1.36	(1.97)				
w/ ethnic students					-48.60**	(18.63)	-47.72*	(18.51)
w/ conflict victims					-15.57	(12.62)	-15.49	(12.57)
Afternoon session					(omitted)			
Governance							-0.44	(1.28)
Grand mean	289.82 ***	(28.91)	204.67*	(81.97)	306.57***	(31.78)	312.01***	(30.71)
Random part (sd):								
School-level	36.03	(6.37)	30.84	(6.91)	27.03	(8.47)	25.92	(8.59)
Student-level	59.78	(2.54)	59.94	(2.53)	60.06	(2.67)	60.47	(2.67)
ICC (schools)	0.27		0.21		0.17		0.16	
Wald test (D1-D5)	F(5, 790.2) = 1.01		F(5,1797.4) = 0.62		F(5, 660.2) = 0.70		F(5, 641.2) = 0.85	
Wald test p-value	0.412		0.687		0.621		0.516	
Wald test (new var.)	F(6, 387.5) = 4.08		F(6,2680.9) = 1.42		F(2, 496.0) = 4.14		F(1, 572.8) = 0.12	
Wald test p-value	0.001		0.201		0.016		0.730	

Standard errors in parenthesis. ***p≤0.001; ** p<0.01; * p<0.05

ANNEX C: Implementation Index

(table starts on next page)

		TEACHER INDEX					STUDENT INDEX				
Indic. #		Final wght	Qu #	1 point	0.5 points	0.25 points	Final wght	Qu #	1 point	0.5 points	0.25 points
	1 Teacher Training										
	1 Pre-service training										
111	**1 Any pre-service training**	0.020	14	v14=1							
	2 Taller de Iniciacion										
112_1	**1** participated	0.010	15	v15=1							
112_2	**2** one week long	0.010	16	v16=3 | v16=2							
	3 Taller de Guias										
113_1	**1** participated	0.007	18	v18=1							
113_2	**2** one week long	0.007	19	v19=3 | v19=2							
113_3	**3** there were learning guides	0.007	21	v21=1							
	4 Workshops follow EN methodology										
114_1	**1** work with manual de docente	0.002	22_1	v22_1=1	v22_1=2						
114_2	**2** work with leaning guides	0.002	22_2	v22_2=1	v22_2=2						
114_3	**3** all types of activities	0.002	22_3	v22_3=1	v22_3=2						
114_4	**4** contextualizacion	0.002	22_4	v22_4=1	v22_4=2						
114_5	**5** group work	0.002	22_5	v22_5=1	v22_5=2						
114_6	**6** gobierno del taller	0.002	22_6	v22_6=1	v22_6=2						
114_7	**7** learning corners	0.002	22_7	v22_7=1	v22_7=2						
114_8	**8** library	0.002	22_8	v22_8=1	v22_8=2						
114_9	**9** strategies for community work	0.002	22_9	v22_9=1	v22_9=2						
	5 Managed to put into practice workshop										
115_1	**1** classroom organization	0.003	23_1	v23_1=1	v23_1=2						
115_2	**2** family/community reunion	0.003	23_2	v23_2=1	v23_2=2						
115_3	**3** student government	0.003	23_3	v23_3=1	v23_3=2						
115_4	**4** learning corners	0.003	23_4	v23_4=1	v23_4=2						
115_5	**5** library	0.003	23_5	v23_5=1	v23_5=2						
115_6	**6** microcentros	0.003	23_6	v23_6=1	v23_6=2						
	2 In-service training and support										
	1 Microcentros and experience exchange										
121_1	**1** are being organized	0.008	26	v26=1	v26=2	v26=3					
121_2	**2** foster exchange between teachers	0.008	28	v28=1							

121_3	**3** regular participation	0.008	27	v27=1 if v28==1							
121_4	**4** ReNueva (Comunidad Virtual)	0.008	25	v25=1							
122	**2 Model schools**	0.033	24	v24=1							
123	**3 Mentoring visits**	0.033	29	v29=1	v29=2	v29=3					
	2 Classroom organization										
	1 Learning corners										
211	**1 in classroom**	0.008	53	v53_1=1	v53_2=1						
	2 Stocked with appropriate materials										
212_1	**1** experimental materials	0.001	55_1	v55_1=1							
212_2	**2** printed materials	0.001	55_2	v55_2=1							
212_3	**3** observational materials	0.001	55_3	v55_3=1							
212_4	**4** manipulation materials	0.001	55_4	v55_4=1							
212_5	**5** materials produced by students	0.001	55_5	v55_5=1							
212_6	**6** commercial didactic materials	0.001	55_6	v55_6=1							
213	**3 Stocked by teachers, students, community**	0.008	57	v57=1 | v57=2							
214	**4 Continually expanded**	0.008	56	v56=2							
	5 For all subject areas										
215_1	**1** Language	0.002	54_1	v54_1=1			0.006	18_1	18_1=1		
215_2	**2** Mathematics	0.002	54_2	v54_2=1			0.006	18_2	18_2=1		
215_3	**3** Social Sciences	0.002	54_3	v54_3=1			0.006	18_3	18_3=1		
215_4	**4** Sciences	0.002	54_4	v54_4=1			0.006	18_4	18_4=1		
215_5	**5** Others	0.002	54_5	v54_5=1							
216	**6 often used**						0.025	19_2	19_2=1	19_2=2	19_2=3
221	**2 Flexible Furniture**	0.040	31_2	v31_2==1 | v31_2==2			0.050	7	7=2		
	3 Classroom library										
231	**1 in classroom**	0.013	49_1, 49_2	v49_1=1	v49_2=1		0.025	17	17=1	17=2	
232	**2 often used**	0.013	50	v50=1	v50=2	v50=3	0.025	19_1	19_1=1	19_1=2	19_1=3
	3 contains wide range of materials										
233_1	**1** learning guides	0.003	51_1	v51_1=1							
233_2	**2** school texts	0.003	51_2	v51_2=1							
233_3	**3** literature	0.003	51_3	v51_3=1							
233_4	**4** reference works	0.003	51_4	v51_4=1							
233_5	**5** others	0.003	51_5	v51_5=1							

	4 EN instruments										
	1 Instruments exist										
241_1	**1** Value board	0.005	70_3	v70_3=1	v70_3=2	v70_3=3	0.017	31_3	31_3=1		
241_2	**2** Buzon de compromisos	0.005	70_4	v70_4=1	v70_4=2	v70_4=3	0.017	31_4	31_4=1		
241_3	**3** Correo amistoso	0.005	70_5	v70_5=1	v70_5=2	v70_5=3	0.017	31_6	31_6=1		
241_4	**4** Diario de confidencias	0.005	70_7	v70_7=1	v70_7=2	v70_7=3					
242	**2 Instruments visibly displayed in classroom**	0.020	69	v69=1							
251	**5 Multigrade classrooms**	0.040	6	v6=1			0.050	4, 5	if 5/4<1		
	3 School and Community										
311	**1 Flat administrative structure**	0.067	61	v61=3							
	2 Parental involvement										
321	**1 Frequent contact**	0.033	72	v72=1	v72=2		0.083	35	35=1	35=2	35=3
322	**2 Travelling journal**	0.033	70_6	v70_6=1	v70_6=2	v70_6=3	0.083	31_5, 34	31_5==1 & 34=3		
323	**3 Participation in school work (application)**						0.083	16_3	16_3=1		
	3 Community Involvement										
331	**1 Map/croquis**	0.017	71_1	v71_1=1							
332	**2 Monograph**	0.017	71_2	v71_2=1							
333	**3 Family book**	0.017	71_3	v71_3=1							
334	**4 Dia de logros**	0.017	74	v74=3							
	4 Learning Guides										
	1 Teacher guides										
411	**1 Teacher has teachers guide**	0.025	34	v34=1							
412	**2 Regular use**	0.025	35	v35=1	v35=2	v35=3					
	2 Student guides										
	1 Guides available in all subjects										
421_1	**1** Language	0.005	38_1	v38_1=1			0.013	11_1	11_1=1		
421_2	**2** Mathematics	0.005	38_2	v38_2=1			0.013	11_2	11_2=1		
421_3	**3** Social Sciences	0.005	38_3	v38_3=1			0.013	11_3	11_3=1		
421_4	**4** Sciences	0.005	38_4	v38_4=1			0.013	11_4	11_4=1		
421_5	**5** Others	0.005	38_5	v38_5=1			0.013	11_5	11_5=1		
	2 Guides frequently used										
422_1	**1** Language	0.005	40_1	v40_1=1	v40_1=2	v40_1=3					
422_2	**2** Mathematics	0.005	40_2	v40_2=1	v40_2=2	v40_2=3					

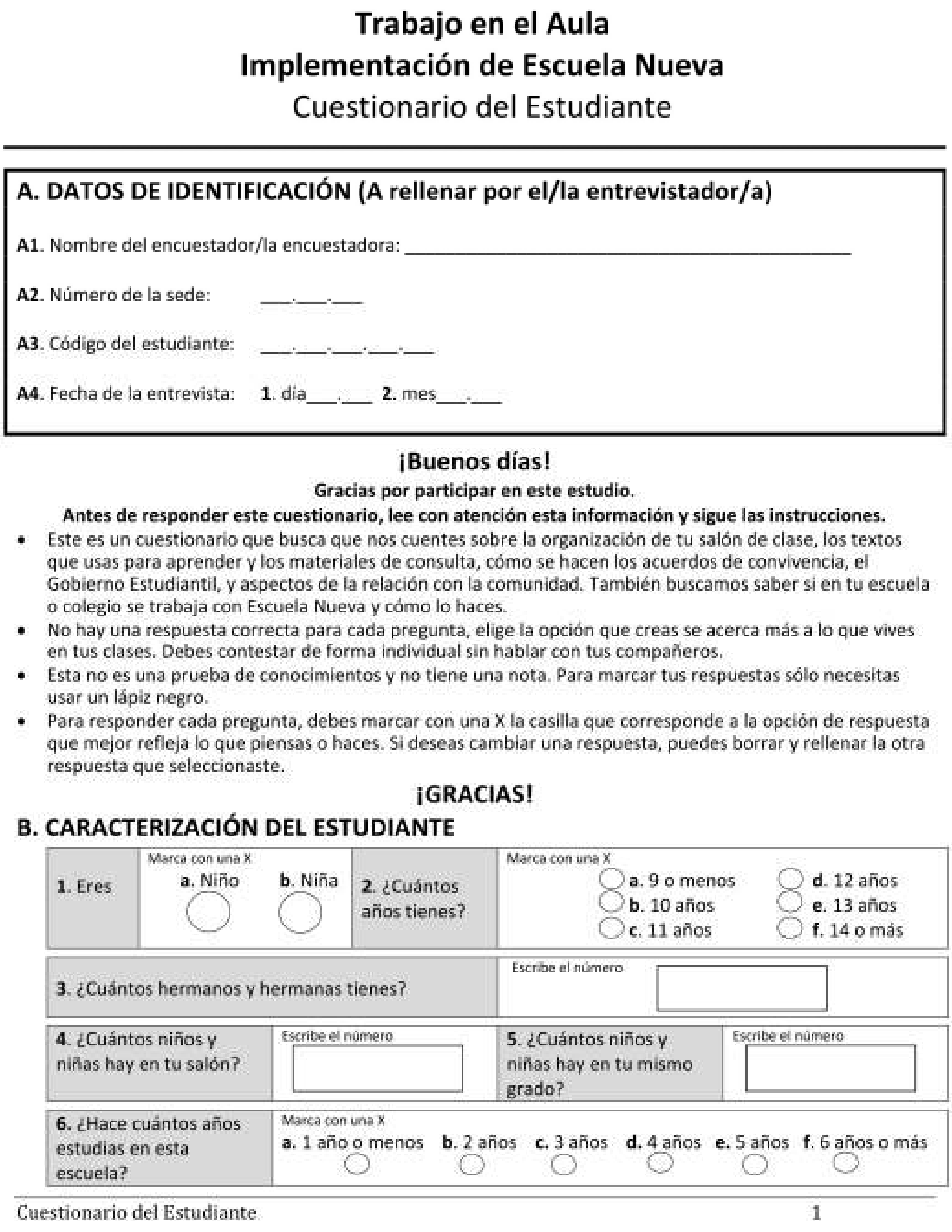
422_3	3 Social Sciences	0.005	40_3	v40_3=1	v40_3=2	v40_3=3				
422_4	4 Sciences	0.005	40_4	v40_4=1	v40_4=2	v40_4=3				
422_5	5 Others	0.005	40_5	v40_5=1	v40_5=2	v40_5=3				
423	**3 One guide per student**						0.063	9	9=3 | 9=4	
	3 Proper use of guides by students									
431	**1 Complement guide with other materials**	0.025	43	v43=3						
432	**2 Do activities in own note book**	0.025	44	v44=3			0.042	12_3	12_3=1	
433	**3 Do not write in guide book**						0.042	12_4	12_4=0	
434	**4 Use alone, in pairs, and in groups**						0.042	10	10=1	
	4 Proper use of guides by teachers									
	1 Modifications to guides									
441_1	1 Adaptions to context	0.008	41	v41=1						
441_2	2 Frequent modifications	0.008	47	v47=1 | v47=2						
	2 Holistic use/all activities									
442_1	1 basic activities	0.006	42_1	v42_1=1	v42_1=2	v42_1=3				
442_2	2 practice activities	0.006	42_2	v42_2=1	v42_2=2	v42_2=3				
442_3	3 application activities	0.006	42_3	v42_3=1	v42_3=2	v42_3=3				
443	**3 Use to promote active learning**	0.017	45	v45=1						
	5 Roles of Students									
	1 Student-centered/active learning									
511	**1 Work alone, in pairs, and groups**						0.025	20	20=2	
511_1	1 work in pairs	0.007	32_2	v32_2=2						
511_2	2 work in small groups	0.007	32_3	v32_3=2						
511_3	3 not always work alone	0.007	32_1	v32_1!=1						
512	**2 Teachers as guides**	0.020	58	v58=3 | v58=2			0.025	24	24=1	
513	**3 Flexible promotion**	0.020	30	v30=1						
514	**4 Assistance self-reported**	0.020	70_1	v70_1=1	v70_1=2	v70_1=3				
514_1	1 each student self-reports						0.013	32	32=3	
514_2	2 students understand purpose						0.013	33	33=4	
515	**5 Peer-to-peer tutoring**	0.020	33	v33=1	v33=2	v33=3	0.025	23	23=1	23=2
	6 Progress report									
516_1	1 in own note book						0.013	13	13=1	
516_2	2 with the professor						0.013	14_3	14_3=1	

	2	School democracy and Shared responsibility										
521	**1**	**Student government**						0.042	26, 27	26=1 & 27=3		
521_1	**1**	Election at start of year	0.017	64	v64=1							
521_2	**2**	students representing students' interests	0.017	65	v65=3							
	2	**Committees**										
522_1	**1**	Standing committees established						0.014	29 - anybody in class!	29_1==1 | 29_2==1 | 29_3==1 | 29_4==1 | 29_5==1		
522_11	1	recreation committee	0.002	67_1	67_1=1							
522_12	2	health committee	0.002	67_2	67_2=1							
522_13	3	environment and cleaning committee	0.002	67_3	67_3=1							
522_14	4	support committee for learning children	0.002	67_4	67_4=1							
522_15	5	conflict resolution committee	0.002	67_5	67_5=1							
522_16	6	other committees	0.002	67_6	67_6=1							
	2	Teacher supports student government										
522_21	1	space for committees	0.011	66_1	v66_1=1	v66_1=2	v66_1=3	0.003	28_3	28_3=1		
522_22	2	explains what it is						0.003	28_1	28_1=1		
522_23	3	motivates students to participate						0.003	28_2	28_2=1		
522_24	4	supports elections						0.003	28_4	28_4=1		
	3	Committees work properly										
522_31	1	committees have leaders	0.004	66_2	v66_2=1	v66_2=2	v66_2=3	0.003	30_1	30_1=1		
522_32	2	committees have plans	0.004	66_3	v66_3=1	v66_3=2	v66_3=3	0.003	30_2	30_2=1		
522_33	3	student assemble evaluates progress	0.004	66_4	v66_4=1	v66_4=2	v66_4=3	0.003	30_4	30_4=1		
522_34	4	put plans in practice						0.003	30_3	30_3=1		
	3	**Shared responsibilities in classroom**										
523_1	**1**	promoted by teacher	0.008	59	v59=1	v59=2	v59=3					
	2	group roles										
523_21	1	group leader	0.002	60_1	v60_1=1	v60_1=2	v60_1=3	0.002	21_1	21_1=1	21_1=2	21_1=3
523_22	2	responsible for materials	0.002	60_2	v60_2=1	v60_2=2	v60_2=3	0.002	21_2	21_2=1	21_2=2	21_2=3
523_23	3	time use	0.002	60_3	v60_3=1	v60_3=2	v60_3=3	0.002	21_3	21_3=1	21_3=2	21_3=3
523_24	4	group work presenter	0.002	60_4	v60_4=1	v60_4=2	v60_4=3	0.002	21_4	21_4=1	21_4=2	21_4=3
523_25	5	mediator	0.002	60_5	v60_5=1	v60_5=2	v60_5=3	0.002	21_5	21_5=1	21_5=2	21_5=3
523_3	**3**	suggestion box	0.008	70_2	70_2=1	70_2=2	70_2=3	0.010	31_2	31_2=1		

523_4	**4** Classroom rules decided collectively	0.008	62	v62=3	0.010	25	25=3
523_5	**5** understanding importance				0.010	22	22=1

ANNEX D: Questionnaires

(survey instruments start on next page)



Trabajo en el Aula
Implementación de Escuela Nueva
Cuestionario del Estudiante

A. DATOS DE IDENTIFICACIÓN (A rellenar por el/la entrevistador/a)

A1. Nombre del encuestador/la encuestadora: ____

A2. Número de la sede: ___.___.___

A3. Código del estudiante: ___.___.___.___.___

A4. Fecha de la entrevista: **1**. día___.___ **2**. mes___.___

¡Buenos días!

Gracias por participar en este estudio.

Antes de responder este cuestionario, lee con atención esta información y sigue las instrucciones.

- Este es un cuestionario que busca que nos cuentes sobre la organización de tu salón de clase, los textos que usas para aprender y los materiales de consulta, cómo se hacen los acuerdos de convivencia, el Gobierno Estudiantil, y aspectos de la relación con la comunidad. También buscamos saber si en tu escuela o colegio se trabaja con Escuela Nueva y cómo lo haces.
- No hay una respuesta correcta para cada pregunta, elige la opción que creas se acerca más a lo que vives en tus clases. Debes contestar de forma individual sin hablar con tus compañeros.
- Esta no es una prueba de conocimientos y no tiene una nota. Para marcar tus respuestas sólo necesitas usar un lápiz negro.
- Para responder cada pregunta, debes marcar con una X la casilla que corresponde a la opción de respuesta que mejor refleja lo que piensas o haces. Si deseas cambiar una respuesta, puedes borrar y rellenar la otra respuesta que seleccionaste.

¡GRACIAS!

B. CARACTERIZACIÓN DEL ESTUDIANTE

1. Eres	Marca con una X **a**. Niño ○ **b**. Niña ○	**2**. ¿Cuántos años tienes?	Marca con una X ○ **a**. 9 o menos ○ **d**. 12 años ○ **b**. 10 años ○ **e**. 13 años ○ **c**. 11 años ○ **f.** 14 o más

3. ¿Cuántos hermanos y hermanas tienes?	Escribe el número □

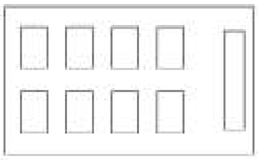
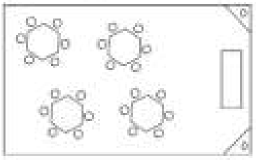
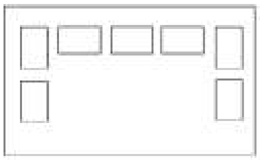
4. ¿Cuántos niños y niñas hay en tu salón?	Escribe el número □	**5**. ¿Cuántos niños y niñas hay en tu mismo grado?	Escribe el número □

6. ¿Hace cuántos años estudias en esta escuela?	Marca con una X **a.** 1 año o menos ○ **b**. 2 años ○ **c**. 3 años ○ **d**. 4 años ○ **e**. 5 años ○ **f**. 6 años o más ○

C. ORGANIZACIÓN DEL SALÓN DE CLASE

7. ¿A cuál de los siguientes gráficos se parece más la organización de tu salón de clase?

Escoge una sola respuesta

a. Todos los niños y niñas en filas y mirando al tablero	**b.** Los niños y niñas sentados en grupos la mayoría del tiempo	**c.** Los pupitres organizados en forma de U
		
○	○	○

D. TEXTOS PARA APRENDER Y GUÍAS DE APRENDIZAJE

8. ¿En tu salón de clase hay Guías de Aprendizaje Escuela Nueva?

Marca con X una sola casilla

○ **a. Sí,** hay Guías de Aprendizaje Escuela Nueva

○ **b. No**, no hay. Trabajamos solamente con otros libros**.**

Si respondiste NO, salta a la pregunta 17

9. ¿Si en tu salón de clase hay Guías de Aprendizaje Escuela Nueva con quién las compartes?

Escoge una sola respuesta

○ **a.** Con ningún compañero, el/la maestro(a) nos pide que cada uno trabaje solo.

○ **b.** Con uno o varios compañeros.

○ **c.** Las uso yo solito porque soy el único alumno en mi grado.

○ **d.** Las uso yo solito porque tenemos guías de sobra en mi salón de clases.

10. ¿Si en tu salón de clase hay Guías de Aprendizaje Escuela Nueva, para qué más las utilizas?

Escoge una sola respuesta

○ **a.** Para trabajar con mis compañeros en los pequeños grupos.

○ **b.** Como material de consulta en la biblioteca.

○ **c.** Para ponerme al día cuando no vengo a clase.

○ **d.** No las uso

11. ¿Utilizas Guías de Aprendizaje en las siguientes materias?

Para cada opción/fila, escoge una sola respuesta

	Sí	No
1. En Lenguaje		
2. En Matemáticas		
3. En Ciencias Sociales		
4. En Ciencias Naturales		
5. En Otras áreas		

12. ¿Cómo usas las Guías de Aprendizaje Escuela Nueva?

Para cada opción/fila, escoge una sola respuesta

	Sí	No
1. Copiando la guía en el cuaderno		
2. Copiando la guía en el tablero		
3. Respondiendo las actividades en el cuaderno		
4. Respondiendo las actividades en la guía		

13. Cuando desarrollas las actividades de las Guías de Aprendizaje Escuela Nueva, ¿dónde registras tu Control de Progreso?

Escoge una sola respuesta

a. En mi cuaderno ○ **b.** En el pizarrón ○ **c.** En el cuaderno del profesor ○

14. ¿Con quién registras en el cuaderno tus avances del Control de Progreso?

Para cada opción/fila escoge una sola respuesta

	Sí	No
1. Lo hago yo solo		
2. Con mis compañeros		
3. Con mi profesor		
4. Con mis padres		

15. ¿Te gusta estudiar con las Guías de Aprendizaje Escuela Nueva?

Escoge una sola respuesta

a. Sí, mucho ○ **b.** Sí, a veces ○ **c.** Casi no ○ **d.** No ○

16. Después de terminar el día de clases y llegar a tu casa, ¿cuál de las siguientes actividades realizas?

Para cada opción/fila, escoge una sola respuesta

	Sí	No
1. Copio textos de la Guía de Aprendizaje para practicar la escritura.		
2. Estudio temas nuevos que no alcance a completar en la escuela.		
3. Realizo las actividades de aplicación con mi familia o algunas personas de la comunidad.		
4. No realizo actividades en casa		

E. MATERIALES DE APOYO

17. La mayor parte del tiempo usas la Biblioteca de:

Escoge una sola respuesta

○ **a.** Tu salón de clase
○ **b.** Tu escuela o colegio
○ **c.** Otra escuela
○ **d.** Tu Municipio
○ **e.** No usas ninguna biblioteca

18. En tu salón de clase o en tu escuela, hay sitios o rincones de aprendizaje con materiales, juegos, recursos, para aprender las materias como:

Para cada opción/fila escoge una sola respuesta

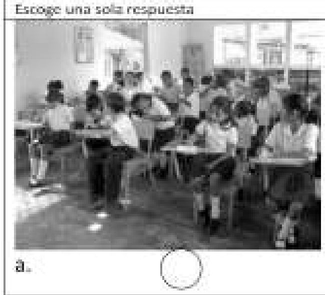
	Sí	No
1. Lenguaje		
2. Matemáticas		
3. Ciencias Sociales		
4. Ciencias Naturales		

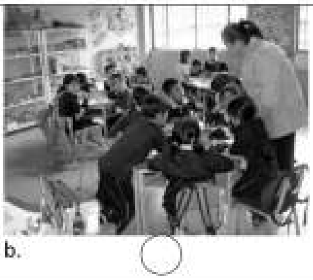
19. Cuando desarrollas las actividades de la clase, cada cuánto consultas o utilizas:				
Para cada opción/fila escoge una sola respuesta				
	a. (Casi) siempre	**b**. Muchas veces	**c.** A veces	**d.** Nunca
1. Los libros de la biblioteca.				
2. Los materiales que están en los sitios o rincones de aprendizaje de las materias				

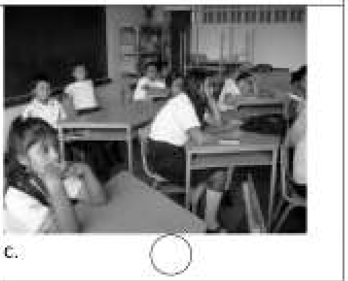
F. TRABAJO EN GRUPOS

20. Durante la clase, ¿cómo te organizas para trabajar la mayoría del tiempo?

Escoge una sola respuesta


a. ○


b. ○


c. ○

21. En tu salón de clase o grupo de trabajo hay:				
Para cada opción/fila escoge una sola respuesta				
	a. Siempre	**b.** Muchas veces	**c.** A veces	**d.** Nunca
1. Encargado de liderar el desarrollo de actividades.				
2. Encargado de traer a la mesa los libros y materiales que se van a usar.				
3. Encargado de controlar el uso del tiempo.				
4. Encargado de presentar o relatar los resultados del trabajo del grupo.				
5. Encargado de ayudar cuando tenemos diferencias.				

22. ¿Por qué es importante que los estudiantes tengan distintas responsabilidades cuando trabajan en grupos?

Escoge una sola respuesta

○**a.** Porque todos pueden colaborar y hacer mejor uso del tiempo.
○**b.** Porque el maestro da mejores calificaciones.
○**c.** Porque así uno puede olvidarse de algunas responsabilidades como cuidar los materiales y llevar el tiempo.
○**d.** En mi salón de clase o en mi grupo no es importante tener distintas responsabilidades.

23. ¿Tu maestro/a te anima a ayudar a otros estudiantes a mejor comprender la materia?

Escoge una sola respuesta

a. Sí, siempre ○ **b**. Sí, a veces ○ **c.** Nunca ○

24. Durante un día normal en la escuela, con qué pasas más tiempo:

Escoge una sola respuesta

○ **a.** Trabajando sólo o en grupos o parejas
○ **b.** Escuchando al maestro/a la maestra

G. ACUERDOS DE CONVIVENCIA

25. ¿Cómo se fijan las Normas de Convivencia en tu salón de clase?

Escoge una sola respuesta

○ **a.** Están establecidas en el Manual de Convivencia de mi escuela o colegio
○ **b.** Todos los años, el maestro las decide y las comparte con los estudiantes
○ **c.** Al inicio del año, los estudiantes y los docentes establecemos las Normas de Convivencia
○ **d.** No tenemos Normas de Convivencia

H. GOBIERNO ESTUDIANTIAL

26. ¿En tu salón de clase o en tu escuela o colegio se eligió el Gobierno Estudiantil este año o el año anterior?

Escoge una sola respuesta

○ **a.** Sí ○ **b.** No

27. El Gobierno Estudiantil es:

Escoge una sola respuesta

○ **a.** El docente al iniciar cada periodo escolar nombra un Presidente y un Vicepresidente.

○ **b.** La elección del personero.

○ **c.** La organización de los estudiantes para trabajar mejor y solucionar necesidades de la escuela o del colegio eligiendo un presidente, vicepresidente, secretario y los comités.

28. Durante este año escolar o en el año anterior tu maestro:

Para cada opción/fila escoge una sola respuesta

	Sí	**No**
1. Explicó a los estudiantes en qué consiste el Gobierno Estudiantil.		
2. Motivó a los estudiantes para que participaran en el Gobierno Estudiantil.		
3. Dio tiempo para que los estudiantes se postularan como candidatos en las elecciones del Gobierno Estudiantil.		
4. Acompañó las elecciones del Gobierno Estudiantil.		

29. ¿Eres miembro de alguno de los siguientes grupos o comités?

Para cada opción/fila escoge una sola respuesta

	Sí	**No**
1. Comité de recreación		
2. Comité de salud		
3. Comité de limpieza y medio ambiente		
4. Comité de apoyo para que otros niños aprendan		
5. Comité de resolución de conflictos		

Si respondiste NO en todas las filas de la pregunta 29,
salta a la pregunta 31

30. El grupo o comité en el que participas:

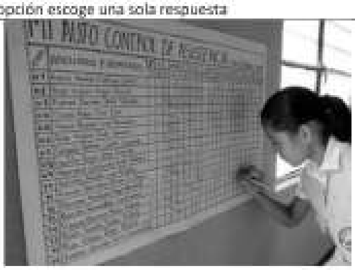
Para cada opción/fila escoge una sola respuesta

	Sí	**No**
1. Tiene un coordinador		
2. Tiene actividades programadas		
3. Desarrolla sus actividades		
4. Se reúne y evalúa el desarrollo de sus actividades		

I. INSTRUMENTOS DEL GOBIERNO ESTUDIANTIL

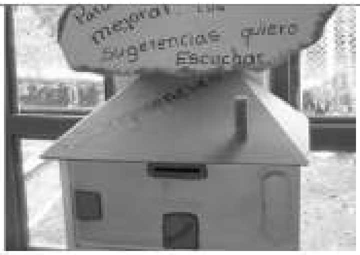
31. ¿Cuáles de los siguientes instrumentos has utilizado en este año escolar o en el año anterior?

Para cada opción escoge una sola respuesta



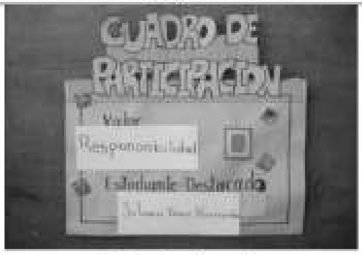
1. Auto control de asistencia

○ **Sí** ○ **No**



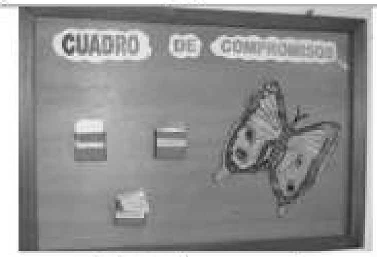
2. Buzón de sugerencias

○ **Sí** ○ **No**



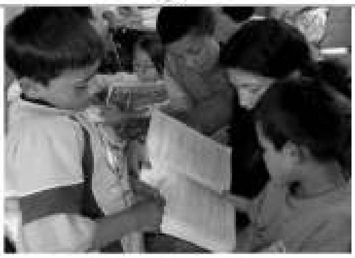
3. Cuadro de valores

○ **Sí** ○ **No**



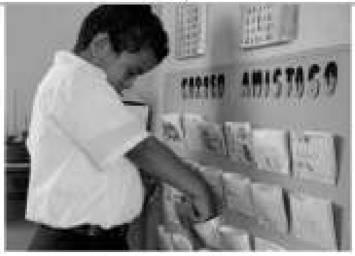
4. Buzón de compromisos

○ **Sí** ○ **No**



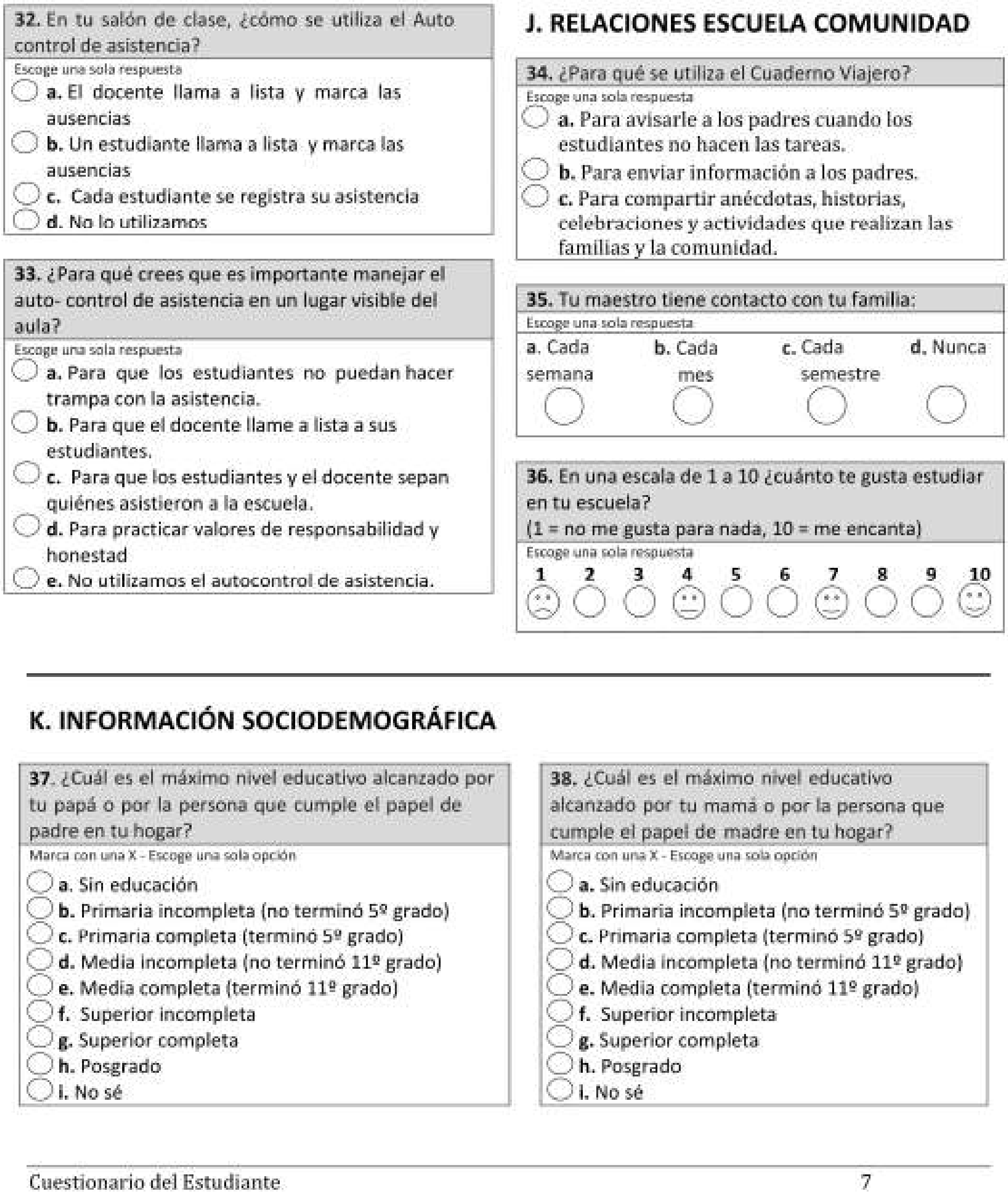
5. Cuaderno viajero

○ **Sí** ○ **No**



6. Correo amistoso

○ **Sí** ○ **No**



32. En tu salón de clase, ¿cómo se utiliza el Auto control de asistencia?

Escoge una sola respuesta

○ **a.** El docente llama a lista y marca las ausencias
○ **b.** Un estudiante llama a lista y marca las ausencias
○ **c.** Cada estudiante se registra su asistencia
○ **d.** No lo utilizamos

33. ¿Para qué crees que es importante manejar el auto- control de asistencia en un lugar visible del aula?

Escoge una sola respuesta

○ **a.** Para que los estudiantes no puedan hacer trampa con la asistencia.
○ **b.** Para que el docente llame a lista a sus estudiantes.
○ **c.** Para que los estudiantes y el docente sepan quiénes asistieron a la escuela.
○ **d.** Para practicar valores de responsabilidad y honestad
○ **e.** No utilizamos el autocontrol de asistencia.

J. RELACIONES ESCUELA COMUNIDAD

34. ¿Para qué se utiliza el Cuaderno Viajero?

Escoge una sola respuesta

○ **a.** Para avisarle a los padres cuando los estudiantes no hacen las tareas.
○ **b.** Para enviar información a los padres.
○ **c.** Para compartir anécdotas, historias, celebraciones y actividades que realizan las familias y la comunidad.

35. Tu maestro tiene contacto con tu familia:

Escoge una sola respuesta

a. Cada semana	**b.** Cada mes	**c.** Cada semestre	**d.** Nunca
○	○	○	○

36. En una escala de 1 a 10 ¿cuánto te gusta estudiar en tu escuela?
(1 = no me gusta para nada, 10 = me encanta)

Escoge una sola respuesta

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
☹	○	○	😐	○	○	🙂	○	○	☺

K. INFORMACIÓN SOCIODEMOGRÁFICA

37. ¿Cuál es el máximo nivel educativo alcanzado por tu papá o por la persona que cumple el papel de padre en tu hogar?

Marca con una X - Escoge una sola opción

○ **a.** Sin educación
○ **b.** Primaria incompleta (no terminó 5º grado)
○ **c.** Primaria completa (terminó 5º grado)
○ **d.** Media incompleta (no terminó 11º grado)
○ **e.** Media completa (terminó 11º grado)
○ **f.** Superior incompleta
○ **g.** Superior completa
○ **h.** Posgrado
○ **i.** No sé

38. ¿Cuál es el máximo nivel educativo alcanzado por tu mamá o por la persona que cumple el papel de madre en tu hogar?

Marca con una X - Escoge una sola opción

○ **a.** Sin educación
○ **b.** Primaria incompleta (no terminó 5º grado)
○ **c.** Primaria completa (terminó 5º grado)
○ **d.** Media incompleta (no terminó 11º grado)
○ **e.** Media completa (terminó 11º grado)
○ **f.** Superior incompleta
○ **g.** Superior completa
○ **h.** Posgrado
○ **i.** No sé

39. ¿De qué material está hecha la mayoría de los pisos de la vivienda en donde vives?

Escoge una sola respuesta

○ **a.** Tierra o arena
○ **b.** Cemento o gravilla
○ **c.** Tabla, tablón o madera burda
○ **d.** Baldosa, tableta, ladrillo o vinilo
○ **e.** Madera pulida, alfombra, tapete, mármol

40. ¿Con qué tipo de servicio sanitario cuenta tu hogar?

Escoge una sola respuesta

○ **a.** Inodoro conectado al alcantarillado
○ **b.** Inodoro conectado a pozo séptico
○ **c.** Inodoro sin conexión
○ **d.** Letrina
○ **e.** No tiene servicio sanitario

41. El servicio sanitario del hogar es:

Escoge una sola respuesta

○ **a.** De uso exclusivo de las personas del hogar
○ **b.** Compartido con personas de otros hogares

42. Incluido tú, ¿cuántas personas viven en tu hogar?

Escoge una sola respuesta

2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Más**
○	○	○	○	○	○	○	○	○	○

43. Contando sala y comedor, ¿cuántos cuartos o piezas tiene la casa o apartamento en que vives? (no cuentes ni cocina, ni baños, ni garaje)

Escoge una sola respuesta

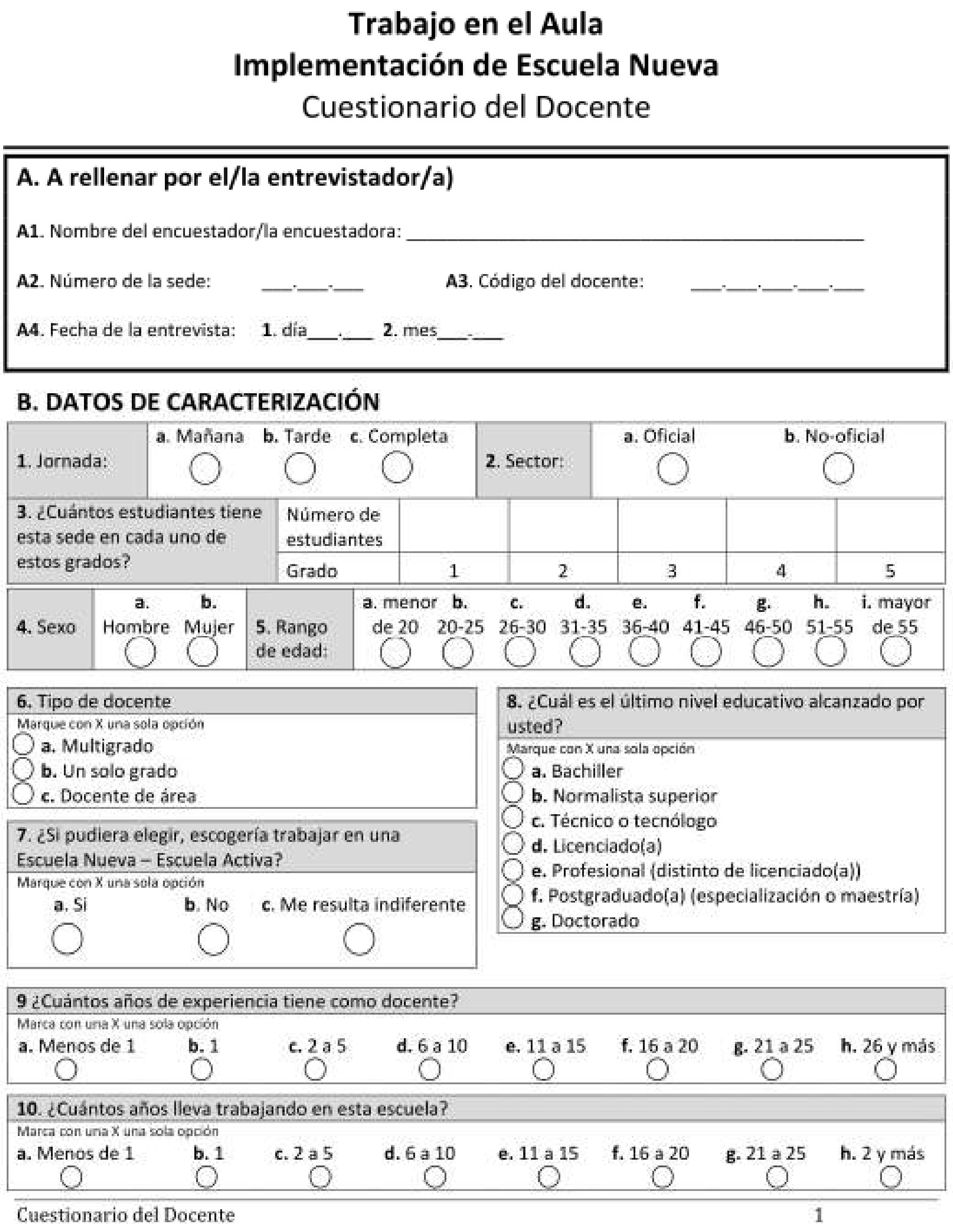
1	**2**	**3**	**4**	**5**	**Más**
○	○	○	○	○	○

44. ¿En cuántos de esos cuartos duermen las personas que viven contigo?

Escoge una sola respuesta

1	**2**	**3**	**4**	**5**	**Más**
○	○	○	○	○	○

L. COMENTARIOS



Trabajo en el Aula
Implementación de Escuela Nueva
Cuestionario del Docente

A. A rellenar por el/la entrevistador/a)

A1. Nombre del encuestador/la encuestadora: ____

A2. Número de la sede: ___.___.___ **A3**. Código del docente: ___.___.___.___.___

A4. Fecha de la entrevista: **1**. día___.___ **2**. mes___.___

B. DATOS DE CARACTERIZACIÓN

1. Jornada:	**a**. Mañana ○	**b**. Tarde ○	**c**. Completa ○	**2**. Sector:	**a**. Oficial ○	**b**. No-oficial ○

3. ¿Cuántos estudiantes tiene esta sede en cada uno de estos grados?	Número de estudiantes					
	Grado	1	2	3	4	5

4. Sexo	**a**. Hombre ○	**b**. Mujer ○	**5**. Rango de edad:	**a**. menor de 20 ○	**b**. 20-25 ○	**c**. 26-30 ○	**d**. 31-35 ○	**e**. 36-40 ○	**f**. 41-45 ○	**g**. 46-50 ○	**h**. 51-55 ○	**i**. mayor de 55 ○

6. Tipo de docente

Marque con X una sola opción

○ **a.** Multigrado
○ **b.** Un solo grado
○ **c.** Docente de área

7. ¿Si pudiera elegir, escogería trabajar en una Escuela Nueva – Escuela Activa?

Marque con X una sola opción

a. Si	**b**. No	**c**. Me resulta indiferente
○	○	○

8. ¿Cuál es el último nivel educativo alcanzado por usted?

Marque con X una sola opción

○ **a.** Bachiller
○ **b.** Normalista superior
○ **c.** Técnico o tecnólogo
○ **d.** Licenciado(a)
○ **e.** Profesional (distinto de licenciado(a))
○ **f.** Postgraduado(a) (especialización o maestría)
○ **g.** Doctorado

9 ¿Cuántos años de experiencia tiene como docente?

Marca con una X una sola opción

a. Menos de 1	**b.** 1	**c.** 2 a 5	**d.** 6 a 10	**e.** 11 a 15	**f.** 16 a 20	**g.** 21 a 25	**h.** 26 y más
○	○	○	○	○	○	○	○

10. ¿Cuántos años lleva trabajando en esta escuela?

Marca con una X una sola opción

a. Menos de 1	**b.** 1	**c.** 2 a 5	**d.** 6 a 10	**e.** 11 a 15	**f.** 16 a 20	**g.** 21 a 25	**h.** 2 y más
○	○	○	○	○	○	○	○

C. FORMACIÓN ESPECÍFICA EN ESCUELA NUEVA

11. ¿Usted es o una vez ha sido docente de Escuela Nueva – Escuela Activa?

Marca con una X una sola opción

○ **a**. Sí ○ **b**. No

Si respondió NO, salte a la pregunta 14

12. Este año escolar 2016 ¿Usted es docente de Escuela Nueva - Escuela Activa?

Marca con una X una sola opción

○ **a**. Sí ○ **b**. No

13. ¿Cuántos años de experiencia tiene como docente en Escuela Nueva?

Marca con una X una sola opción

○ **a.** Menos de 1
○ **b.** 1
○ **c.** 2 a 5
○ **d.** 6 a 10
○ **e.** 11 a 15
○ **f.** 16 a 20
○ **g.** 21 a 25
○ **h.** 25 y más

14. Después de su capacitación institucional previa al servicio, ¿usted ha asistido a un taller o curso de formación y de enseñanza continua?

Marca con una X una sola opción

○ **a**. Sí ○ **b**. No

Si respondió NO, salte a la pregunta 24

15. ¿Usted ha asistido a un **taller de iniciación** de Escuela Nueva?

Marca con una X una sola opción

○ **a**. Sí ○ **b**. No

Si respondió NO, salte a la pregunta 18

16. ¿Qué duración tuvo el taller de iniciación de Escuela Nueva?

Marca con una X una sola opción

a. 1 a 2 días	**b**. 3 a 5 días	**c**. Más de 5 días
○	○	○

17. ¿De qué institución eran los capacitadores que orientaron el taller de iniciación de Escuela Nueva?

Marca con una X una sola opción

○ **a.** Ministerio de Educación
○ **b.** Secretaría de Educación
○ **c.** Universidades
○ **d.** Fundación Escuela Nueva Volvamos a la Gente
○ **e.** Otra entidad ¿Cuál? ______

18. ¿Usted ha asistido a un taller de **uso de guías de aprendizaje** Escuela Nueva?

Marca con una X una sola opción

○ **a**. Sí ○ **b**. No

Si respondió NO, salte a la pregunta 22

19. ¿Qué duración tuvo el taller de uso de guías de aprendizaje?

Marca con una X una sola opción

a. 1 a 2 días	**b**. 3 a 5 días	**c**. Más de 5 días
○	○	○

20. ¿De qué institución eran los capacitadores que orientaron el taller de uso de guías de aprendizaje?

Marca con una X una sola opción

○ **a.** Ministerio de Educación
○ **b.** Secretaría de Educación
○ **c.** Universidades
○ **d.** Fundación Escuela Nueva Volvamos a la Gente
○ **e.** Otra entidad ¿Cuál? □

21. Cuando recibió la capacitación, ¿contaba con guías de aprendizaje Escuela Nueva?

Marca con una X una sola opción

○ **a.** Sí ○ **b.** No

22. Durante el taller (los talleres), usted tuvo la oportunidad de:

(Para cada opción/fila marque con X una sola respuesta)

	a. **Todo el tiempo**	b. **Algunas veces**	c. **Nunca**
1. Trabajar con el Manual para el docente Escuela Nueva			
2. Trabajar con las guías de aprendizaje Escuela Nueva			
3. Desarrollar actividades siguiendo la secuencia de: actividades básicas, actividades de práctica y actividades de aplicación			
4. Recibir orientación sobre la contextualización y planeación de la clase con las guías de aprendizaje Escuela Nueva			
5. Organizarse y trabajar en pequeños grupos con otros docentes			
6. Organizar y vivenciar el Gobierno del Taller y los instrumentos del Gobierno Estudiantil			
7. Organizar Rincones de Aprendizaje			
8. Organizar Biblioteca Aula			
9. Preparar estrategias de sensibilización a padres y comunidad			

23. Después de culminar el taller (los talleres), usted logró en su aula:

(Para cada opción/fila marque con X una sola respuesta)

	a. **Completa-mente**	b. **Parcial-mente**	c. **No se logró**
1. Organizar el aula			
2. Organizar una reunión con los padres de comunidad			
3. Organizar el Gobierno Estudiantil			
4. Organizar Rincones de Aprendizaje			
5. Organizar la Biblioteca Aula			
6. Organizar microcentros			

24. ¿Ha visitado alguna Escuela Nueva Demostrativa?

Marca con una X una sola opción

○ **a.** Sí ○ **b.** No

25. ¿Participa usted en la Comunidad Virtual de la Fundación Escuela Nueva?

Marca con una X una sola opción

ReNueva

a. Sí ○ **b.** No ○

26. ¿Se organizan reuniones de maestros en su comunidad educativa, con el fin de intercambio de experiencia (por ejemplo, Microcentros)?

Marca con una X una sola opción

a. Cada mes ○ **b.** Cada dos meses ○ **c.** Cada semestre ○ **d.** Nunca ○ **e.** No sé ○

27. Durante el actual o pasado año escolar, ¿ha participado en Microcentros de Escuela Nueva?

Marca con una X una sola opción

○ **a.** Sí ○ **b.** No

Si respondió NO, salte a la pregunta 29

28. ¿En qué consisten los Microcentros en los que ha participado?

Marca con una X una sola opción

○ **a.** Intercambio de conocimientos y experiencias entre docentes con eventual acompañamiento de directivos docentes.

○ **b.** Reunión de padres de familia, estudiantes y docentes para dialogar sobre el ambiente escolar.

○ **c.** Reunión entre docentes y padres de familia para la entrega de informes académicos.

○ **d.** Jornadas pedagógicas.

29. ¿Con qué frecuencia ha recibido visitas de acompañamiento de un tutor o coordinador en su escuela durante este año escolar o el año escolar anterior?

Marca con una X una sola opción

a. Dos veces o más al semestre ○ **b.** Una vez al semestre ○ **c.** Una vez al año ○ **d.** Nunca ○

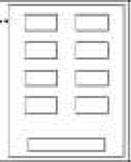
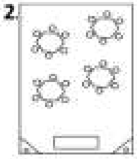
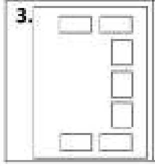
30. ¿Esta sede pone en práctica un mecanismo de promoción flexible?

Marca con una X una sola opción

a. Sí ○ **b.** No ○ **c.** No sé ○

31. ¿A cuál de los siguientes gráficos se parece la organización de su salón de clases?

Para cada opción/fila marque con X una sola respuesta

	a. Siempre	**b.** Muchas veces	**c.** A veces	**d.** Nunca
1. 				
2. 				
3. 				

32. ¿Con qué frecuencia los estudiantes se encuentran trabajando...

Para cada opción/fila marque con X una sola respuesta

	a. Siempre	**b.** Muchas veces	**c.** A veces	**d.** Nunca
1. Individual-mente?				
2. En parejas?				
3. En grupos?				

33 ¿Usted anima a sus estudiantes a ayudarse y apoyarse mutuamente con el trabajo?

Marca con una X una sola opción

a. Siempre ○ **b.** Muchas veces ○ **c.** Algunas veces ○ **d.** Nunca ○

D. MATERIALES ESCUELA NUEVA

34. ¿Usted dispone del Manual del Docente Escuela Nueva?

Marca con una X una sola opción

○ **a**. Sí ○ **b**. No

Si respondió NO,
salte a la pregunta 37

35. Durante el año escolar, ¿con qué frecuencia utiliza usted el Manuel del Docente Escuela Nueva?

Marca con una X una sola opción

a. Siempre	**b**. Muchas veces	**c.** Algunas veces	**d**. Nunca
○	○	○	○

36. ¿Quién es el autor de su Manual del Docente Escuela Nueva?

Marca con una X una sola opción

○ **a.** Ministerio de Educación
○ **b.** Secretaria de Educación
○ **c.** Universidades
○ **d.** Fundación Escuela Nueva Volvamos a la Gente
○ **e.** Otra entidad ¿Cuál?

37. Actualmente la sede a la que usted pertenece tiene Guías de Aprendizaje Escuela Nueva?

Marca con una X una sola opción

○ **a**. Sí ○ **b**. No

Si respondió NO,
salte a la pregunta 48

38. ¿En cuáles de las siguientes áreas tiene Guías de Aprendizaje Escuela Nueva?

Para cada opción/fila marque con X una sola respuesta

	a. **Sí**	b. **No**
1. En Lenguaje		
2. En Matemáticas		
3. En Ciencias Sociales		
4. En Ciencias Naturales		
5. En otras áreas		

39. ¿Quién es el autor de las Guías de Aprendizaje Escuela Nueva?

Marca con una X una sola opción

○ **a.** Ministerio de Educación
○ **b.** Secretaría de Educación
○ **c.** Universidades
○ **d.** Fundación Escuela Nueva Volvamos a la Gente
○ **e.** Otra entidad:

40. ¿Con qué frecuencia usa usted las Guías de Aprendizaje Escuela Nueva?

Para cada opción/fila marque con X una sola respuesta

	a. Siempre	**b.** Muchas veces	**c.** A veces	**d.** Nunca
1. En Lenguaje				
2. En Matemáticas				
3. En Ciencias Sociales				
4. En Ciencias Naturales				
5. En otras áreas				

41. ¿Cómo planea la clase para el uso de las Guías de Aprendizaje Escuela Nueva?

Marca con una X una sola opción

○ **a.** Contextualizando la guía de aprendizaje al medio.
○ **b.** Teniendo las guías no hay necesidad de planear las clases.

42. Dadas las restricciones de tiempo, ¿cuáles de las siguientes actividades de las Guías de Aprendizaje Escuela Nueva usted planea completar totalmente?

Para cada opción/fila marque con X una sola respuesta

	a. Siempre	**b.** Muchas veces	**c.** A veces	**d.** Nunca
1. Actividades básicas				
2. Actividades de práctica				
3. Actividades de aplicación				

43. Durante la jornada escolar los estudiantes:

Marca con una X una sola opción

○ **a.** Trabajan con textos escolares.

○ **b.** Trabajan únicamente con las Guías de Aprendizaje Escuela Nueva.

○ **c.** Trabajan con las Guías de Aprendizaje Escuela Nueva y consultan otros textos escolares como complemento.

○ **d.** Trabajan con textos escolares en su mesa de trabajo y consultan las Guías de Aprendizaje Escuela Nueva como complemento.

44. ¿Cómo desarrolla las Guías de Aprendizaje Escuela Nueva durante la clase?

Marca con una X una sola opción

○ **a.** Haciendo que los estudiantes transcriban las actividades de la guía en el tablero.

○ **b.** Haciendo que los estudiantes transcriban las actividades de la guía en el cuaderno.

○ **c.** Los niños comprenden las preguntas y actividades y las desarrollan en el cuaderno.

45. Con el uso de las Guías de Aprendizaje Escuela Nueva en su clase, usted logra:

Marca con una X una sola opción

○ **a.** Fomentar el aprendizaje activo y la promoción flexible.

○ **b.** Tener material de apoyo para los estudiantes que facilita la escritura en el cuaderno.

○ **c.** Mejorar el material de consulta en la biblioteca.

○ **d.** Mantener a los estudiantes trabajando solos en sus puestos.

46 ¿El uso de las Guías de Aprendizaje Escuela Nueva en su clase facilita su labor como docente?

Marca con una X una sola opción

a. Siempre	**b.** Muchas veces	**c.** Algunas veces	d. Nunca
○	○	○	○

47 ¿Con qué frecuencia usted le hace adaptaciones o modificaciones a las Guías?

Marca con una X una sola opción

a. Siempre	**b.** Muchas veces	**c.** Algunas veces	d. Nunca
○	○	○	○

E. BIBLIOTECA AULA

48. ¿Sus estudiantes tienen acceso a una biblioteca?

Marca con una X una sola opción

○ **a.** Sí ○ **b.** No

Si respondió NO, salte a la pregunta 52

49. ¿Dónde queda la Biblioteca a la que tienen acceso?

Para cada opción/fila marque con X una sola respuesta

	a. **Sí**	b. **No**
1. En el aula de clase		
2. Dentro de la escuela		
3. En otra escuela u otro edificio		

50 ¿Qué tan frecuente es el uso de la Biblioteca por parte de los estudiantes?

Marca con una X una sola opción

a. Siempre	**b.** Muchas veces	**c.** Algunas veces	**d.** Nunca
○	○	○	○

51. ¿Qué contiene la biblioteca?

Para cada opción/fila marque con X una sola respuesta

	a. **Sí**	b. **No**
1. Guías de Aprendizaje Escuela Nueva		
2. Textos escolares		
3. Literatura		
4. Textos para consulta e investigación		
5. Otras		

F. RINCONES DE APRENDIZAJE

52. ¿Sus estudiantes tienen acceso a Rincones de Aprendizaje o Centros de Recursos?

Marca con una X una sola opción

○ **a**. Sí ○ **b**. No

Si respondió NO, salte a la pregunta 58

53. ¿En dónde están organizados los Rincones de Aprendizaje o Centros de Recursos?

Para cada opción/fila marque con X una sola respuesta

	a. **Sí**	b. **No**
1. En el aula de clase		
2. Dentro de la escuela		
3. En otra escuela		

54. ¿Tiene Rincones de Aprendizaje en las siguientes áreas?

Para cada opción/fila marque con X una sola respuesta

	a. **Sí**	b. **No**
1. En Lenguaje		
2. En Matemáticas		
3. En Ciencias Sociales		
4. En Ciencias Naturales		
5. En otras áreas		

55. ¿Con qué tipo de material cuentan los estudiantes en los Rincones de Aprendizaje?

Para cada opción/fila marque con X una sola respuesta

	a. **Sí**	b. **No**
1. Material experimental		
2. Material impreso		
3. Material observación		
4. Material manipulación		
5. Material producido por los estudiantes.		
6. Material didáctico comercial		

56. ¿Cuándo se organizan los Rincones de Aprendizaje?

Marca con una X una sola opción

○ **a.** Los organizo al principio del año antes de que lleguen los estudiantes aprovechando mejor el tiempo.

○ **b.** Los organizo continuamente de acuerdo con la planeación de las guías de aprendizaje con la participación de estudiantes, padres de familia y docentes.

○ **c.** Los organizo al final del año para poder exhibir los trabajos de los niños.

57. ¿Quién organiza los Rincones de Aprendizaje?

Marca con una X una sola opción

○ **a.** Los padres de familia con el material que se reúne en los microcentros.

○ **b.** Yo junto con los estudiantes de la clase.

○ **c.** Yo solo en las jornadas pedagógicas.

○ **d.** Los estudiantes solos en las jornadas pedagógicas

G. ROLES DENTRO DEL GRUPO

58. ¿Cuál es su rol en clase respecto al aprendizaje de los estudiantes?

Marca con una X una sola opción

○ **a.** Presento e material de la lección, y dirijo y controlo las actividades que desarrollan los estudiantes en la clase. Todo lo que hacen los estudiantes responde a mis indicaciones.

○ **b.** Acompaño y retroalimento el trabajo de los estudiantes en la clase.

○ **c.** Los estudiantes trabajan en grupos orientados por mí y/o por las guías de aprendizaje.

59. ¿Usted promueve en cada uno de los grupos de trabajo roles y responsabilidades individuales?

Marca con una X una sola opción

a. Siempre	**b.** Muchas veces	**c.** Algunas veces	**d.** Nunca
○	○	○	○

60. En los grupos o mesas de trabajo de sus estudiantes hay:

(Para cada opción/fila marque con X una sola respuesta)

	a. Siempre	**b.** Muchas veces	**c.** A veces	**d.** Nunca
1. Monitores o ayudantes encargados de liderar el grupo, impulsar y coordinar el desarrollo de las actividades.				
2. Un encargado de llevar a la mesa materiales como guías de aprendizaje de Escuela Nueva, materiales del centro de recursos, de la biblioteca y retornarlos a su sitio.				
3. Un encargado de acompañar y alertar a los compañeros sobre los horarios acordados y el uso del tiempo.				
4. Un relator encargado de presentar los resultados del trabajo del grupo.				
5. Un líder que ayuda a mediar en los desacuerdos y recurre al comité pertinente o al maestro si es necesario para darle solución.				

61. En su escuela, ¿quién normalmente toma las decisiones administrativas?

Marca con una X una sola opción

○ **a.** El director de la escuela

○ **b.** El director después de consultar a los docentes

○ **c.** Toda la comunidad educativa (docentes, padres, y el director)

H. ACUERDOS DE CONVIVENCIA

62. ¿Cómo promueve la construcción de los acuerdos de convivencia en su salón de clase o escuela?

Marca con una X una sola opción

○ **a.** No es necesario, ya están establecidos en el Manual de Convivencia de la escuela.

○ **b.** Todos los años, los maestros los decidimos y los compartimos con los estudiantes.

○ **c.** Al iniciar el año escolar, los estudiantes y los docentes establecemos los acuerdos de convivencia.

○ **d.** No tenemos acuerdos de Convivencia

I. GOBIERNO ESTUDIANTIL

63. ¿En su escuela se elige un Gobierno Estudiantil?

Marca con una X una sola opción

○ **a**. Sí ○ **b**. No

Si respondió NO, salte a la pregunta 67

64. ¿En qué momento del año escolar se elige el Gobierno Estudiantil?

Marca con una X una sola opción

○ **a.** Al iniciar el año escolar

○ **b.** En cualquier momento del año escolar

○ **c.** Al finalizar el año escolar

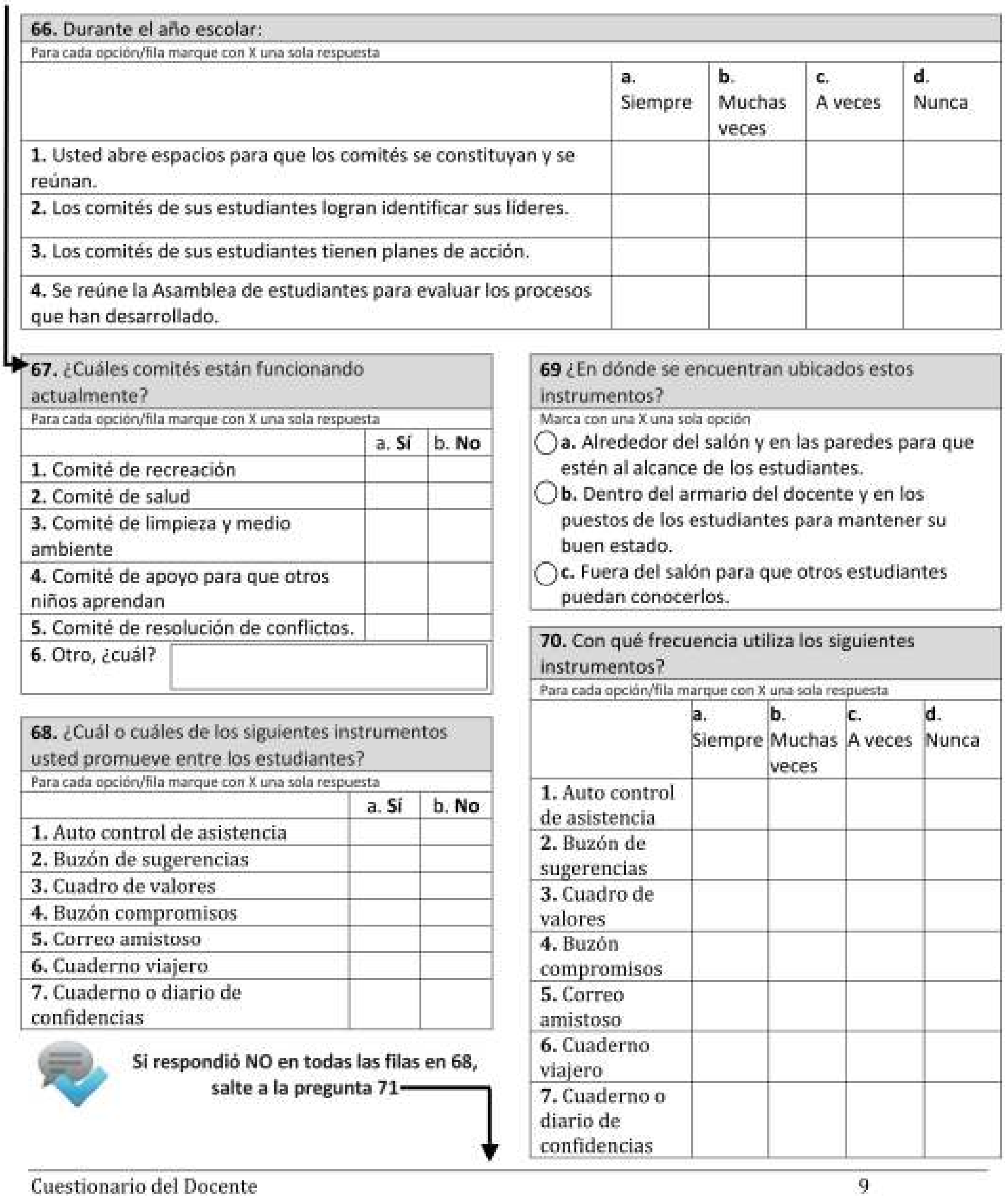
65. ¿En qué consiste el Gobierno Estudiantil?

Marca con una X una sola opción

○ **a.** El Gobierno es un órgano de participación democrática de todos los estamentos de la comunidad educativa para participar en la dirección de las instituciones de educación. Está compuesta por el Consejo Directivo, Consejo Académico y el rector.

○ **b.** Es el órgano encargado de velar por los derechos y deberes de una comunidad estudiantil y es liderado desde los estudiantes por el Personero.

○ **c.** El gobierno estudiantil se crea por y para los estudiantes, permite que estos participen en la organización de las actividades del día a día y está integrado por la asamblea general, donde están todos los estudiantes y la junta directiva, que está conformada a su vez por el presidente, el vicepresidente, el secretario y los comités.



66. Durante el año escolar:

Para cada opción/fila marque con X una sola respuesta

	a. Siempre	b. Muchas veces	c. A veces	d. Nunca
1. Usted abre espacios para que los comités se constituyan y se reúnan.				
2. Los comités de sus estudiantes logran identificar sus líderes.				
3. Los comités de sus estudiantes tienen planes de acción.				
4. Se reúne la Asamblea de estudiantes para evaluar los procesos que han desarrollado.				

67. ¿Cuáles comités están funcionando actualmente?

Para cada opción/fila marque con X una sola respuesta

	a. **Sí**	b. **No**
1. Comité de recreación		
2. Comité de salud		
3. Comité de limpieza y medio ambiente		
4. Comité de apoyo para que otros niños aprendan		
5. Comité de resolución de conflictos.		
6. Otro, ¿cuál?		

68. ¿Cuál o cuáles de los siguientes instrumentos usted promueve entre los estudiantes?

Para cada opción/fila marque con X una sola respuesta

	a. **Sí**	b. **No**
1. Auto control de asistencia		
2. Buzón de sugerencias		
3. Cuadro de valores		
4. Buzón compromisos		
5. Correo amistoso		
6. Cuaderno viajero		
7. Cuaderno o diario de confidencias		

Si respondió NO en todas las filas en 68, salte a la pregunta 71

69 ¿En dónde se encuentran ubicados estos instrumentos?

Marca con una X una sola opción

○ **a.** Alrededor del salón y en las paredes para que estén al alcance de los estudiantes.

○ **b.** Dentro del armario del docente y en los puestos de los estudiantes para mantener su buen estado.

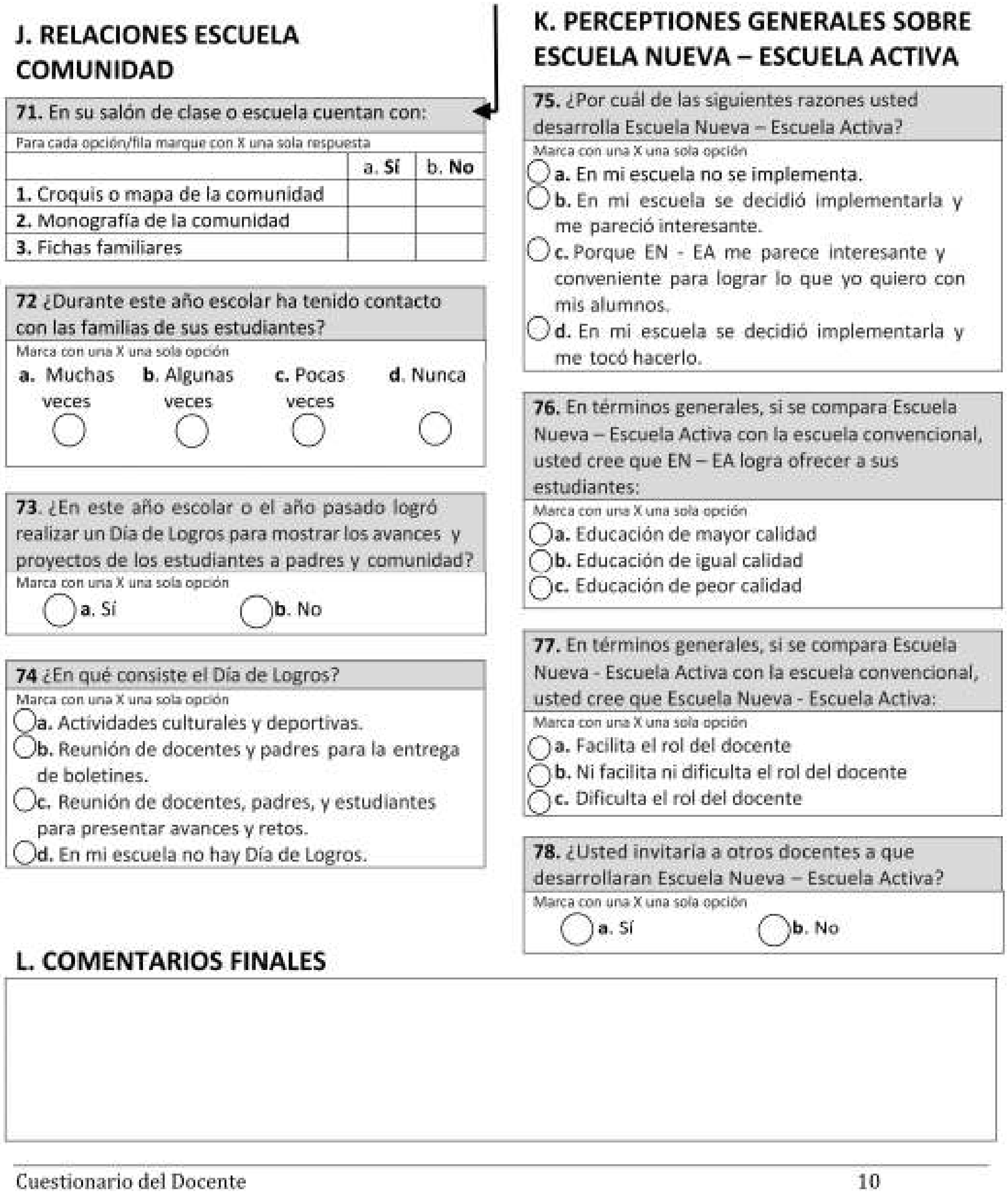
○ **c.** Fuera del salón para que otros estudiantes puedan conocerlos.

70. Con qué frecuencia utiliza los siguientes instrumentos?

Para cada opción/fila marque con X una sola respuesta

	a. Siempre	b. Muchas veces	c. A veces	d. Nunca
1. Auto control de asistencia				
2. Buzón de sugerencias				
3. Cuadro de valores				
4. Buzón compromisos				
5. Correo amistoso				
6. Cuaderno viajero				
7. Cuaderno o diario de confidencias				

Cuestionario del Docente 9



J. RELACIONES ESCUELA COMUNIDAD

71. En su salón de clase o escuela cuentan con:

Para cada opción/fila marque con X una sola respuesta

	a. **Sí**	b. **No**
1. Croquis o mapa de la comunidad		
2. Monografía de la comunidad		
3. Fichas familiares		

72 ¿Durante este año escolar ha tenido contacto con las familias de sus estudiantes?

Marca con una X una sola opción

a. Muchas veces ○ **b.** Algunas veces ○ **c.** Pocas veces ○ **d.** Nunca ○

73. ¿En este año escolar o el año pasado logró realizar un Día de Logros para mostrar los avances y proyectos de los estudiantes a padres y comunidad?

Marca con una X una sola opción

○ **a.** Sí ○ **b.** No

74 ¿En qué consiste el Día de Logros?

Marca con una X una sola opción

○ **a.** Actividades culturales y deportivas.

○ **b.** Reunión de docentes y padres para la entrega de boletines.

○ **c.** Reunión de docentes, padres, y estudiantes para presentar avances y retos.

○ **d.** En mi escuela no hay Día de Logros.

K. PERCEPTIONES GENERALES SOBRE ESCUELA NUEVA – ESCUELA ACTIVA

75. ¿Por cuál de las siguientes razones usted desarrolla Escuela Nueva – Escuela Activa?

Marca con una X una sola opción

○ **a.** En mi escuela no se implementa.

○ **b.** En mi escuela se decidió implementarla y me pareció interesante.

○ **c.** Porque EN - EA me parece interesante y conveniente para lograr lo que yo quiero con mis alumnos.

○ **d.** En mi escuela se decidió implementarla y me tocó hacerlo.

76. En términos generales, si se compara Escuela Nueva – Escuela Activa con la escuela convencional, usted cree que EN – EA logra ofrecer a sus estudiantes:

Marca con una X una sola opción

○ **a.** Educación de mayor calidad

○ **b.** Educación de igual calidad

○ **c.** Educación de peor calidad

77. En términos generales, si se compara Escuela Nueva - Escuela Activa con la escuela convencional, usted cree que Escuela Nueva - Escuela Activa:

Marca con una X una sola opción

○ **a.** Facilita el rol del docente

○ **b.** Ni facilita ni dificulta el rol del docente

○ **c.** Dificulta el rol del docente

78. ¿Usted invitaría a otros docentes a que desarrollaran Escuela Nueva – Escuela Activa?

Marca con una X una sola opción

○ **a.** Sí ○ **b.** No

L. COMENTARIOS FINALES

References

Anderson, Joan B. 2005. “Improving Latin America’s School Quality: Which Special Interventions Work?” *Comparative Education Review* 49 (2): 205–29. doi:10.1086/428720.

Atención al Ciudadano del ICFES. 2016. “Respuesta: Información Sobre La Ponderación Muestral Utilizada En Las Pruebas Saber 3°, 5° Y 9° de 2013,” April 5.

Ayala Garcia, Jhorland, Shirly Marrugo Llorente, and Bernardo Saray Ricardo. 2011. “Antecedentes Familiares Y Rendimiento Academico En Los Colegios Oficiales de Cartagena. (Family Background and Academic Performance in Cartagena Public Schools. With English Summary.).” *Economia Y Region* 5 (2): 43–85. doi:http://publicaciones.unitecnologica.edu.co/index.php/revista-economia-region/issue/archive.

Baessa, Yetilú de, Ray Chesterfield, and Tanya Ramos. 2006. “Active Learning and Democratic Behavior in Guatemalan Rural Primary Schools.” Washington, D.C. http://www.escuelanueva.org/portal/images/pdf/monitoreo/12.Baessa2006.pdf.

Baron, Juan D. 2012. “Diferencias En Las Caracteristicas de Los Estudiantes Y La Brecha de Rendimiento Academico Entre Barranquilla Y Bogota: Una Descomposicion Semiparametrica. (Differences in Student Characteristics and the Academic Achievement Gap between Barranquilla and Bogota: A Semi-Parametric Decomposition Approach.).” *Ensayos Sobre Politica Economica* 30 (68): 164–215. doi:http://www.banrep.gov.co/es/serie-de-publicaciones/revista-ensayos-sobre-pol-tica-econ-mica-espe.

Becker, Gary S. 1965. “A Theory of the Allocation of Time.” *The Economic Journal* 75 (299): 493–517. doi:10.2307/2228949.

Behrman, Jere R., and Nancy Birdsall. 1983. “The Quality of Schooling: Quantity Alone Is Misleading.” *American Economic Review* 73 (5): 928–46.

Benveniste, Luis A., and Patrick J. McEwan. 2000. “Constraints To Implementing Educational Innovations: The Case of Multigrade Schools.” *International Review of Education/Internationale Zeitschrift Fuer Erziehungswissenschaft/Revue Internationale de l’Education* 46 (1): 31–48.

Bonilla Mejia, Leonardo, and Luis Armando Galvis. 2012. “Profesionalizacion Docente Y Calidad de La Educacion Escolar En Colombia. (Teacher Training and the Quality of School Education in Colombia).” *Ensayos Sobre Politica Economica* 30 (68): 114–63. doi:http://www.banrep.gov.co/es/serie-de-publicaciones/revista-ensayos-sobre-pol-tica-econ-mica-espe.

Brown, K.G., and Martin A.B. 1989. “Student Achievement in Multigrade and Single Grade Classes.” *Education Canada* 29 (2): 10–13.

Caballero Rojas, Diana. 2009. “Investigaciones Sobre Escuela Nueva.”

Cañette, Isabel, and Yulia Marchenko. 2017. “Stata FAQ: Combining Results Other than Coefficients in E(b) with Multiply Imputed Data.” Accessed January 20.

http://www.stata.com/support/faqs/statistics/combine-results-with-multiply-imputed-data/.

Carcamo Vergara, Carolina, and Jose Antonio Mola Avila. 2012. “Diferencias Por Sexo En El Desempeno Academico En Colombia: Un Analisis Regional. (Differences in the Academic Performance by Sex in Colombia: A Regional Analysis.).” *Economia Y Region* 6 (1): 133–69. doi:http://publicaciones.unitecnologica.edu.co/index.php/revista-economia-region/issue/archive.

Carle, Adam C. 2009. “Fitting Multilevel Models in Complex Survey Data with Design Weights: Recommendations.” *BMC Medical Research Methodology* 9: 49. doi:10.1186/1471-2288-9-49.

Carstens, Ralph, and Dirk Hastedt. 2010. “The Effect of Not Using Plausible Values When They Should Be: An Illustration Using TIMSS 2007 Grade 8 Mathematics Data. Draft.” IEA Data Processing and Research Center.

Casassus, Juan, Sandra Cusato, Juan Enrique Froemel, and Juan Carlos Palafox. 2000. “Primer Estudio Internacional Comparativo Sobre Lenguaje, Matemática Y Factores Asociados, Para Alumnos Del Tercer Y Cuarto Grado de La Educación Básica.” Santiago de Chile: UNESCO Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación. http://www.escuelanueva.org/portal/images/pdf/monitoreo/7.Llece2000.pdf.

Cepeda-Cuervo, Edilberto, and Vicente Núñez-Antón. 2013. “Spatial Double Generalized Beta Regression Models: Extensions and Application to Study Quality of Education in Colombia.” *Journal of Educational and Behavioral Statistics* 38 (6): 604–28.

Chaux, Enrique. 2009. “Citizenship Competencies in the Midst of a Violent Political Conflict: The Colombian Educational Response.” *Harvard Educational Review* 79 (1): 84–93,167.

Chaux, Enrique, and Ana M. Velásquez. 2009. “Peace Education in Colombia: The Promise of Citizenship Competencies.” In *Colombia: Building Peace in a Time of War*, edited by Virginia Marie Bouvier, 159–72. Washington, D.C.: United States Institute of Peace.

Chesterfield, Ray. 1994. “Indicators of Democratic Behaviour in Nueva Escuela Unitaria (Neu) Schools.” Academy for Educational Development.

Citizenship Advisory Group. 1998. “Education for Citizenship and the Teaching of Democracy in Schools. The Crick Report.” London, UK: Qualifications and Curriculum Authority.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Colbert, Vicky. 2009. “Improving Education Quality and Access in Colombia through Innovation and Participation: The Escuela Nueva Model.” *Journal of Education for International Development* 3 (3).

———. 2015. “Escuela Nueva: Una Contribución a La Calidad Y a La Equidad En La Educación Para El Siglo XXI.” In *Equidad: Perspectivas Para Colombia*, by Fundación para la Educación y Desarrollo Social (FES), 206–33. Santiago de Cali, Colombia: Fundación para la Educación y Desarrollo Social (FES).

DANE. 2016. “Cuentas Anuales Departamentales - Colombia. Producto Interno Bruto (PIB) Definitivo Y 2014 Provisiona.” Bogotá, Colombia: Departamento Administrativo Nacional de Estadística.

Davier, Matthias von, Eugenio Gonzales, and Robert J. Mislevy. 2009. “What Are Plausible Values and Why Are They Useful?” In *Issues and Methodologies in Large-Scale Assessments*, edited by Matthias von Davier and Dirk Hastedt, 2:9–36. IERI Monograph Series. IEA-ETS Research Institute.

Davies, Lynn, Christopher Williams, Hiromi Yamashita, and Aubrey Ko-Man Hing. 2006. "Inspring Schools: Impact and Outcomes. Taking Up the Challenge of Pupil Participation." London, UK: Carnegie Young People Initiative and Esmée Fairbairn Foundation.

Dewey, John. 1916. *Democracy and Education - An Introduction to the Philosophy of Education*. Macmillan.

Diazgranados, Silvia, and James Noonan. 2015. "The Relationship of Safe and Participatory School Environments and Supportive Attitudes toward Violence: Evidence from the Colombian Saber Test of Citizenship Competencies." *Education, Citizenship and Social Justice* 10 (1): 79–94.

DNP. 2014. "Evaluación Del Desempeño Integral de Los Municipios Y Distritos, Vigencia 2013." Bogotá, Colombia: Departamento Nacional de Planeación.

Federación Colombiana de Municipios. 2016. "Indicadores Municipales." https://www.fcm.org.co/NuestrosProyectos/Paginas/Indicadores-Municipales.aspx.

Forero-Pineda, Clemente, Daniel Escobar-Rodriguéz, and Danielken Molina. 2006. "Escuela Nueva's Impact on the Peaceful Social Interaction of Children in Colombia." In *Education for All and Multigrade Teaching: Challenges and Opportunities*, edited by A.W. Little, 265–300. Springer.

Freire, Paulo. 1970. *Pedagogy of the Oppressed*. Continuum.

Gaviria, Alejandro, and Jorge Hugo Barrientos Marín. 2001. "Determinantes de La Calidad de La Educación En Colombia." 002301. Archivos de Economía. Bogotá, Colombia: Departamento Nacional de Planeación. http://ideas.repec.org/p/col/000118/002301.html.

Glewwe, Paul W., Eric A. Hanushek, Sarah D. Humpage, and Renato Ravina. 2011. "School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010."

Grilli, Leonardo, and Carla Rampichini. 2012. "Selection Bias in Linear Mixed Models." *METRON* 68 (3): 309–29. doi:10.1007/BF03263542.

Harel, O. 2009. "The Estimation of R2 and Adjusted R2 in Incomplete Data Sets Using Multiple Imputation." *Journal of Applied Statistics* 36 (10): 1109–18.

Herz, Barbara, and Gene B Sperling. 2004. "What Works in Girls' Education: Evidence and Policies from the Developing World." New York, NY: Council on Foreign Relations.

Hill, Carolyn J., Howard S Boom, Alison Rebeck Black, and Mark W. Lipsey. 2007. "Empirical Benchmarks for Interpreting Effect Sizes in Research." MDRC.

Hincapié, Diana. 2014. "Essays on Education Policy and Student Achievement in Colombia." Dissertation, Washington, D.C.: George Washington University.

Holdsworth, Roger. 2000. "Schools That Create Real Roles of Value for Young People." *Prospects* 30 (3): 349–62. doi:10.1007/BF02754058.

Huang, Francis L., and Tonya R. Moon. 2009. "Is Experience the Best Teacher? A Multilevel Analysis of Teacher Characteristics and Student Achievement in Low Performing Schools." *Educational Assessment, Evaluation and Accountability* 21 (3): 209–34.

ICFES. 2010. "Pruebas SABER 5o. Y 9o. Database." Instituto Colombiano para el Fomento de la Educación Superior.

———. 2011. "SABER 5o. Y 9o. 2009: Informe de Resultados de La Aplicación a Estudiantes de Algunos Modelos Educativos." Bogotá, Colombia: ICFES.

———. 2015a. "Establecimientos Educativos: Guía de Interpretación Y Uso de Resultados de Las Pruebas SABER 3, $^{o}5^{o}$ Y 9^{o}." Bogotá, Colombia: ICFES.

———. 2015b. “Información de La Prueba Saber 3°, 5° Y 9°.” September 13. http://www.icfes.gov.co/instituciones-educativas2/pruebas-saber-3-5-y-9/informacion-de-la-prueba-saber3579.

———. 2016a. “Consulta de Resultados Pruebas SABER 3° 5° 9°.” http://www2.icfesinteractivo.gov.co/ReportesSaber359/.

———. 2016b. “SABER 3°, 5° Y 9°: Resultados Nacionales 2009 - 2014.” Bogotá, Colombia: ICFES.

Iregui B, Ana María, Ligia Melo B., and Jorge Ramos F. 2006. “Evaluación Y Análisis de Eficiencia de La Educación En Colombia.” Bogotá, Colombia: Banco de la Republica.

Jiménez, Manuela, Enrique Chaux, D Andrade, and A Bustamante. 2008. “Socio-Emotional Competencies in Violent Contexts: Evaluation of the Multicomponent Program Aulas En Paz (Classrooms in Peace).” In . Budapest, Hungary.

Jola S., Andres Fernando. 2011. “Determinantes de La Calidad de La Educacion Media En Colombia: Un Analisis de Los Resultados PISA 2006 Y Del Plan Sectorial ‘Revolucion Educativa’.” *Coyuntura Economica: Investigacion Economica Y Social* 41 (1): 25–61. doi:http://www.fedesarrollo.org.co/publicaciones/publicaciones-periodicas/coyuntura-economica/ediciones-anteriores/.

Juárez and Associates. 2003. “The Effects of Active Learning Programs in Multigrade Schools on Girls’ Persistence in and Completion of Primary School in Developing Countries.” Girls’ Education Monitoring System. USAID.

Kim, Jee-Seon, Carolyn J. Anderson, and Bryan Keller. 2013. “Multilevel Analysis of Assessment Data.” In *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, edited by Leslie Rutkowski, Matthias von Davier, and David Rutkowski. Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences. CRC Press Taylor & Francis Group.

Kline, Rachel. 2002. “A Model for Improving Rural Schools: Escuela Nueva in Colombia and Guatemala.” *Current Issues in Comparative Education* 2 (2): 170–81.

Kolenikov, Stas. 2010. “Mean and Standard Deviation of Multiply-Imputed Values?” *Statalist*. http://www.stata.com/statalist/archive/2010-09/msg01289.html.

Kremer, Michael, and Alaka Holla. 2009. “Improving Education in the Developing World: What Have We Learned from Randomized Evaluations?” *Annual Review of Economics* 1 (1): 513–45.

Land, Susan M., and Michael J. Hannafin. 2000. “Student-Centered Learning Environments.” In *Theoretical Foundations of Learning Environments*, edited by David Jonassen and Susan M. Land. Mahwah, N.J.: L. Erlbaum Associates.

Lipsey, Mark W. 1990. *Design Sensitivity: Statistical Power for Experimental Research*. Newbury Park, Calif: SAGE Publications.

Loera, A., and Noel F. McGinn. 1992. “La Repitencia En La Escuela Primaria Colombiana: Resultados de Uno Exploración Sobre Los Factores Asociados a La Repetencia Y Las Políticas de Promoción.” Education Development Discussion Paper. Cambridge, Mass: Harvard Institute for International Development.

Mager, Ursula, and Peter Nowak. 2012. “Effects of Student Participation in Decision Making at School. A Systematic Review and Synthesis of Empirical Research.” *Educational Research Review* 7 (1): 38–61. doi:10.1016/j.edurev.2011.11.001.

Manzano Lopez, Dennys Jazmin, and Jorge Raul Ramirez Zambrano. 2012. “Interrelacion Entre La Desercion Escolar Y Las Condiciones Socioeconomicas de Las Familias: El Caso de La Cuidad de Cucuta (Colombia). (Interrelation between the School Dropouts and Socioeconomic Conditions Families: The Case of Cucuta [Colombia].).” *Revista de*

Economia Del Caribe 10 (July): 203–32. doi:http://rcientificas.uninorte.edu.co/index.php/economia/issue/archive.

Marchant, Gregory J. 2015. "How Plausible Is Using Averaged NAEP Values to Examine Student Achievement?" *Comphrehensive Psychology* 4 (1).

Marlowe, Bruce A., and Marilyn L. Page. 2005. *Creating and Sustaining the Constructivist Classroom*. Corwin Press.

Marzano, Robert J., Barbara B. Gaddy, and Ceri Dean. n.d. "What Works in Classroom Instruction." Aurora, CO: Mid-Continent Regional Educational Laboratory.

Mason, DeWayne A., and Robert B. Burns. 1997. "Reassessing the Effects of Combination Classes." *Educational Research and Evaluation* 3 (1): 1–53. doi:10.1080/1380361970030101.

McArdle, John J., Thomas S. Paskus, and Steven M. Boker. 2013. "A Multilevel Multivariate Analysis of Academic Performances in College Based on NCAA Student-Athletes." *Multivariate Behavioral Research* 48 (1): 57–95.

McCoach, D. Betsy, and Anne C. Black. 2012. "Introduction to Estimation Issues in Multilevel Modeling." *New Directions for Institutional Research* 2012 (154): 23–39. doi:10.1002/ir.20012.

McEwan, Patrick J. 1998. "The Effectiveness of Multigrade Schools in Colombia." *International Journal of Educational Development* 18 (6): 435–52. doi:10.1016/S0738-0593(98)00023-6.

McEwan, Patrick J. 2008. "Evaluating Multigrade School Reform in Latin America." *Comparative Education* 44 (4): 465–83. doi:10.1080/03050060802481504.

McGinn, Noel F. 1998. "Resistance to Good Ideas: Escuela Nueva in Colombia." In *Education Reform in the South in the 1990s*, edited by Lene Buchert, 29–52. Paris, France: UNESCO.

Merrill, M David. 2002. "First Principles of Instruction." *Educational Technology Research and Development* 50 (3): 43–59.

Mina, Alejandro. 2004. "Factores asociados al logro educativo a nivel municipal." Documentos CEDE. Bogotá, Colombia: Centro de Estudios sobre Desarrollo Económico.

Mincer, Jacob. 1958. "Investment in Human Capital and Personal Income Distribution." *Journal of Political Economy* 66 (4): 281–302. doi:10.2307/1827422.

Ministerio de Educación Nacional de Colombia. 2016. "Cierre de Brechas Con Un Enfoque Regional." February 6. http://www.mineducacion.gov.co/1759/w3-article-278741.html.

Misión Social del Departamento Nacional de Planeación de Colombia. 1997. "La Calidad de La Educación Y El Logro de Los Planteles Educativos." *Planeación & Desarrollo* 28 (1): 25–51.

Montessori, Maria. 1912. *The Montessori Method*. New York: Frederick A. Stokes Company.

Moore, Audrey-Marie Schuh, Ana Florez, and Eva Grajeda. 2010. "Evaluation of Education Programs Developed by the Public and Private Alliance between the Coffee Growers Committee of Caldas and the State Government of Caldas, Colombia. Final Report." Academy for Educational Development.

Mulkeen, Aidan G., and Cathal Higgins. 2009. "Multigrade Teaching in Sub-Saharan Africa : Lessons from Uganda, Senegal, and The Gambia," August. https://openknowledge.worldbank.org/handle/10986/5952.

Mullis, Ina V.S., Michael O. Martin, Pierre Foy, and Kathleen T. Drucker. 2012. "PIRLS 2011 International Results in Reading." Chestnut Hill, MA, USA/Amsterdam, the Netherlands: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College/International Association for the Evaluation of Educational Achievement (IEA).

Nuñez, Jairo, Roberto Steiner, Ximena Cadena, and Renata Pardo. 2002. "¿Cuáles Colegios Ofrecen Mejor Educación En Colombia?" 193. Archivos de Economía. Bogotá, Colombia:

República de Colombia, Departamento Nacional de Planeación, Dirección de Estudios Económicos.

Obwoya, Kinyera Sam, William Epeju, Frances Nakiwala, and Ginyakol P Okello. 2004. “Report of Evaluation on Multigrade School Education in Uganda (Draft).” Kyambogo University.

O’Connell, Ann A., and Sandra J. Reed. 2012. “Hierarchical Data Structures, Institutional Research, and Multilevel Modeling.” *New Directions for Institutional Research* 2012 (154): 5–22. doi:10.1002/ir.20011.

OECD. 2016. *Education in Colombia*. Paris: Organisation for Economic Co-operation and Development. http://www.oecd-ilibrary.org/content/book/9789264250604-en.

Piaget, Jean. 1953. *To Understand Is to Invent*. New York, NY: Grossmann.

Pitt, Jennifer. 2002. “Civic Education and Citizenship in Escuela Nueva Schools in Colombia.” Masters Thesis, Toronto, Canada: University of Toronto.

Psacharopoulos, George, Carlos Rojas, and Eduardo Velez. 1992. “Achievement Evaluation of Colombia’s Escuela Nueva: Is Multigrade the Answer?” Policy Research Working Paper. Washington, D.C.: The World Bank.

Rabe-Hesketh, Sophia, and Anders Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata, Volume I: Continuous Responses, Third Edition*. 3 edition. College Station, Tex: Stata Press.

Ramos, Cecilia, Ana María Nieto, and Enrique Chaux. 2007. “Classrooms in Peace: Preliminary Results of a Multi-Component Program.” *Inter-American Journal of Education for Democracy* 1 (1): 35–58.

Rangel, Claudia, and Christy Lleras. 2010. “Educational Inequality in Colombia: Family Background, School Quality and Student Achievement in Cartagena.” *International Studies in Sociology of Education* 20 (4): 291–317.

Raudenbush, Stephen W, and Anthony S Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks: Sage Publications.

Richards, Lyn. 2005. *Handling Qualitative Data : A Practical Guide*. London Thousand Oaks, CA: SAGE Publications.

Rojas, Carlos, and Z Castillo. 1988. “Evaluación Del Programa Escuela Nueva En Colombia.” Bogotá, Colombia: Instituto SER de Investigaciones.

Rubin, Allen, and Earl R Babbie. 2007. *Essential Research Methods for Social Work*. Belmont, CA: Thomson/Brooks/Cole.

Ryoo, J.H. 2011. “Model Selection with the Linear Mixed Model for Longitudinal Data.” *Multivariate Behavioral Research* 46: 598–624.

Sarmiento Gómez, Alfredo. 2006. “Una Estrategia Para Aumentar La Retención de Los Estudiantes.” Ministerio de Educación Nacional, Departamento Nacional de Planeación. http://www.escuelanueva.org/portal/images/pdf/monitoreo/63.UNAESTRATEGRETENCIONDEESTUDNTES.pdf.

Snijders, Tom A. B., and Roel J. Bosker. 2011. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. SAGE.

Somers, Marie-Andree, Patrick J. McEwan, and J. Douglas Willms. 2004. “How Effective Are Private Schools in Latin America?” *Comparative Education Review* 48 (1): 48.

StataCorp. 2013. *Stata Multiple-Imputation Reference Manual, Release 13*. College Station, Tex: Stata Press.

Steiner, Rudolf. 1919. *The Foundations of Human Experience. Vol 1. The Foundations of Waldorf Education*. Allgemeine Menschenkunde Als Grundlage Der Pädagogik. Pädagogischer Grundkurs. Stuttgart, Germany.

Torres, Rosa María. 1992. “Alternatives in Formal Education: Colombia’s Escuela Nueva Programme.” *Prospects - Quaterly Review of Education* 22 (4): 512–20.

UCLA Statistical Consulting Group. 2017. “Stata FAQ: How Can I Estimate R-Squared for a Model Estimated with Multiply Imputed Data?” Accessed January 22. http://www.ats.ucla.edu/stat/stata/faq/mi_r_squared.htm.

Ultanir, Emel. 2012. “An Epistemological Glance at the Constructivist Approach: Constructivist Learning in Dewey, Piaget, and Montessori.” *International Journal of Instruction* 5 (2): 195–212.

UNESCO. 1998. “First Comparative International Study on the Quality of Education.” Santiago de Chile: UNESCO.

UNESCO EFA. 2013. “World Inequality Database on Education.” http://www.education-inequalities.org/.

University of Bristol. 2013. “Centre for Multilevel Modelling: Learning Environment for Multilevel Methods and Applications.” https://www.cmm.bris.ac.uk/lemma/.

Uribe, Maria Camila. 1998. “Eficiencia en el gasto público de educación.” 96. Bogotá, Colombia: República de Colombia, Departamento Nacional de Planeación, Unidad de Análisis Macroeconómico. http://www.escuelanueva.org/portal/images/pdf/monitoreo/22.Uribe2008.pdf.

Veenman, Simon. 1997. “Combination Classrooms Revisited.” *Educational Research and Evaluation* 3 (3): 262–76. doi:10.1080/1380361970030304.

Velez, Eduardo. 1991. “Colombia’s ‘Escuela Nueva ’: An Educational Innovation.” A View from LATHR Nr. 9. Washington, D.C.: The World Bank. http://www.escuelanueva.org/portal/images/pdf/monitoreo/48.COLOMBIASESCNVAANEDUCINNOVATION.pdf.

World Bank. 2009. “La Calidad de La Educación En Colombia: Un Análisis Y Algunas Opciones Para Un Programa de Política.” Washington, D.C. and Bogotá, Colombia: Banco Mundial, Unidad de Gestión del Sector de Desarrollo Humano. Oficina Regional de América Latina y el Caribe.

———. 2010. “Quality of Education in Colombia - Achievements and Challenges Ahead: Analysis of the Results of TIMSS 1995-2007.” Report No. 54351-CO. Washington, D.C.: The World Bank. https://openknowledge.worldbank.org/handle/10986/2916.

———. 2016. “World Development Indicators.” http://data.worldbank.org/data-catalog/world-development-indicators.

Zambrano Jurado, Juan Carlos. 2013. “Multilevel Analysis of School Performance in Mathematics for Fourth Grade of Basic Education in Colombia.” *Sociedad Y Economía*, no. 25 (December): 205–35.

Biography

Katharina Hammler was born in Graz, Austria. She holds Magister degrees in Economics and Socioeconomics from WU (Vienna University of Economics and Business Administration) and in Political Sciences from the University of Vienna. Katharina came to the USA in 2011 as a Fulbright student to pursue her doctorate at Tulane. While at Tulane, Katharina worked as a graduate teaching assistant and as a research assistant on projects focusing on higher education in Africa, e-learning, and tax incidence analysis. Additionally, she has worked as a Monitoring and Evaluation consultant for a USAID-funded education project in Northeast Nigeria, taught at the American University of Nigeria, and served as the Director of Monitoring and Evaluation for Fundación Paraguaya, where her work has focused on multidimensional poverty measurement and micro finance. Katharina currently resides in New Orleans.