Empirical Essays on Education Policy Evaluation

Zelda Brutti

Thesis submitted for assessment with a view to obtaining the degree of Doctor of Economics of the European University Institute

Florence, 12 September 2016

European University Institute
Department of Economics

Empirical Essays on Education Policy Evaluation

Zelda Brutti

Thesis submitted for assessment with a view to obtaining the degree of
Doctor of Economics of the European University Institute

Examining Board

Prof. Jerome Adda, Bocconi University & EUI, Supervisor
Prof. Andrea Ichino, EUI
Prof. Éric Maurin, Paris School of Economics
Prof. Jörn-Steffen Pischke, London School of Economics

Researcher declaration to accompany the submission of written work

I Zelda Brutti certify that I am the author of the work "Empirical Essays on Education Policy Evaluation" I have presented for examination for the PhD thesis at the European University Institute. I also certify that this is solely my own original work, other than where I have clearly indicated, in this declaration and in the thesis, that it is the work of others.

I warrant that I have obtained all the permissions required for using any material from other copyrighted publications.

I certify that this work complies with the *Code of Ethics in Academic Research* issued by the European University Institute (IUE 332/2/10 (CA 297).

The copyright of this work rests with its author. [quotation from it is permitted, provided that full acknowledgement is made.] This work may not be reproduced without my prior written consent. This authorisation does not, to the best of my knowledge, infringe the rights of any third party.

Statement of inclusion of previous work (if applicable):

I confirm that chapter 3 was jointly co-authored with Fabio Sánchez Torres and I contributed 80% of the work.

Signature and Date:

Zelda Brutti, 1 September 2016

Abstract

This thesis is composed of three empirical essays that share the theme of education policy evaluation.

Chapter 1 looks at cognitive spillovers between siblings who live in the same household. The research question asked is whether such spillovers exist, and what their magnitude may be. Using longitudinal data on children aged 5 to 14 and living in the United States, I identify positive, significant and sizable cognitive spillovers between siblings. In the light of these findings, education policies that increase the cognitive achievement of any target population of children may also indirectly benefit their cohabiting siblings.

Chapter 2 evaluates the outcome of a decentralization reform that has affected public education in Colombia. Municipalities with more than 100 thousand inhabitants were made autonomous in the management of public education on their territories, while cities below that threshold were not given autonomy. I exploit this quasi-experimental setup to evaluate the impact of local autonomy on educational quality. I find that autonomy was beneficial for municipalities that were highly developed at the time of the reform, but the impact was negative for low-developed ones, increasing inequality across regions. These results sound a note of caution about the potential consequences of decentralization reforms, in educational contexts similar to the Colombian one.

Chapter 3 is coauthored with Fabio Sánchez Torres at the University of Los Andes, and looks at how quality assurance mechanisms in the careers of pub-

lic school teachers have affected student performance in Colombia. The main novelties consist in a selective entry examination, a probation period and permanent evaluation processes. We find that teachers who operate under the new regulation noticeably improve pupils' achievement with respect to colleagues who follow the old rules, which feature very few incentives for teaching quality. Our findings may provide education policy guidance to several Latin American countries, in which current teacher regulations resemble the former Colombian one.

Acknowledgments

This thesis is an important achievement to me, and I feel profoundly grateful to all those people that gave me help, motivation and encouragement towards its completion.

I thank Jérôme Adda, Steve Pischke and Andrea Ichino for their invaluable guidance during my doctoral studies. I thank the European University Institute and the London School of Economics for providing me with stimulating and pleasant research environments. By environments I mean the places and resources, but especially the people giving life and energy to these institutions. I thank my fellow PhD students for all those hours of sharing ideas, discussions and sometimes desperation (!), that have made me grow academically and personally.

I thank the OEAD and the Luca d'Agliano Foundation for their generous financial support during my studies.

Last but foremost, I thank those people whose multidimensional support has been too great to be properly acknowledged here, and to whom my deepest gratitude goes.

To my past,

Maria, Mirko and Brando

and to my future,

Luis

Contents

Chapter 1

Sibling Effects in Cognitive Achievement:

Are Smart Brothers and Sisters Good for You?

Abstract

In this paper I investigate whether children's cognitive achievement is influenced by the cognitive achievement of their siblings. I apply system GMM to estimate a dynamic model of cognitive skill production. I allow for endogeneity of sibling performance and of other regressors, for reciprocal influence between younger and older sibling, for unobserved individual and family effects, and for state dependence of cognitive achievement. I find evidence for positive and significant sibling effects on both verbal reasoning and mathematics test scores, with results consistent across the two skills. The results support theories of cumulative achievement processes, as current achievements are shown to depend on past ones. Influence goes from older siblings to younger ones and viceversa, in contradiction with the assumption of unidirectional effect from old to young, which underlies identification in related studies.

1 Introduction

One of the most active debates in the child development literature revolves around defining the elements that concur to the production of cognitive skills in early years of life. Endowments at birth, parental investment, family environment and early schooling are among the most widely discussed factors. In this paper I focus on the role of siblings. I investigate whether the cognitive performance of siblings influences a child's achievements. Numerous studies in sociology and developmental psychology show that children are more responsive to siblings than to other peers such as friends or classmates[1]. Nevertheless, the amount of Economics research dedicated to peer effects among siblings is surprisingly small.

It is helpful to clarify what kind of peer effects this piece of work aims at identifying. We are dealing with what Manski [1993] defines as "endogenous peer effects", i.e. those occurring when individual outcomes are influenced by *the outcomes or the behavior* of peers[2]. This definition encompasses direct and indirect channels of influence. Siblings could influence each other directly through teaching or imitation, but also indirectly in several manners. The way a child's cognitive achievement compares to his sibling's may modify parental investment behavior. Having an ambitious and high achieving sibling might lead to a higher number of books being kept at home. The set of potential channels of influence is large, and it is not within the scope of this paper to identify which ones are operating. Data at hand allow me to rule out the channel of parental investment feedbacks; beyond that, all channels through which the level of cognitive achievement of the sibling may influence a child are included in the estimated effect.

[1] See for example Ardelt and Day [2002], who find that older siblings have a stronger effect than peers and parents on deviant behavior of adolescents. Also Azmitia and Hesser [1993] show through an experimental setting that younger children are more likely to imitate and learn a task from older siblings than from older peers or parents, and obtain better results if they do so.

[2] The challenge is to detach endogenous effects from other two categories of peer influences: "exogenous effects" and "correlated effects". Exogenous effects measure how *the characteristics* of the peer group influence the individual outcome. In the case of siblings, relevant characteristics could be the sibling's age, gender, birth order, health. Correlated effects instead represent the similarity in behavior that is explained by *background characteristics and environmental factors* that peers share. In the case of siblings, we can think of family effects: growing up in the same family environment drives correlation in life outcomes.

Two are the main observations that generate interest for the results of the exercise I here propose. Firstly, do policies which are aimed at improving educational achievement of a specific population of children have spillover effects on other siblings back in the households? In other words, is there a "social multiplier" effect in place in this context? The answer is highly relevant for program budgeting and cost-benefit analyses. Secondly, siblings are frequently used as control groups to evaluate the effectiveness of policy interventions[3]. If performance spillovers do actually take place, this strategy would turn out to be flawed.

Both the outcome I am considering and the identification methodology I am employing introduce novelty to the research on endogenous sibling effects. I carry out the empirical analysis on a large samples of US families, using longitudinal data from the National Longitudinal Survey of Youth 1979 Cohort (NLSY79). The outcome I look at is the performance in cognitive tests measuring skills in math and verbal knowledge, at ages between 3 and 14. Early results of this sort are well known to have strong predictive power on future academic and life outcomes, and are therefore highly valuable as targets for policy intervention. I explore how test scores of a child are impacted by the test scores that the sibling achieved in the same test session. The choice of considering contemporaneous score achievements, and not achievements distant in time, is motivated by the fact that the strongest channels of sibling influence during childhood are everyday interaction and reciprocal observation, rather than memories or past records[4]. These test scores are taken as measures of the children's cognitive ability in the areas of verbal reasoning and in mathematics. The scope of the exercise is to isolate the influence of the ability of one child on the ability of the other, removing confounding factors such as family effects, characteristics of the sibling other than his or her cognitive performance, and parental investment. If not appropriately controlled for, these and other elements give rise to the issues of endogeneity that typically affect peer and sibling studies. Beyond exploiting the rich information provided by the

[3]See for example Currie and Thomas [1995] and Garces et al. [2002], evaluating the impact of the *Head Start* program on later life outcomes.

[4]For more formal discussions on this topic, see for example Green et al. [1994] with experimental evidence on the high time discounting of children, and the papers they cite.

NLYS79 panel, I employ a system GMM estimation methodology to overcome these issues. Importantly, this methodology also allows to account for a dynamic skill formation process, i.e. for the fact that past achievement plays a structural role in explaining achievement today. The assumptions underlying identification are different and less restrictive than the ones made in related studies. I find evidence for positive, large and significant influence between siblings' cognitive performances, for both numerical and verbal skills. Influence goes both from the older sibling to the younger and from the younger to the older.

The rest of the paper is organized as follows. Section 2 places the paper in the context of the existing related literature; Section 3 describes the data resources used for the empirical exercise; Section 4 addresses the empirical strategy used to identify sibling influence; Section 5 presents and discusses the estimation results and Section 6 concludes.

2 Related Literature

Research on cognitive skill formation has been attracting increasing numbers of contributions in Economics, from the earliest works of Becker and Lewis [1973] and Becker and Tomes [1986] to present, witnessing continuous improvement especially on the empirical side, thanks to the progress in econometric methods and to the increasing quality of available datasets. In recent years, Cunha and Heckman [2009] and Aizer and Cunha [2012] study how birth endowments and later investments interact in the production of human capital in childhood. Todd and Wolpin [2007] test several specifications of skill production functions, accounting for family unobservables and history dependence. History dependence of cognitive achievements, i.e. the fact that past achievements influence current ones, is nowadays extensively supported by the literature on life cycle skill formation. Back in the late Seventies, Boardman and Murnane [1979] present and estimate a model featuring cumulative patterns in cognitive achievement. More recently, multiple pieces of research by Heckman [2000, 2006] bring new evidence on the fact that skill formation is a dynamic process in which early inputs strongly affect the productivity of later inputs. Cunha and

Heckman [2007, 2008] and Cunha, Heckman and Schennach (2010) formulate and estimate multistage production functions for children's cognitive and non-cognitive skills, showing that the return to cognitive investment during a given period of the child's life depends on the skill stock inherited from the former period. The estimation method I adopt in this paper allows for dependence of current achievement on its past value.

The role of peers is one of the most recent aspects that research on cognitive achievement has been looking at. I will omit the discussion of literature on exogenous and correlated peer effects, and focus instead on contributions exploring endogenous effects, as my empirical exercise does. Sacerdote [2001] exploits the random-pairing of Dartmouth college students into rooms, and proposes a structural model to isolate causal influences between roommates' grades; he finds evidence for (modest) positive peer influence between GPAs of students sharing the room. In a similar paper, Zimmerman [2003] proposes evidence for not large but significant effects between students' SAT scores; in particular he finds stronger effects for verbal achievement scores, with respect to mathematics. Effects of larger magnitude are reported by Carrell et al. [2009], who use data on US Air Force Academy students and find that peer results do significantly affect individual academic achievement. Patacchini et al. [2011] exploit friend nominations in AddHealth data and find that friends significantly influence the length of each others's periods of education. Zimmer and Toma [2000], Hoxby [2000], Betts and Zau [2002], Hanushek et al. [2003], Robertson and Symons [2003], Arcidiacono and Nicholson [2005] and Vigdor and Nechyba [2007] are further examples of recent studies addressing the influence of schoolmates' abilities and achievements on own achievement. For additional models, estimation techniques and empirical findings about peer effects in education, one can refer to the excellent and thorough review by Sacerdote [2011].

We finally come to the sparse literature that has attempted to isolate endogenous sibling influences. Powers and Cherng-tay Hsueh [1997] look at how the probability of premarital childbearing of young women is influenced by the occurrence of the same event for their elder sisters; the effect they isolate is positive and significant. Ouyang [2004] and Altonji

et al. [2010] focus on risky teenage behaviors: their results indicate that having an older sibling who smokes, engages in unprotected sexual activities or consumes alcohol or drugs increases the probability for an adolescent to behave similarly. Oettinger [2000] finds that the probability of graduating from high school on time is higher if older siblings have achieved the same goal before.

To some up, one strand of research has investigated how the cognitive performance of peers such as friends, classmates and college roommates influences individual cognitive performance. Another strand has looked at how specific behaviors of siblings influence a child's own behavior. In this paper I fill the gap between the two branches and investigate whether the cognitive performance of siblings has a causal effect on a child's own achievement.

3 Data

3.1 NLSY79 (National Longitudinal Survey of Youth)

The NLSY79 [5] is an ongoing longitudinal project that follows the lives of a sample of American youth born between 1957-64. The panel collects a vast and detailed range of information about the families: demographics, education, employment, income, wealth, housing, expenditures, health, marriage, childbearing, and several other topics. The NLSY79 Child and Young Adult Cohort[6] follows biological children of the women in the NLSY79. Data are now available from 1986 to 2010, with 13 survey rounds having been administered to the child sample. These look at individuals that had not turned 15 by the end of the calendar year in which each survey is conducted. Cognitive assessments of these children are taken in the form of Peabody Individual Achievement Tests (PIAT), administered to individuals aged 5 to 14. I examine the results of the PIAT Math and the PIAT Reading tests; PIAT

[5]Bureau of Labor Statistics, U.S. Department of Labor. National Longitudinal Survey of Youth 1979 cohort, 1979-2010 (rounds 1-24) [computer file]. Produced and distributed by the Center for Human Resource Research, The Ohio State University. Columbus, OH: 2012.

[6]Bureau of Labor Statistics, U.S. Department of Labor, and National Institute for Child Health and Human Development. Children of the NLSY79, 1979-2010 [computer file]. Produced and distributed by the Center for Human Resource Research, The Ohio State University. Columbus, OH: 2012.

scores are standardized in order to capture individual performance compared to the average performance of the age group. I select sibling pairs for which at least 2 rounds of results are available. I could make use of a final sample of about 3,900 children[7]. for the empirical analysis.

3.2 Relevant information

The Choice of Sibling Pairs

The NLSY study observes all children in the household. In 38% of the cases, families were 2-children families. In 33% of the cases, a third sibling was born during the observation span; 17% of the families recorded 4 children; 12% recorded more. For the main empirical analysis I associate to each child only his or her closest-in-age sibling. Dealing with multiple siblings complicates model and estimation to great extent. One might have concerns about confounding effects coming from the influence of the siblings excluded from estimation; to placate these concerns I run a sensitivity analysis on a reduced sample of 2-children families only, which I present in the Appendix. The reduced specification implies a major sample reduction without major changes in the estimated coefficients. Full family size is also included among the control variables in the analysis.

Parental investment: The HOME score

A large theoretical and empirical literature shows that parental investment is a crucial factor in the production of cognitive skills in children. The Home Observation for Measurement of the Environment (HOME) Inventory is one of the most prominently used instruments to measure parental investment. The HOME score is constructed using both mother reports and interviewer observations about the overall quality of the home environment, emotional and verbal responsiveness of the mother, maternal acceptance and involvement with the

[7]This figure is reduced to about 3,400 for the Reading test, due to a larger number of missing values for these results.

child, organization of the environment, presence of materials for learning, and variety of stimulation[8]. The HOME measure is taken for each child in the household individually, at each survey round, along with the cognitive tests and the household interview. It has been employed in literature to investigate intrahousehold differences in parental investment - attributing these to gender, birth endowments, disabilities, different levels of cognitive achievement, and to study the effects of such investments (see for example Todd and Wolpin [2007], Cunha and Heckman [2008], Coneus et al. [2010], Zuppann [2013]). Given the widespread acceptance of its reliability and accuracy[9], the HOME measurement has been introduced in all major household interviews that touch issues such as intrahousehold relationships, child development, and child welfare - including the NLSY79. In my analysis I use the HOME score as a control for parental investment, to account for potential changes in attitude of the parents towards one or both siblings, and to be thus able to exclude the parental investment channel from the influence between siblings.

4 Empirical Specification and Estimation Method

In his 1993 seminal work on peer effects, Manski shows the general conditions under which we have a chance to isolate endogenous peer effects from exogenous and correlated ones. He concludes his theoretical analysis claiming *"there may be realistic prospects for inference on endogenous effects if the attributes defining reference groups and those directly affecting outcomes are moderately related"*. This means that the two sets of attributes should be neither statistically unrelated nor a function of each other. Moreover he adds that the composition of reference (peer) groups needs to be known a priori, and not to be extrapolated from the data. In the setup of my analysis, I am considering a well defined and a priori known reference group: siblings in the household. The first condition identified by Manski is also satisfied: there is a partial relation between the attributes characterizing

[8]Description from "NLS Handbook, 2005, Chapter 4: Children of the NLSY79".

[9]Mott [2004] provides a valuable review of the intensive usage that social science research, and especially child development studies, have made of the HOME scale over the past years.

the reference group (the children in the household) and the set of attributes that directly impact on individual test scores. In fact, all siblings in the household are characterized by some shared family characteristics (e.g. mother education and family income), but there are other variables which impact individual outcomes without being systematically related to the reference group (e.g. each child's birth endowments, ability and health dynamics). As a result, the two sets of attributes are correlated but not functionally dependent from each other.

After having established that identification of sibling influence is possible, I will proceed looking at which are the factors that make it challenging. The results of an ordinary least squares regression of own achievement on sibling's achievement is biased by the following factors:

1. Reverse causality. The influence between the two children may flow in both directions, which would lead to an overestimation of the true sibling effect.

2. Family effects. Both siblings are influenced by some common characteristics and by dynamics affecting the family as a whole. This fact inflates naive estimations of the sibling effect, as it introduces spurious correlation between children's outcomes.

3. Individual fixed effects. Each child has inborn characteristics, such as ability and personality traits, which are unobserved and potentially correlated with observed regressors. As a result, OLS estimates are biased and inconsistent, with the direction of the bias depending on the sign of such correlations.

4. Dependence on past achievement. As discussed above in Section 2, theoretical and empirical evidence shows that past cognitive achievement plays a structural role in explaining current achievement. OLS estimation of a linear model featuring an autoregressive term is biased - as well as fixed effects estimation, especially with short panels.

Equation (1) shows the model I estimate in the empirical analysis; it is written out for one

of the two siblings and it is specular for the other.

$$y^a_{ft} = \rho y^a_{f(t-1)} + \beta y^b_{ft} + \sum_{i=1}^{I} \gamma_i x^a_{ift} + \sum_{j=I+1}^{J} \gamma_j x^b_{jft} + \sum_{k=J+1}^{K} \gamma_k x_{kft} + \alpha^a_f + \epsilon^a_{ft} \qquad (1)$$

where index t denotes time; index f denotes families; indexes a and b distinguish the two siblings; y^a_{ft} and y^b_{ft} are current cognitive achievements; $y^a_{f(t-1)}$ is lagged cognitive achievement; x^a_{ft} and x^b_{ft} are observed time-varying individual characteristics of the two siblings; x_{ft} are observed time-varying family characteristics; α^a_f represent observed and unobserved, time-invariant, individual and family characteristics; ϵ^a_{ft} is the idiosyncratic shock. The x^a_{ft}, x^b_{ft} and x_{ft} are not necessarily exogenous: they might be predetermined (uncorrelated with the contemporaneous error term, but correlated with past ones) or endogenous (correlated also with the contemporaneous error term).

I estimate the model applying the system GMM (SGMM) estimator, whose final version is the result of the contributions by Arellano and Bond [1991], Arellano and Bover [1995] and Blundell and Bond [1998], to which the reader may refer for thorough discussions - as well as to the excellent pedagogical papers by Roodman (Roodman [2009a] and Roodman [2009b]) . The estimator's native field of application is the short and wide panel data case, i.e. a situation in which few time periods are observed for a large number of individuals (I face $T = 13$ and N=3,900 using the NLSY). The SGMM estimation procedure can be concisely summarized as follows:

- The equation specified by the researcher is estimated combining two different procedures: I will refer to the two procedures as "in deviations" and "in levels".
- "In deviations" procedure. A forward orthogonal deviations transform is applied to

the equation[10], thus eliminating fixed effects from the error term[11]. The right hand side variables that are believed to be endogenous after the transformation are instrumented. The matrix of instruments is composed of contemporaneous and lagged values of exogenous regressors[12]; lagged values of predetermined regressors; lags 2 and deeper of the endogenous regressors[13]. The set of moment conditions underlying this instrumenting procedure is[14]:

$$E\left[y_{f(t-1)}\epsilon^*_{ft}\right] = 0 \quad \text{for } 2 \leq t \leq T-1 \quad \text{, i.e. } (T-1) \text{ cond.} \tag{2}$$

where y indicates any endogenous regressor. The assumption needed for 2 to hold is absence of serial correlation in the idiosyncratic error term ϵ_{ft}. If ϵ_{ft} follows an AR(1) process of the sort $\epsilon_{ft} = \rho\epsilon_{f(t-1)} + w_{ft}$, then $y_{f,t-1}$ is correlated to ϵ_{ft} through $\epsilon_{f(t-1)}$. Then $y_{f,t-1}$ is also correlated to ϵ^*_{ft} and condition 2 fails. Valid instruments are then y_{t-2} and deeper lags.

- "In levels" procedure. The equation is kept untransformed, and the right hand side variables that are potentially correlated with the error term through fixed effects are instrumented. The matrix of instruments is composed of differences and lagged differences of exogenous and predetermined regressors12, and of lagged differences of the endogenous regressors. The fixed effect components are absent in these differenced

[10] To each of the first $T-1$ observations, we subtract the mean of the remaining future observations available in the panel[Arellano and Bover, 1995]. Formally, $x^*_{ft} = x_{ft} - \frac{\sum_{i=t+1}^{T} x_i \mathbb{1}(x_i \neq .)}{\sum_{i=t+1}^{T} \mathbb{1}(x_i \neq .)}$, where the function $\mathbb{1}(x_i \neq .)$ assumes value 1 when x_i is nonmissing and 0 otherwise. This method preserves lack of correlation among the transformed variables if the original variables are not autocorrelated and have constant variance (Arellano and Honore [2001], page 3256) and minimizes data loss in case of missing observations across the panel.

[11] An alternative method is to first-difference the equation. Due to the fact that only 3 waves of Child Development Supplement data are available from the PSID dataset, I minimize data loss by choosing the orthogonal deviations transform for my analysis, thus I focus on that option in the rest of the discussion.

[12] and of other external instruments, if available

[13] Specifically, the instrument set is the standard one proposed by Holtz-Eakin, Newey and Rosen (1988) Holtz-Eakin et al. [1988].

[14] In the case of first-differencing the conditions would be: $E\left[y_{f(t-l)}\Delta\epsilon_{ft}\right] = 0$for each $t \geq 3, l \geq 2$, i.e.$(T-2)(T-1)/2$cond.

regressors, making them uncorrelated with the present error term and thus valid instruments. The set of moment conditions underlying this procedure is:

$$E\left[\Delta y_{it-1}\epsilon_{it}\right] = 0 \quad \text{for each} t \geq 3 \quad \text{, i.e. } (T-2) \text{ cond.} \tag{3}$$

Call y_{i1} the first observed individual outcome. Call $\bar{y_i}$ the value to which each individual outcome process converges over time. Condition 3 holds if the deviations of y_{i1} from $\bar{y_i}$ are not correlated with $\bar{y_i}$ itself, or in other words with the individual fixed effect. This assumption is formalized in Blundell and Bond [1998] and further discussed in the next subsection.

The combination of these two procedures makes the SGMM estimator suitable for empirical specifications featuring fixed effects, state dependence, regressors that are endogenous because of reverse causality. SGMM coefficient estimates have causal interpretation, given that assumptions 2 and 3 hold. Compared to standard in-differences estimators such as the Arellano and Bond [1991] GMM estimator, SGMM has the advantages of mitigating the weak instrument problem of which the former suffer[15], of increasing efficiency of the estimation and of allowing the estimation of coefficients on time invariant regressors.

4.1 Assumptions

Attempts to answer questions similar to the one I am addressing here have been based on a diverse set of assumptions. Ouyang [2004] and Altonji et al. [2010] assume that influence goes only from the older to the younger sibling and exclude the other direction. Ouyang also overlooks the presence of individual fixed effects; Altonji assumes away parental investment dynamics. Oettinger [2000] does not use panel data, omits fixed effects of any sort and uses child-specific characteristic to instrument graduation probability. Here below I briefly

[15] See the discussion in Blundell and Bond [1998].

discuss the assumptions on which my identification strategy is based.

4.1.1 No residual serial correlation

For condition (2) to hold and the "in deviations" procedure to be valid, there should be no serial correlation in the residual error term, conditional on observables and after purging out the fixed effect. The assumption will hold if we are effective in eliminating fixed effects and controlling for the factors that are likely to provoke correlation over time in outcomes. The validity of the assumption can be tested through a post-estimation AR test. An AR(2) test on the differenced residuals will spot AR(1) relationships in levels. I show the test results for the NLSY79 data in Section 5. The null hypothesis of no residual serial correlation is not rejected at the standard significance levels.

4.1.2 No correlation between initial deviations and individual fixed effects

For condition (3) to hold and the "in levels" procedure to be valid, we need a restriction on the outcomes observed at the first period. For a formal discussion, see Blundell and Bond [1998]. The underlying concept is quite straightforward. Say each child has a long term outcome to which he/she converges over time, and this long term level is influenced by the individual fixed effect. It should not be the case that the deviation of the first-period outcomes from the individual long term level is correlated to the individual fixed effect. Violation of this condition would reintroduce the fixed effect into the differenced regressors used as instruments, as the growth rate over time (and thus the differences) would correlate with the fixed effect through the first-period deviation. In fact this condition is not deemed restrictive in our setup. Some children will have a lower overtime baseline outcome, and other children will have a higher one. There is no research evidence or intuitive reason to think that one of these groups will be further away from its baseline than the other group at the first wave of assessments. Postestimation overidentification tests that verify the validity of the set of differenced instruments are reported in the result tables.

4.1.3 No residual correlation across individuals

The SGMM estimator assumes no correlation across individuals in the idiosyncratic disturbances. Both the AR test for residual serial correlation and the computation of robust standard errors on coefficient estimates rely on this assumption. The inclusion of time dummies in the model should suffice to alleviate concerns about correlation across families, given they are randomly sampled across the US. Somewhat more worrisome is the possible residual pairwise correlation between siblings. Correlation deriving from time-invariant factors is wiped out in the SGMM estimation procedure; what remains to discuss are time-variant shocks that are common to the two siblings and not eliminated along with the fixed effects. This is where the high quality of datasets such as the NLSY79 becomes relevant. We can observe changes in family composition, relationship status of parents, income dynamics, health of parents, neighborhood ratings[16]. If we believe that these measures are able to account for the effects of shocks affecting wealth and income (job losses, financial fortunes or misfortunes), the family history (births, deaths, adoptions, parents splitting up), in health and in the living environment, we should be able to trust the share of residual unobserved correlation to be reassuringly small.

4.1.4 Which regressors are exogenous, predetermined, endogenous?

Exogenous regressors are uncorrelated with current and past error terms. Predetermined regressors are uncorrelated with present errors, but might be correlated with past ones. Endogenous regressors might be correlated with present and past error terms. The classification of regressors into the categories of exogenous, predetermined and endogenous determines how they are handled during the SGMM estimation procedure. Specifically, it determines how they contribute to the instrument matrix - as described in the SGMM summary in Section 4. Misclassifying a regressor as exogenous or predetermined when it is in fact endogenous, or exogenous when it is in fact predetermined, means introducing invalid

[16]For the sake of simplicity, health of parents and neighborhood ratings have been left out in the final specification of the model, as results turned out to be insensitive to these factors.

instruments into the SGMM instrument matrices. The other direction of misclassification, i.e. classifying an exogenous or predetermined factor as endogenous, or an exogenous as predetermined, implies unnecessary loss of valid instruments but no bias is introduced. I adopt a conservative classification, to avoid invalidating the set of instruments. I classify as exogenous gender, ethnicity and the time dummies. The level of education of the mother and the lagged test score are predetermined. Finally, I classify as endogenous the performance of the sibling, health indicators, family composition variables, relationship status, income measures, and HOME scores. The results of the post estimation overidentification tests provide additional support for the validity of these choices.

5 Results

5.1 Verbal Skills Achievement Test

Table (1) shows the estimation results for the Verbal achievement tests. The first column shows the SGMM estimation results, and is followed by Fixed Effects and pooled OLS results.

I find that a child's performance in the verbal skill tests is positively affected by a higher performance of his or her sibling. The estimations on the two datasets yield similar estimations of the coefficient of interest[17]: a child's test performance is increased by around 0.40 standard points when the performance of the sibling is increased by one standard point.

I find evidence for positive state dependence in the test performance, a fact that provides additional empirical support to the theories of cumulative cognitive skill formation. I find no evidence towards significant age gap effects. Other stylized facts the estimation meets are that females achieve better verbal scores than males; blacks and hispanics do worse on average; better health ratings induce better cognitive performance.

Notice that the fixed effects estimations yield negative state dependence, and an estimated

[17]Recall that the results of both tests are standardized, and thus the sizes of the estimated coefficients are comparable across the two panels.

sibling effect half the size with respect to the SGMM results. As demonstrated in the early work by Nickell [1981], applying fixed effects to short, dynamic panels with a large number of individuals induces significant bias in the coefficient on the lagged dependent variable. The bias is shown to be invariably negative when the structural state dependence is positive. Also, the fixed effects model does not adequately handle the issues of inverse causality between the results of the two siblings. The pooled OLS estimation, which ignores altogether the presence of fixed effects, performs somewhat better and yields state dependence of the correct sign, but smaller coefficients on the sibling effect.

Table (1) also reports the results of the overidentifying restrictions tests on the validity of the instruments used in the SGMM procedure, and the AR tests for residual autocorrelation[18]. The Hansen overidentification test reports on whether the instruments, as a group, appear exogenous[19]. The Difference-in-Hansen tests considers two subgroups of instruments separately: differenced instruments that are used for the equation in levels, and the levels instruments used for the transformed equation[20]. The outcomes of all three tests induce confidence in the validity of the instruments used. As to the test for residual autocorrelation, recall from section 4.1.1 that we expect to find autocorrelation of first order, but do not wish to find autocorrelation of second order. Autocorrelation of second order in differences implies autocorrelation in levels, and this would discredit the validity of the instruments used in relation with the transformed equation. Gratifyingly, the test results meet both the expectations of presence of AR(1) and absence of AR(2) in first differences.

[18]Recall the panel from the PSID is too short to obtain results for these tests.

[19]Note: Roodman [2009b] warns about the risks of relying on the results of the Hansen test without further inquiry. The test is weakened when the number of instruments used in the SGMM procedure is high relative to the sample size, leading to systematic under-rejection of the null of valid instruments. Literature does not provide a commonly accepted rule of thumb for a "safe" ratio between sample size and instrument count, but the figures that have come into consideration by Roodman and others are well below my ratios of 11 (NLSY79 sample) and 44 (PSID sample), which thus appear adequate even by conservative standards.

[20]The test compares the values that the Hansen statistic takes with and without the relevant set of instruments. The Difference-in-Hansen test on differenced instruments is indicative about the validity of the assumption underlying the "in levels" part of the SGMM estimation, i.e. that initial deviations from long run outcomes are not residually correlated with individual fixed effects.

Table 1: Verbal Skills

	SGMM	Fixed Effects	Pooled OLS
Lagged score	0.189***	-0.351***	0.407***
	(0.03)	(0.02)	(0.02)
Sib.score	0.404***	0.169**	0.182***
	(0.05)	(0.05)	(0.02)
Sibsc.*Agegap	0.001*	-0.003	-0.000
	(0.00)	(0.00)	(0.00)
Time dummies	Yes	Yes	Yes
Demographics	Yes	Yes	Yes
Family vars	Yes	Yes	Yes
Health, HOME	Yes	Yes	Yes
SS	3,424	3,424	3,424
Fstat p-value	0.000		
N. of instruments	367		
Hansen overid	308.504		
(pval)	(0.910)		
Diff.H, levels	82.648		
(p-val)	(0.459)		
Diff.H, transformed	246.009		
(p-value)	(0.887)		
AR(2) in fd	-1.610		
(p-value)	(0.107)		

5.2 Math Skills Achievement Test

Table (2) reports the estimation results for the Mathematics achievement tests. The result patterns are the same as for the verbal skill tests, and the size of the sibling effect is of the same magnitude as the one that was found for the verbal tests.

Positive dependence on past achievement is again confirmed, and no age gap effects are spotted. Racial minorities perform less well and better health status has positive influence on results. Girls appear to do worse than boys in mathematics, while they were better in verbal skills.

The results of the Fixed Effects and Pooled OLS estimations follow the same patterns as in the case of verbal skills: state dependence appears negative using fixed effects; the pooled OLS does not perform badly but consistently underestimates the coefficients on sibling performance.

The Hansen overidentification test on the instrument set as a whole supports instrument validity. The result of the difference-in-Hansen test for the levels equation is less supportive, while the instrument set for transformed equation does not give reason to worry. Also the result on the autocorrelation test does again meet our positive expectations.

Table 2: Mathematics Skills

	SGMM	Fixed Effects	Pooled OLS
Lagged score	0.228***	-0.251***	0.434***
	(0.03)	(0.02)	(0.01)
Sib.score	0.359***	0.154***	0.179***
	(0.05)	(0.04)	(0.01)
Sibsc.*Agegap	-0.000	-0.002	0.000
	(0.00)	(0.00)	(0.00)
Time dummies	Yes	Yes	Yes
Demographics	Yes	Yes	Yes
Family vars	Yes	Yes	Yes
Health, HOME	Yes	Yes	Yes
SS	3,948	3,948	3,948
Fstat p-value	0.000		
N. of instruments	372		
Hansen overid	364.352		
(pval)	(0.250)		
Diff.H, levels	104.807		
(p-val)	(0.053)		
Diff.H, transformed	299.191		
(p-value)	(0.183)		
AR(2) in fd	-1.337		
(p-value)	(0.181)		

5.3 General remarks on the results

As noted by Roodman [2009b], estimators that employ exclusively "internal" instruments to achieve identification (in this case, lagged levels and differences of the regressors) should always be treated with caution and preferably limited to cases in which alternative strategies are unavailable. The identification of endogenous sibling influences arguably represents one of these cases. The environment we face when we approach the research question brings about challenges that the SGMM estimator is able to overcome, under specified conditions. Overall the estimation results are consistent across the two skill categories and across the two samples employed, which induces confidence about their robustness. Even though not always completely satisfactory, in general the postestimation tests on the validity of the instruments used in the SGMM procedure appear to support the soundness of the identification obtained.

Table (**??**) in the Appendix shows the SGMM results for reduced samples of 2-children families only. We can see that the coefficients of interest do not vary by much with respect to the full sample specifications. Two-children families account for somewhat more than 38% of the full sample a priori.

5.4 Younger Vs. Older Siblings

In this section I look at whether influence in cognitive achievement goes only from the older brother to the younger, or flows also into the opposite direction. As anticipated in the review of Section 2, the assumption of unidirectional influence from the older to the younger sibling has been used by some authors to achieve identification of the sibling effect. The assumption seems somewhat dubious, especially in light of research on child development by sociologists and educational scientists. Among these, experimental evidence by Azmitia and Hesser [1993] shows that the performance of older siblings in performing a task is positively affected when younger siblings are eager to learn and prompt them to explaining it in detail. Also Hartup [1989] describes how influence is bidirectional between young and old, and how it depends on respective ability levels. More in general, we would expect interactions between individuals to have repercussions on all participants, and not on one party alone. Table (3) shows estimations for verbal and mathematics skills, with Panel A looking at the effect of the older sibling on the younger and Panel B looking at the opposite direction. The sample sizes obviously drop with respect to previous estimations, due to the split into younger and older siblings[21]. The main finding is that results show significant influence both from old to young and from young to old, the former being larger than the latter. Regarding the postestimation tests on the validity of instruments, it is good to notice that the ratios 'sample size to instrument count' are reduced due to the sample splits. Nevertheless, since all test results are comfortably above acceptance levels, we should be able to be once again faithful about the reliability of the estimations. In conclusion, from the analysis I drew evidence for bidirectional influence on cognitive achievement between younger and older siblings.

[21]The PSID sample is evenly split into two, as only two children per household are recorded - one is the younger and the other is the older. Recall that the NLSY79 follows instead all children in the household, and I associated to each child his/her closest-in-age sibling. If there are more than 2 children in the family, the number of children who are associated to a younger sibling and the number of those associated to an older sibling are not necessarily equal and depend on the birth spacing patterns.

Table 3: Influences from Older to Younger, and viceversa

	Panel A: Older on Younger		Panel B: Younger on Older	
	Verbal	Math	Verbal	Math
Lagged score	0.177***	0.234***	0.155***	0.200***
	(0.04)	(0.03)	(0.04)	(0.04)
Sib.score	0.342***	0.342***	0.239***	0.146*
	(0.06)	(0.05)	(0.06)	(0.06)
Sibsc.*Agegap	0.002**	0.001	-0.000	-0.001*
	(0.00)	(0.00)	(0.00)	(0.00)
Time dummies	Yes	Yes	Yes	Yes
Demographics	Yes	Yes	Yes	Yes
Family vars	Yes	Yes	Yes	Yes
Health, HOME	Yes	Yes	Yes	Yes
SS	2,130	2,515	1,262	1,401
Fstat p-value	0.000	0.000	0.000	0.000
N. of instruments	364	368	330	341
Hansen overid	312.965	347.300	298.900	309.892
(pval)	(0.851)	(0.440)	(0.603)	(0.586)
Diff.H, levels	76.006	103.535	77.783	86.054
(p-val)	(0.606)	(0.046)	(0.329)	(0.141)
Diff.H, transformed	257.708	288.773	236.052	260.532
(p-value)	(0.738)	(0.286)	(0.664)	(0.427)
AR(2) in fd	-1.261	-1.588	-0.956	-0.877
(p-value)	(0.207)	(0.112)	(0.339)	(0.380)

6 Conclusion

Identifying if and how siblings influence each other is always a challenging econometric exercise. Identification is hampered by inverse causality between the outcomes of the two siblings and by presence of individual and family fixed effects. When cognitive achievement is the outcome of interest, we also have to allow for dependence of current achievements on past achievements. External instruments are hardly available in this framework, as most circumstances potentially changing cognitive progress of a child would affect the other children in the household too. In this paper I turned to the System GMM (SGMM) estimator, which combines estimation in levels and estimation in orthogonal deviations, using differences and levels of past values of the regressors as "internal" instruments. Even though estimators of this sort should be handled with care, I find that the framework of my analysis complies with the assumptions underpinning valid identification through the SGMM procedure. These assumptions are different and arguably less restrictive than those used in past related studies, which include the absence of fixed effects and unidirectional influence from the older sibling to the younger. Given that the SGMM assumptions hold, the estimator is fit to overcome the aforementioned obstacles to the identification of endogenous sibling effects.

The estimation results obtained on verbal and mathematics skills are consistent between

each other. I find that a child's test performance is increased by 0.35-0.40 standard points when the performance of the sibling is increased by one standard point. The effect is larger from older siblings to younger ones, but it is present and significant also from younger to older. The results also show support for the theories of cumulative dynamics in cognitive achievement, as current score achievements are revealed to positively depend on past ones.

References

A. Aizer and F. Cunha. The production of human capital: Endowments, investments and fertility. *NBER Working Paper No. 18429*, Sep 2012.

J.G. Altonji, S. Cattan, and I. Ware. Identifying sibling influence on teenage substance use. *NBER Working Paper No. 16508*, 2010.

P. Arcidiacono and S. Nicholson. Peer effects in medical school. *Journal of Public Economics*, 89 (2-3):327–350, Feb 2005.

M. Ardelt and L. Day. Parents, siblings, and peers: Close social relationships and adolescent deviance. *Journal of Early Adolescence*, 22(3):310–349, Aug 2002.

M. Arellano and S. Bond. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies*, 58:277–97, 1991.

M. Arellano and O. Bover. Another look at the instrumental variables estimation of error-components models. *Journal of Econometrics*, 68:29–51, 1995.

M. Arellano and B. Honore. Panel data models: Some recent developments. In J.J. Heckman and E. Leamer, editors, *Handbook of Econometrics*, volume 5, pages 3229–3296. Elsevier, 2001.

M. Azmitia and J. Hesser. Why siblings are important agents of cognitive development: A comparison of siblings and peers. *Child Development*, 64:430–444, 1993.

G.S. Becker and H.G. Lewis. On the interaction between the quantity and quality of children. *Journal of Political Economy*, 81(2):S279–S288, 1973.

G.S. Becker and N. Tomes. Human capital and the rise and fall of families. *Journal of Labor Economics*, 4(3, p.II):1–39, Jul 1986.

J.R. Betts and A. Zau. Peer groups and academic achievement: Panel evidence from administrative data. Unpublished manuscript, 2002.

R. Blundell and S. Bond. Initial conditions and momen restrictions in dynamic panel data models. *Journal of Econometrics*, 87:11–143, 1998.

A.E. Boardman and R.J. Murnane. Using panel data to improve estimates of the determinants of educational achievement. *Sociology of Education*, 52(2):113–121, Apr 1979.

S.E. Carrell, R.L. Fullerton, and J.E. West. Does your cohort matter? Measuring peer effects in college achievement. *Journal of Labor Economics*, 27(3):439–464, 2009.

K. Coneus, M. Laucht, and K. Reuss. The role of parental investments for cognitive and noncognitive skill formation - Evidence for the first 11 years of life. *ZEW Discussion Paper No. 10-028*, 2010.

F. Cunha and J.J. Heckman. The technology of skill formation. *American Economic Review*, 97 (2):31–47, May 2007.

F. Cunha and J.J. Heckman. Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources*, 43(4):738–782, 2008.

F. Cunha and J.J. Heckman. Investing in our young people. *Rivista Internazionale di Scienze Sociali*, 117(3):387–418, 2009.

F. Cunha, J.J. Heckman, and S.M. Schennach. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3):883–931, 2010.

J. Currie and D. Thomas. Does head start make a difference? *The American Economic Review*, 85(3):341–364, 1995.

E. Garces, D. Thomas, and J. Currie. Longer-term effects of head start. *American Economic Review*, 92(4):999–1012, Sep 2002.

L. Green, A.F. Fry, and J. Myerson. Discounting of delayed rewards: A life-span comparison. *Psychological Science*, 5(33), 1994.

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

W.W. Hartup. Social relationships and their developmental significance. *American Psychologist*, 44(2):120–126, Feb 1989.

J.J. Heckman. Policies to foster human capital. *Research in Economics*, 54(1):3–56, Mar 2000.

J.J. Heckman. Skill formation and the economics of investing in disadvantaged children. *Science*, 312(5782):1900–1902, Jun 2006.

D. Holtz-Eakin, W. Newey, and H.S. Rosen. Estimating vector autoregressions with panel data. *Econometrica*, 56:1371–1395, 1988.

C. Hoxby. Peer effects in the classroom: Learning from gender and race variation. *NBER Working Paper No. 7867*, 2000.

C.F. Manski. Identification of endogenous social effects: The reflection problem. *Review of Economic Studies*, 60:531–542, 1993.

F.L. Mott. The utility of the HOME-SF scale for child development research in a large national longitudinal survey: The National Longitudinal Survey of Youth 1979 Cohort. *Parenting: Science and Practice*, 4(2-3):259–270, 2004.

S. Nickell. Biases in dynamic models with fixed effects. *Econometrica*, 49:1399–1416, 1981.

G.S. Oettinger. Sibling similarity in high school graduation outcomes: Causal interdependency or unobserved heterogeneity? *Southern Economic Journal*, 66(3):631–648, Jan. 2000.

L. Ouyang. Sibling effects on teen risky behaviors. Unpublished manuscript, Nov. 2004.

E. Patacchini, E. Rainone, and Y. Zenou. Dynamic aspects of teenage friendships and educational attainment. *CEPR Discussion Papers - n.8223*, 2011.

D. A. Powers and J. Cherng-tay Hsueh. Sibling models of socioeconomic effects on the timing of first premarital birth sibling models of socioeconomic effects on the timing of first premarital birth. *Demography*, 34(4), Nov. 1997.

D. Robertson and J. Symons. Do peer groups matter? peer group versus schooling effects on academic attainment. *Economica*, 70(277):31–53, Feb 2003.

D.M. Roodman. How to do xtabond2: An introduction to "Difference" and "System" gmm in stata. *Stata Journal*, 9(1):86–136, Mar 2009a.

D.M. Roodman. A note on the theme of too many instruments. *Oxford Bulletin of Economics and Statistics*, 71(1):135–158, 2009b.

B. Sacerdote. Peer effects with random assignment: Results for Darthmouth roommates. *Quarterly Journal of Economics*, 116(2):681–704, May 2001.

B. Sacerdote. Peer effects in education: How might they work, how big are they and how much do we know thus far? In *Handbook of Economics of Education*, volume 3, pages 249–277. Elsevier, 2011.

P. Todd and K.I. Wolpin. The production of cognitive achievement in children: Home, school and racial test score gaps. *Journal of Human Capital*, 1(1):91–136, 2007.

J.L. Vigdor and T.S. Nechyba. Peer effects in North Carolina public schools. In L. Woessman and P.E. Peterson, editors, *Schools and the Equal Opportunity Problem*, pages 73–101. The MIT Press, 2007.

R.W. Zimmer and E.F. Toma. Peer effects in public schools across countries. *Journal of Policy Analysis and Management*, 19:75–92, 2000.

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

A. Zuppann. Children's cognitive abilities and intrahousehold parental investment. *Department of Economics, University of Houston Working Paper Series*, Mar 2013.

A Appendix

A.1 Two-Children Families Only

Table 4: Verbal Skills, 2-children families

	SGMM	Fixed Effects	Pooled OLS
Lagged score	0.110*	-0.414***	0.393***
	(0.05)	(0.05)	(0.04)
Sib.score	0.485***	0.151	0.177***
	(0.06)	(0.11)	(0.03)
Sibsc.*Agegap	-0.001*	-0.002	-0.000
	(0.00)	(0.00)	(0.00)
Time dummies	Yes	Yes	Yes
Demographics	Yes	Yes	Yes
Family vars	Yes	Yes	Yes
Health, HOME	Yes	Yes	Yes
SS	1,101	1,101	1,101
Fstat p-value	0.000		
N. of instruments	329		
Hansen overid	300.267		
(pval)	(0.566)		
Diff.H, levels	79.291		
(p-val)	(0.345)		
Diff.H, transformed	240.021		
(p-value)	(0.542)		
AR(2) in fd	-0.949		
(p-value)	(0.343)		

Table 5: Math Skills, 2-children families

	SGMM	Fixed Effects	Pooled OLS
Lagged score	0.196***	-0.276***	0.421***
	(0.04)	(0.03)	(0.02)
Sib.score	0.317***	0.068	0.187***
	(0.06)	(0.08)	(0.02)
Sibsc.*Agegap	-0.001	0.001	-0.000
	(0.00)	(0.00)	(0.00)
Time dummies	Yes	Yes	Yes
Demographics	Yes	Yes	Yes
Family vars	Yes	Yes	Yes
Health, HOME	Yes	Yes	Yes
SS	1,302	1,302	1,302
Fstat p-value	0.000		
N. of instruments	334		
Hansen overid	298.509		
(pval)	(0.655)		
Diff.H, levels	79.947		
(p-val)	(0.327)		
Diff.H, transformed	241.436		
(p-value)	(0.605)		
AR(2) in fd	0.064		
(p-value)	(0.949)		

Chapter 2
Cities Drifting Apart:
Heterogeneous Outcomes of Decentralizing Public Education

Abstract

Looking at the decentralized provision of public education in a middle income country, this paper estimates the impact of local autonomy on service quality, finding large heterogeneity in the effect across different levels of local development. Colombian municipalities were assigned to administer their public education service autonomously solely on the basis of whether they exceeded the 100 thousand inhabitants threshold. Exploiting this discontinuity, I estimate the impact that autonomy has had on student test scores across municipalities, using a regression discontinuity design and fixed-effects regression on a discontinuity sample. I find a test score gap arising between autonomous municipalities in the top quartile and those in the bottom quartile of the development range, in a trend that reinforces over time. From analysis of detailed municipal balance sheet data, I show that the autonomous high-developed municipalities invest in education more than the ad hoc transfers they receive, supplementing these with own financial resources. Indicators of municipal administration quality also show significant differences between the two groups of cities, helping to explain the education outcome patterns.

1 Introduction

Decentralization of public service provision has been at the top of policy agendas in numerous countries over the last decades, involving services such as education, health, public transport and the supply of energy, water and sewerage systems. In developing and middle-income countries, responsibilities are often handled from a central or regional level down to municipalities[1]. Such reform are expected to yield welfare benefits through better local preference matching, higher governor accountability and increases in the efficiency of service delivery[2]. Welfare losses may instead derive from inadequate management skills of local authorities, increases in administrative and coordination costs, corruption among local bureaucrats or local elites resource capture[3]. These positive and negative repercussions may materialize in different proportions across different regions in the reforming country, giving rise or exacerbating regional inequalities. In this paper I show that entrusting Colombian municipalities with managerial autonomy over local public education has yielded heterogeneous results on local educational outcomes, depending on the level of municipal development at the time of the responsibility takeover.

In this empirical analysis I benefit of an unusually clean decentralization criterion: autonomy over the education service was assigned to cities solely depending on whether they exceeded the 100 thousand inhabitants threshold. This decision rule relieves the analysis from the issues that typically hinder identification of the effects of higher autonomy: non-random selection into autonomy by local authorities, and other nationwide phenomena occurring along with decentralization. In this way, this study introduces innovation to existing literature on the topic. A second valuable aspect of studying the Colombian case is that it yields insight into a context in which decentralization was purely administrative, and not mingled, as it often happens, with changes on the fiscal or political front: managerial authority was transferred, but local taxation and local representation were left unchanged.

Using panel data on standardized student test scores over a period of 10 years after the reform, I show that higher autonomy has proven beneficial for highly developed municipalities, but not so for less developed municipalities. Average test scores in high-developed autonomous municipalities have started to significantly exceed those of their non-autonomous

[1] Recent examples are the experiences of Chile, Argentina, Bolivia, Brazil and Colombia in Latin America; India, Thailand, Vietnam and the Philippines in Southeast Asia; South Africa, Senegal, Ethiopia and Uganda in Africa; Ukraine, Serbia and Bulgaria in Eastern Europe

[2] On informational advantage and heterogeneity in preferences, see seminal work by Musgrave [1959] and Oates [1972]. On accountability, monitoring and elections see Crook and Manor [1998], Manor [1999] and Blair [2000]

[3] Administrative costs are addressed in Breton and Scott [1978] and Panizza [2004]. Corruption and local elites capture are extensively discussed by Bardhan and Mookherjee [2000, 2002, 2005, 2006].

counterparts, in magnitudes that are growing over time. Low-developed autonomous municipalities instead appear to be progressively losing terrain with respect to their non-autonomous counterparts, even though effects are smaller and not always statistically significant. Both test score gains and losses seem to be larger for students with a more advantaged socioeconomic background: they gain more in highly developed cities but lose more in low developed ones, increasing score dispersion in the former group of cities and decreasing it in the latter.

In the second part of the analysis, exploring municipal balance sheet data, I show that high and low developed municipalities implement different spending decisions. Autonomous municipalities of the upper development quartile invest on local education not only the ad-hoc transfers they receive from the government, but also add own financial resources to their per-pupil education budget. Municipalities in the lowest development quartile only spend the education resources they receive from the central government, or somewhat less. The two groups also show significant differences in terms of municipal management and law compliance indicators, in directions that are consistent with the results on student test scores.

The three ways in which this paper adds to existing work in the field are its quasi-experimental estimation setup, the ability to focus on solely administrative power shifts, and the provision of suggestive evidence on channels that drive the heterogeneity in outcomes across local development levels. Implications of the findings may represent relevant references for future public service decentralization reforms to be implemented in low and middle income contexts similar to the Colombian one, especially in presence of significant subnational heterogeneity in development levels and local wealth.

2 Selected Literature

Heterogeneity in the effects of decentralization is modeled by Bardhan and Mookherjee [2000, 2002, 2005, 2006], who show how the combination of strong local elites and weak local institutions implies decentralization to yield under-provision of services to the local poor. Channels for diversity of impacts across places and people are illustrated also in the reviews by Kaiser [2006] and, with a special focus on developing countries, by Juetting et al. [2005]. These reviews and the vast majority of empirical literature fail to establish any clear link between decentralization and poverty reduction, and document higher advantages for the rich with respect to the poor in decentralized contexts. Some studies describe correlations between indicators of local welfare and the spending decisions of local politicians, but do not establish causal relationships between the two. Reinikka and Svensson [2004]

find that decentralized school grants in Uganda were subject to local elite capture, but less so in better-off communities. Local governments are found to be more responsive to citizen's needs when the electorate is more informed and when better institutions are in place in studies by Besley and Burgess [2002] on India and Ferraz and Finan [2011] on Brazil. Faguet and Sanchez [2008, 2014] look at Colombian municipalities' balance sheet data and construct original aggregate measures of decentralization, then finding negative association between dependence on central government transfers and expenditure on education, and positive association with public school enrollment rates. There are studies that aim at isolating causal effects of decentralization processes at different levels of local development, but focusing on fiscal decentralization or at contexts in which administrative decentralization came along with important fiscal and political changes. Hammond and Tosun [2011] apply a spatial error model on the US and find that fiscal decentralization (as proxied by government fragmentation) led to gains in employment and economic growth for metropolitan counties but insignificant to negative impacts for non-metropolitan ones. The fixed effects analysis by Zhang [2006] shows that fiscal decentralization in China has promoted regional inequality, mainly due to inequalities in tax bases and thus in fiscal burden, and in the development of nonfarm activities between jurisdictions. Contrary to mainstream findings, Faguet [2004] finds that after a large fiscal and political decentralization process the poor and marginalized communities of Bolivia benefited and adapted their expenditure structure to local needs. Closest to this paper in terms of reform context analyzed and in terms of outcomes looked at is Galiani et al. [2008], who show that transferring a number of Argentinian schools from a central to a provincial management yielded positive results in terms of test scores only for schools located in non-poor municipalities. This paper differs from the study on Argentina in scale of the reform (transferring to local authorities some additional schools versus the whole education service), in the level of government being looked at (regional versus municipal), and in the availability of a quasi-experimental setup for Colombia but not for Argentina[4].

3 Decentralization in Colombia and the 2001 Reform

Starting in the 1980s, Colombia has been undergoing a progressive decentralization process involving governance and administration, fiscal structure, and the delivery of public services; various authors have looked at the outcomes of these gradual processes, some in a

[4]"The transfer schedule was determined through bilateral negotiations between the federal government and each province" (Galiani et al., 2008, sec.3§3)

qualitative and some in a quantitative fashion[5]. The reform in 2001 kept the political and fiscal scenarios unchanged and enforced administrative decentralization[6], reallocating local authorities' responsibilities towards the delivery of public services[7].

3.1 Pre-reform context and reform motivations

Colombia is currently structured into local authorities as follows: there are thirty-two departments[8], 1,118 municipalities located within departments and four special districts (see maps in Section A.1 in the Appendix). Local authorities enjoy decisional and spending autonomy over a wide range of matters, although the necessary financial resources chiefly consist of central government transfers deriving from national tax revenues[9]. Central government transfers have historically been accounting for around 90% of the total education expenditure (nationwide average), and the remaining 10% is contributed by local authorities, with some local variability in these figures (Borjas and Acosta, 2000, p.6; Iregui B. et al., 2006, p.31;Santa Maria S. et al., 2009, pp.19-20). Up to the 2001 reform, the law had departments and municipalities jointly in charge of public education, entitled to hire personnel and invest in infrastructure and equipment[10]; as a result the division of responsibilities over the management of public education was vague and far from transparent (Borjas and Acosta [2000]). De facto, being the direct recipients of the bulk of education transfers[11], departments were the primary players on the education sector[12]. The elimina-

[5]Keeping the focus on education outcomes, Borjas and Acosta [2000], Vergara and Simpson [2001] and Caballero [2006] comprehensively illustrate dynamics and descriptive trends of decentralizing the public education system over the nineties, agreeing on generally undistinguished results.

[6]Sometimes this type of administrative decentralization is labeled as 'devolution' in literature, referring to situations in which the activities of subnational units of government are substantially outside the direct control of the central government [Rondinelli et al., 1983].

[7]Educational outcomes of the 2001 reform are explored in the descriptive Colombian central bank report by Lonzano et al. [2007], who conclude that the post-reform years have witnessed progress in attendance rates but disappointing results in terms of quality and efficiency. Also Cortés [2010] focuses on the 2001 reform, uses enrollment data up to 2006 and compares municipalities who gained education autonomy to the remaining, finding that the former significantly increased enrollments of publicly subsidized pupils into private schools.

[8]These represent the regional level, equivalent to "states" in the US, or "provinces" in Argentina.

[9]Colombia is considered among the administratively most decentralized countries in Latin America, but is fiscally very centralized (Alesina et al. [2000];Toro [2006]).

[10]Law 60 / 1993 (distributing competencies across levels of government and assigning resources accordingly), Law 115 / 1994 (the 'comprehensive education act'), and respective follow-up decrees.

[11]See Table 1

[12]For example, departmental payrolls included 85-90% of all public school teachers [Corte Constitucional, 1997, par.16], and departments decided on their allocation across municipalities. Municipalities were then responsible for allocating teachers across schools within their territory, and hired the remaining 10-15% that were not on departmental payrolls [Gómez et al., 2001]. Departments also had the final word on

tion of any responsibility overlap for the sake of accountability was one of the main goals of the 2001 reform; further goals were improving efficiency and reducing waste in the use of public resources, eliminating the yearly fluctuations in financial transfers, and updating some obsolete distribution criteria[13].

3.2 Reform content

Regarding the management of public education, the enactment of Law 715/2001 yielded the fundamental change of a clear-cut allocation of responsibility over the service to either municipalities or departments. Municipalities which counted 100 thousand or more inhabitants in the year 2002 became "certified in education" (certified municipalities, from now onwards), meaning responsible for the public education service on their territories. The education transfers from the central government, which are assigned on a per-pupil base, started to flow into their treasuries. Municipalities with fewer than 100 thousand inhabitants were not certified, and their public education is run by the departments they belong to. The next subsection further clarifies the concept of autonomy and discusses the shift in responsibilities. The forty municipalities certified in 2001 account for around one third of Colombia's population and pupil share; their size ranges from 105 thousand to over 2 million inhabitants[14][15].

The 2001 reform affected not only the education service, but also the provision of healthcare, water and sewerage and other smaller public services. Nonetheless, it was only for the education sector that this reform separated municipalities into autonomous and not, and used the 100 thousand inhabitants rule. Another task performed by the 2001 reform was updating the formulas used by the central government to compute financial transfers financing local public services; section 3.2.1 below provides further details and discussion on this aspect.

education proposals by municipalities, as these had to be taken in accordance with departments and under their supervision (Law 60 / 1993). Also see the DDTS [2004] report, p.6.

[13]For the official document motivating the reform, see: "Exposición de motivos 715 de 2001 Nivel Nacional", Congreso de Colombia, Gaceta del Congreso 294 de 2000. For further discussions of this matter see among others Sarmiento and Vargas, 1997; Alesina et al. [2000]; Borjas and Acosta [2000]; Vergara and Simpson [2001] and the technical report by DNP [2002].

[14]See their locations in panel c), Section A.1 Appendix.

[15]The reform provided for a transition period of two years (2002 and 2003), during which certified local authorities took over the school infrastructure, started the effective management of the service, and had the opportunity to reorganize staffing plans on their territories. During these two years temporary transfer amounts were set, and from 2004 onwards the new transfer system became fully operational.

3.2.1 Local authorities' competencies and transfers before and after the reform

Table 1 summarizes competencies of local authorities before and after the 2001 reform, and indicates percentages of education transfers flowing to their treasuries.

As illustrated in the table, the reform left the role of the central government unchanged but polarized both financial transfers and managerial responsibilities of local authorities. From receiving a narrow share of transfers and being subject to departmental supervision, certified municipalities transitioned into a situation of full managerial and financial autonomy, while non-certified ones gave up their already limited powers to the respective departments[16]. For the rest of the analysis I will consider certified municipalities as "treated" by the decentralization reform and the non-certified counterparts as "untreated", since both figures and anecdotal evidence indicate that a truly substantial change in regime has happened for the former group but not for the latter. How a violation of this premise would affect the interpretation of empirical results is discussed in Section 5.3.

The reform also brought an adjustment in the allocation formulas of education resources to local authorities. In broad outlines, up to 2001 the vast majority of transfers were assigned based on number and seniority of teachers employed, with some adjustment based on number of inhabitants, local poverty and administrative efficiency. From 2002 onwards the allocation criteria were tilted towards a student headcount base, but with number of teachers still playing a key role, and again with some adjustment for local poverty and population density; these changes applied to transfers to all local authorities, certified and not[17]. Both before and after the reform, transfers are meant to be exclusively used for the service to which they are dedicated, administered in separate accounts and thus not fungible with respect to the remaining revenues and expenses of the local authority [MEN, 2003].

3.3 Further relevant aspects

3.3.1 Population and population cutoff

The population figures that were used for the 2001 reform were issued by the National Statistics Office (DANE). The counts were not prepared ad hoc for the reform but issued on

[16]With only a 3% of total funds still flowing to non-certified municipalities, with pre-set destination. These funds need to be spent entirely on school infrastructure and school material, according to departments' directions [MEN, 2003 ; DDTS, 2004, p.7; Law 715/2001, art.16].

[17]The transition to a transfer system giving more weight to student head-counts should have, if anything, favored municipalities characterized by low levels of local development,, as such areas have historically been disadvantaged in terms of teacher provision [Corte Constitucional, 1997, par.19]. Evidence on central government transfers having become more redistributive over time is presented in Table 15 in the Appendix.

Table 1: Education responsibilities and transfers by level of government

Central Government

Curriculum, teacher wages, general guidelines, transfer to local authorities

Local Authorities

Up to 2002 (Law 60/1993)		**From 2002 onwards (Law 715/2001)**	
		Certified Municipalities	
Transfers:	84% to department 16% to municipality	Transfers:	100% to municipality
Teacher hiring, training and placement; School infrastructure and materials; School transport and any extra education programs	Departments and municipalities, under departments' supervision	Teacher hiring, training and placement; School infrastructure and materials; School transport and any extra education programmes	Municipality only
		Non-Certified Municipalities	
		Transfers:	97% to department 3% to municipality
		Teacher hiring, training and placement; School infrastructure and materials; School transport and any extra education programmes	Department only (maintenance duties for municipality)

Author's illustration, based on Laws 60/1993, 115/1994 and 715/2001 (República de Colombia); Borjas and Acosta [2000]; DNP[2002]. Percentages are author's derivation: pre-reform is based on 2001 data in DNP[2002], p.16; post-reform is based on 2004 data in DNP[2004a] and DNP[2004b]. Percentages for departments include the four special districts.

the occasion of the 1993 general census, as forward projections. Certification was assigned to those municipalities that according to the projections for the year 2002 were exceeding 100 thousand inhabitants. The cutoff was sharply implemented, and no exceptions were made in either direction; the way population figures arose allows us to set aside any potential suspicion of targeted count manipulation.

Beyond its use in the 2001 reform, the 100 thousand inhabitant cutoff does not play any significant role in Colombia's legislation and it is never used in other matters involving municipal public service provision. The cutoff appears in the municipal classification that is performed every fiscal year by the central government. Current inhabitant count and current revenues are jointly used for the classification, and, given appropriate current revenues, 100 thousand inhabitants may represent the lower bound for a 'first' category city[18]. This categorization is updated every year and is used for setting limits to salaries of the mayor, of council members and administrative staff and limits to general administrative expenditures; the changes are minor across category thresholds. The smaller municipalities (categories fourth to sixth) are entitled to special support transfers.

Figure 7 in the Appendix shows smoothness of various municipal characteristics around 100 thousand inhabitants. Most notably, student test scores just before the reform (in 2001) do not exhibit discontinuities at the 2002 treatment cutoff. If we believe test scores to reflect a range of underlying municipal characteristics, especially those affecting education outcomes, the lack of discontinuities in pre-reform scores injects further confidence on the absence of any relevant transitions occurring at 100 thousand inhabitants. Further falsification and robustness tests on these aspects are performed along the empirical analysis.

3.3.2 Districts and special municipalities

Districts are local authorities whose nature is mixed between departments and municipalities. Already before 2002, districts were drawing the totality of their financial entitlements for education directly into their treasuries and managing them autonomously. The four Colombian districts are Bogotá, Barranquilla, Cartagena and Santa Marta, and they are excluded from the analysis.

There are two municipalities[19] whose freedoms on local education policy had been formally enhanced in 1999-2000, even though the substantial implications of the procedure remained

[18] Law 136 / 1994 and Law 617 / 2000. The seven categories and their relative inhabitant cutoffs are: Special (500,001 or above), First (100,001 to 500,000), Second (50,001 to 100,000), Third (30,001 to 50,000), Fourth (20,001 to 30,000), Fifth (10,001 to 20,000) and Sixth (10,000 or below).

[19] The municipalities of Armenia (department of Quindio) and San Juan de Pasto (department of Nariño).

unclear. I exclude these two cities from the analysis.

4 Data

4.1 Test scores

Colombia has a long running tradition of standardized testing in public schools; ICFES is the government agency in charge of conducting and assessing the tests across the whole country. The most complete and frequent test score data refers to the Saber11 examination, which is administered to all students completing high school[20], and which is widely accepted as the reference examination to evaluate the quality of Colombian secondary education. Saber11 evaluates a range of school subjects; test scores range from 0 to 100 in each subject and are standardized by subject at the national level, to a mean of 50 and a standard deviation of 10. This is, each student's score is informative about his/her position relative to the national average in that subject. Individual-level Saber11 test scores are made available by ICFES for the years 2000 to 2012, with information about the school and municipality to which each student belongs, and some information student background.

4.2 Municipal Development Measures

The development level of Colombian municipalities is being evaluated periodically by government agencies: relevant data is collected by the National Statistics Office (DANE) and the summative indicators are calculated by the National Planning Department (DNP). Up to the year 2013, the most informative and widely used indicator on local development was the Municipal Development Index (hereafter MDI[21]). The MDI ranges from 0 to 100 and expresses a composite measure of municipal development; it considers 'social' or 'life quality' variables such as coverage of energy, water and sewerage systems, literacy rates and poverty ratios, and 'financial status' variables such as per capita tax revenue and public spending, and dependency on central government transfers; the higher index value, the better local development. I use the 2001 MDI index to measure the local development of municipalities at the time of the reform. As can be seen in Figures 1, 5 and 6, size and

[20] This is, students completing 11 years of schooling. The first 9 years are compulsory, the last 2 are optional.

[21] Translation from the original *Índice de Desarrollo Municipal (IDM).* Data on the index is provided for public use by the Colombian National Planning Department (*DNP - Departamento Nacional de Planeación*). A new "Overall Performance Index" (*Índice de Desempeño Integral (IDI))* has been issued starting in 2006 and has now replaced the IDM (2013 onwards).

local development level are overall positively correlated but with high variation at all size ranges. Municipalities which obtained certification in 2002 had MDI values ranging from 28 to 70; the empirical analysis will use the distribution of development of certified cities to determine development quartiles.

5 Empirical framework

The aim is to identify the impact of municipal autonomy over the management of local education on student test scores, especially looking out for heterogeneous patterns that the effect might display across different levels of local development. The next subsections first introduce and then discuss the two identification strategies adopted.

5.1 Sharp Regression Discontinuity Design

The fact that in 2001 certification was assigned solely on the basis of the 100 thousand municipal population cutoff sets the conditions for a sharp regression discontinuity (RD) design. The subsequent methodological summary draws on the excellent outlines by Imbens and Lemieux [2008] and Pettersson-Lidbom [2008], to which I refer for a more detailed discussion of the RD methodology. Consider the equation (I) $Y_i = \alpha + \ \tau \ C_i + \epsilon_i$, where Y_i represents average test scores in municipality i and C_i is a dummy signaling whether municipality i was certified in education ($C_i = 1$) or not ($C_i = 0$). The consistent estimation of the treatment effect τ is hindered by the fact that most likely certification C_i is correlated with other municipal characteristics enclosed in ϵ_i. In our setup though, we know that the sole assignment rule for C_i was population count P_i, and specifically whether P_i exceeded $c = 100\,000$ or not, such that $C_i = \mathbb{1}\{P_i > c\}$ where $\mathbb{1}\{.\}$ is the indicator function. In this case we have that conditioning on population P_i will remove any correlation between C_i and ϵ_i, so that treatment C_i is as good as randomly assigned conditional on P_i[22]. Thus the ideal specification of a 'control function' $h(P_i)$ is such that its insertion into our equation (I) will completely purge it from any dependence between C_i and ϵ_i [Heckman and Robb, 1985]. In practice it is difficult to guess the ideal functional form of $h(P_i)$ and thus a common approach is to use flexible functions such as high order polynomials of P_i[23]. In

[22] Other ways to express this is saying that the 'conditional mean independence' or 'selection on observables' or 'unconfoundedness' assumption holds, $E[\epsilon_i|C_i, P_i] = E[\epsilon_i|P_i]$.

[23] This is sometimes referred to as 'global polynomial series' estimator. Other estimators of the treatment effect in a RD setting are kernel estimators, and estimators based on trimming data close to the boundary such as local linear regression or other nonparametric methods. In a recent working paper, Gelman and Imbens [2014] recommend to prefer the latter group to the global polynomial method I employ here. These methods are data thirsty and require high numbers of observations close to the boundary, a luxury that is

order to ensure that τ is capturing only the effect of the treatment, we also need all other characteristics of the observed units not to change discontinuously at the treatment cutoff. In our case, we need municipal characteristics other than certification in education not to exhibit 'jumps' at 100 thousand inhabitants, otherwise we would not know which part of the estimated effect on test scores is due to certification and which part is due to the other changes happening at 100 thousand inhabitants. Smoothness of municipal characteristics was discussed in section 3.3.1, and further evidence on pre-reform smoothness of test scores is given through the falsification tests in section A.4.

I estimate the model

$$Y_i = \alpha + \ \tau^{RD} \ C_i + f(P_i) + \epsilon_i \tag{1}$$

where $f(P_i)$ is approximated by a third oder polynomial in P_i, and interpret τ^{RD} as the average treatment effect of certification in education. I then introduce the municipal development variable D_i, expressing the MDI indicator illustrated in section 4.2. Heterogeneity across levels of local development can be explored either by applying 1 to different subsamples, or by introducing an interaction term between certification and development, obtaining

$$Y_i = \alpha + \tau_0^{RD} C_i + \tau_1^{RD} C_i \cdot D_i + \beta D_i + f(P_i) + \epsilon_i \tag{2}$$

where $\tau_0^{RD} + \tau_1^{RD} D_i$ can be interpreted as the average treatment effect of certification at development level D_i. Section 6 shows the estimation results of 1 on the full sample of municipalities and on four subsamples split by development of certified cities (upper half and lower half, highest quartile and lowest quartile), and the estimation results of 2.

Regression discontinuity designs are notably data demanding and rarely free of obstacles (more on this in the discussion section 5.3). The fact that only forty cities obtained certification in 2002 and that their population sizes are not all clustered at 100 thousand inhabitants poses difficulties in terms of available sample size and precision. I am able to reduce sampling variance by using ten post-reform years of data (2002-2012); including year fixed effects is not necessary in this setup as the test score outcomes are standardized each year according to the yearly national performance. Further discussion of the RD estimation is available in Section 5.3.

unavailable in this setting.

5.2 Fixed Effect Regression on a Discontinuity Sample

An alternative method for estimating the average treatment treatment effect of certification in education on student test scores is applying the fixed effects (FE) concept, which exploits over-time variation in the performance of municipalities that acquire certification. The basic FE model reads as follows:

$$Y_{it} = \alpha + \tau^{FE} C_{it} + \boldsymbol{M_i} + \mathbf{T_t} + \epsilon_{it} \tag{3}$$

where test scores in municipality i and year t, Y_{it}, are regressed on certification status C_{it}, which takes value 1 in years from 2002 onwards for municipalities who obtained certification, and is 0 otherwise. Municipality fixed effects are $\boldsymbol{M_i}$ and time fixed effects are $\boldsymbol{T_t}$; the effect of certification is captured by τ^{FE}.

I limit the sample to municipalities with a number of inhabitants close to the certification cutoff, both from the left and from the right, in order to avoid confounders such as municipal characteristics varying with population size to threaten identification. Following Angrist and Lavy [1999], I refer to this as our 'discontinuity sample'. We are thus estimating 3 on a sample of cities that are similar to each other in size, out of which some acquired education autonomy ($C = 1$) in 2002 and some did not.

For the main specification in the empirical analysis I use municipalities between 80 thousand and 130 thousand inhabitants - which results in a sample of thirty cities, eleven of which acquired certification in 2002 and nineteen did not[24], and whose population counts and development indices are illustrated in Figure 1 with dark bars and light bars respectively[25]. Table 2 shows some relevant summary statistics separately for certified and non-certified municipalities, and highlights the similarity of the two groups in terms of pre-reform characteristics - including pre-reform test score levels. The thirty cities in the discontinuity sample account for about 8.35% of the student population enrolled in primary and secondary school in 2012[26].

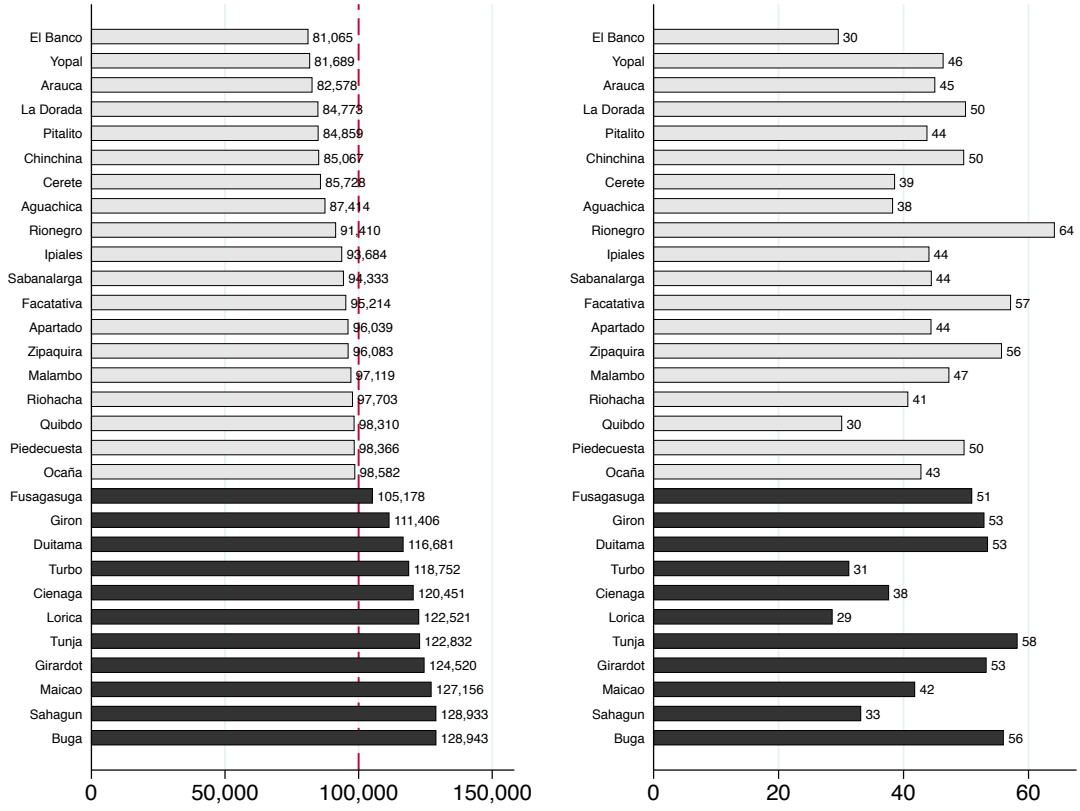
In order to pursue our goal of identifying heterogeneity in the effects of autonomy by levels of local development, model 3 is augmented with an interaction term between certification status C_{it} and development measure D_i, obtaining

[24]Results are robust to extending or restricting the sample; regressions on different samples, including the same sample used for the RD estimation, are presented in Table in the Appendix.

[25]Figure 6 in the Appendix shows the two distributions for a wider range of municipalities.

[26]Author's calculation, based on enrollment data provided by MEN (Ministry of Education).

Figure 1: Population and MDI distribution of the 30 municipalities around the inhabitant cutoff



$$Y_{it} = \alpha + \tau_0^{FE} C_{it} + \tau_1^{FE} C_{it} \cdot D_i + \boldsymbol{\gamma} \boldsymbol{M_i} + \boldsymbol{\delta} \mathbf{T_t} + \epsilon_{it} \tag{4}$$

where the effect of certification at development level D_i will be given by the estimates of $\tau_0^{FE} + \tau_1^{FE} D_i$. Section 6 shows results of this fixed effects approach.

5.3 Discussion of identification

As anticipated in section 5.1, the probably most salient difficulty of applying the RD methodology to the context of this paper is the limited sample size available, which exposes the analysis to both low power and potentially excessive small sample variation. In fact the confidence levels at which the null is being rejected in the result section are not particularly impressive[27]. Moreover, also cities distant from the treatment cutoff are used for estimation

[27]The limited amount of available data points also prescribes parsimony in the number of model parameters to estimate. The parsimonious regression model 1 is therefore chosen for the main specification

Table 2: Municipalities in the discontinuity sample (80,000 - 130,000 inhabitants)

	Certified (11)	Non certified (19)	Difference	
Population in 1992	99,998	73,182	26,816	***(1,820)
Population in 2002	120,670	91,043	29,627	***(692)
Population in 2012	127,756	112,305	15,451	***(3,705)
Municipal Development Index (MDI) 2001	40.99	40.59	0.40	(1.92)
Unsatisfied Basic Needs indicator (UBN) 1993	45.18	45.34	-0.16	(0.96)
Saber 11 Math score 2001	40.38	40.45	-0.07	(0.39)
Saber 11 Language score 2001	45.27	45.33	-0.06	(0.62)
Public primary school gross enrollment rates 2001	0.67	0.67	-0.00	(0.06)
Public secondary school gross enrollment rates 2001	0.61	0.61	0.00	(0.05)

Standard error of mean difference in parentheses; *p<0.10 **p<0.05 ***p<0.01

of the RD model[28], which relies on the assumption that the population polynomial $f(P_i)$ is able to 'control' for those municipal characteristics that vary with size and may confound the effect of autonomy on test scores. The less one needs to move away from the cutoff (the larger the sample close to the cutoff), the more likely it is that such confounders are properly eliminated. Despite these drawbacks, results display a fairly stable path across subsamples and are robust to different specifications (see section A.9.1).

As anticipated in Section 3.2.1, one may hold that the 2001 reform implied not only an increase in autonomy for the cities which obtained certification, but also some loss of autonomy for the cities which did not, leading to a certain extent of 'inverse treatment' in the control group[29]. In the case the reform had induced changes in the education trends of 'untreated' municipalities too, both the regression discontinuity and the fixed effects

instead of the RD approach that leaves two different sets of parameters on the two sides of the treatment cutoff (see Imbens and Lemieux [2008] for a discussion). Table 17 in the Appendix shows result of the latter method too.

[28] All cities between 10 and 500 thousand inhabitants.

[29] For the reasons and context explained in Sections 3.2 and 3.2.1, the author's assessment is that this scenario ought not to be excessively worried about.

analysis would be affected through alterations in their control groups. One would then need reject the interpretation of estimated results as the effects of an 'increase in autonomy' (or effects of decentralization), and rather look at results as the effects of 'an autonomy gap' (or effects of authority polarization). These interpretational issues are however confined to the background as one keeps in mind the primary objective of capturing heterogeneity in the effect of autonomy across different levels of the local development spectrum.

6 Results

6.1 Regression Discontinuity results

6.1.1 Baseline results

Table 3 shows estimation results for the regression discontinuity models (1) and (2), on Mathematics and Spanish student test scores on the 2002-2012 period. In all cases I have excluded from the analysis municipalities of special sizes, namely those below 10 thousand and above 500 thousand inhabitants (municipal categories "Special" and "Sixth" - see footnote 18). Column (1) of each panels report the outcome of model (1). The estimate of the average effect of certification in education on municipal test scores is close to zero. Columns (2) to (5) of each panel explore heterogeneities in the effect, considering different subsamples of the 2001 MDI distribution of certified cities: model (1) is applied to the lower and upper 50% development range and to the bottom and top 25% of the range. As anticipated in Section 4.2, development quartiles are constructed referring to the distribution of local development of *certified* cities. Figure 5 in the Appendix illustrates how in correspondence of lower levels of this distribution we find a larger number of smaller, non-certified cities. A larger 'control group' at low levels of development is what causes low-development sample sizes to be larger than the high-development ones throughout the analysis.

Looking at the result pattern, a test score gap appears to be opening between the most and the least developed autonomous cities. More precisely, high developed cities who become autonomous do better and low-developed cities do worse than their non-certified counterparts[30]. The effect on the high developed group is larger and more precisely estimated. Table 12 in the Appendix shows that in the pre-reform period (years 2000 and 2001 data) this pattern is not visible, and neither there is any evidence for differential score growth

[30]Keeping in mind that test scores are nationally standardized, the post-reform bifurcation could arise because certified municipalities change their performance and non-certified ones remain static, or the other way round, or a mix of both effects. Refer to the discussion in Section 5.3.

in the different development subsamples. In Section A.6 in the Appendix I perform formal tests on the difference between pre-reform and post-reform RD coefficients.

The magnitudes of the effects are sizable: negative 1.5 points (0.15 student standard deviations) for certified cities in the least developed quartile and positive 2 points (0.2 student standard deviations) for certified cities in the most developed quartile. The three panels in the first column of Figure 2 depicts these estimation results graphically.

Columns (6) in Table 3 show the estimation of model (2), where certification status is linearly interacted with the development percentile to which each municipality belongs, as an alternative way to capture heterogeneity in the effect. This second estimation approach confirms the pattern previously emerged: the effect of certification is increasing in MDI values, starting negative for low MDI values and becoming positive at higher ones.

6.1.2 Consolidation over time

After looking at the average effect on scores over the whole post-reform period 2002-2012, I will now concentrate on years further away from the reform date. Cohorts of high school students taking the Saber 11 exam in later years have been exposed to the reform for longer[31]; moreover one should allow for certified municipalities to gradually implement their medium and long-horizon education plans. Table 4 shows how the estimated effect on test scores grows in absolute values as model (1) is run on periods further and further away from the reform date (time periods starting in 2004, 2007 and 2010). Figure 3 uses point estimates and confidence intervals from Table 4 to illustrate how certified cities in the top 25% development range progressively increase their score gap with respect to their non-certified counterparts, while the opposite happens for cities in the bottom 25% development range. Looking at the last period (2010-2012), the point estimates have reached about a third of a student standard deviation into both directions. In Section A.5 of the Appendix I show analogous over-time changes in the effect estimations obtained using mutually exclusive year bins instead of progressively later time periods.

[31] In the spirit of the exercise performed by Galiani et al. [2008] in their paper on Argentinian school decentralization.

Table 3: Effect of certification on Saber11 test scores - Main RD estimation

(a) Mathematics

	(1) All	(2) Bot. 25%	(3) Bot. 50%	(4) Top 50%	(5) Top 25%	(6) Interact.
Certified	0.17	-1.58	0.40	0.73	2.20**	-2.37***
	(0.63)	(0.99)	(0.94)	(0.87)	(0.86)	(0.84)
Certif.*MDI'01						0.06***
						(0.02)
MDI '01						0.07***
						(0.01)
.						
F(Population)	Yes	Yes	Yes	Yes	Yes	Yes
N	7,572	6,536	7,100	472	275	7,561
R-sq.	0.013	0.003	0.003	0.011	0.050	0.084

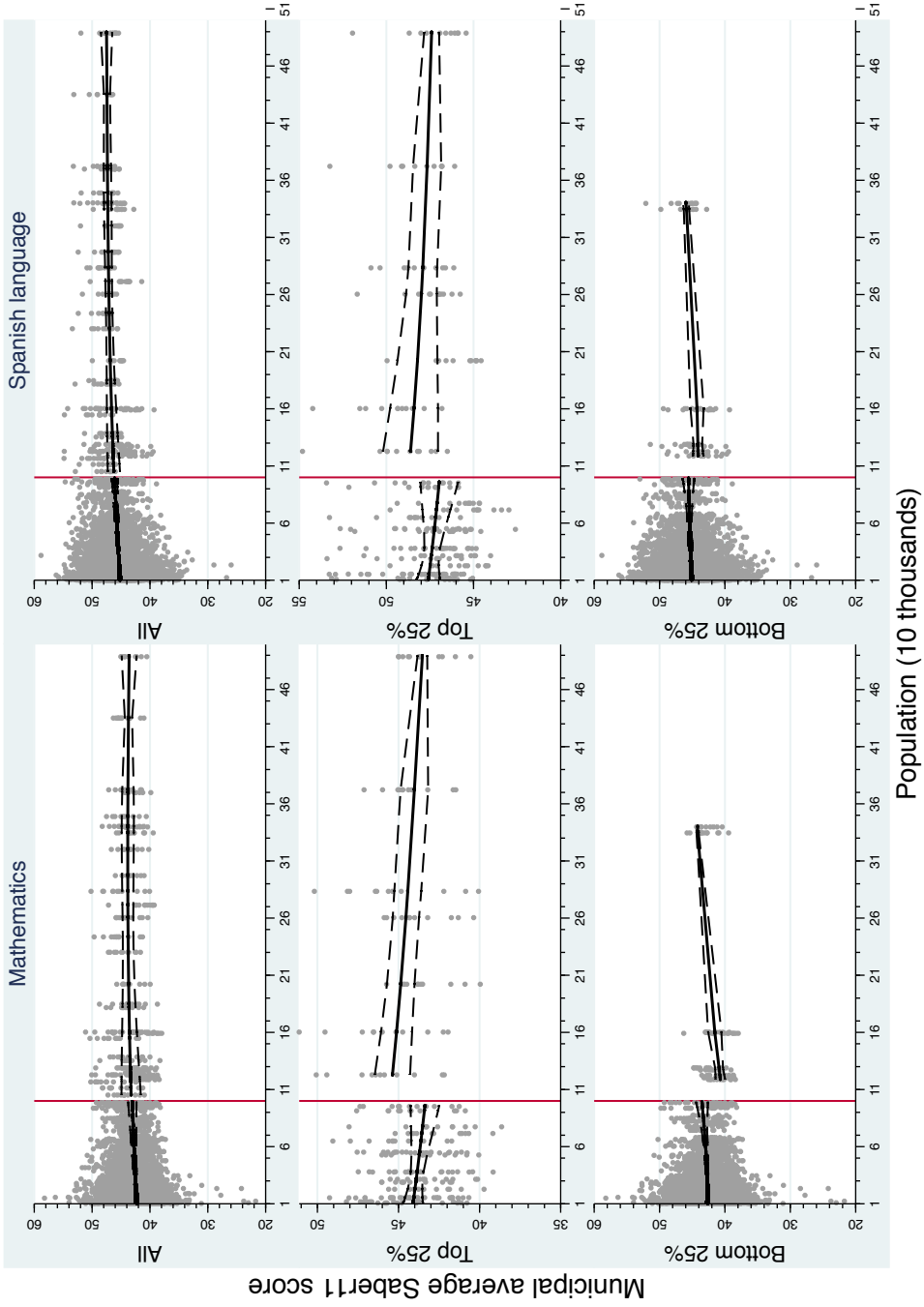
Standard errors clustered by municipality in parentheses
* p<.10 ** p<.05 *** p<.01

(b) Spanish Language

	(1) All	(2) Bot. 25%	(3) Bot. 50%	(4) Top 50%	(5) Top 25%	(6) Interact.
Certified	0.07	-1.55	0.32	0.52	1.81	-2.11**
	(0.66)	(1.00)	(0.94)	(0.90)	(1.14)	(0.96)
Certif.*MDI'01						0.05**
						(0.02)
MDI '01						0.09***
						(0.01)
.						
F(Population)	Yes	Yes	Yes	Yes	Yes	Yes
N	7,572	6,536	7,100	472	275	7,561
R-sq.	0.018	0.003	0.004	0.011	0.035	0.123

Standard errors clustered by municipality in parentheses
* p<.10 ** p<.05 *** p<.01

Figure 2: Certification on Saber 11 scores - Graphical results



Municipal averages of Saber 11 scores against population. Left column for Mathematics, right column for Spanish language. Solid lines are predicted scores, dashed lines are 95% confidence intervals on the prediction.

Table 4: Certification on Saber 11 test scores - progress over time

[Regression Discontinuity Estimation]

(a) Top 25% MDI '01

	Mathematics			Spanish Language		
	(1)	(2)	(3)	(4)	(5)	(6)
	Post 2004	Post 2007	Post 2010	Post 2004	Post 2007	Post 2010
Certified	2.37**	3.00***	3.80**	1.64	1.36	1.92
	(0.92)	(1.06)	(1.52)	(1.09)	(1.04)	(1.29)
.						
F(Population)	Yes	Yes	Yes	Yes	Yes	Yes
N	225	150	75	225	150	75
R-sq.	0.069	0.137	0.199	0.038	0.126	0.137

(b) Bottom 25% MDI '01

	Mathematics			Spanish Language		
	(1)	(2)	(3)	(4)	(5)	(6)
	Post 2004	Post 2007	Post 2010	Post 2004	Post 2007	Post 2010
Certified	-1.80	-2.23*	-3.17**	-1.47	-1.61	-2.03*
	(1.12)	(1.29)	(1.60)	(1.02)	(1.06)	(1.12)
.						
F(Population)	Yes	Yes	Yes	Yes	Yes	Yes
N	5,344	3,609	1,809	5,344	3,609	1,809
R-sq.	0.003	0.005	0.007	0.003	0.004	0.007

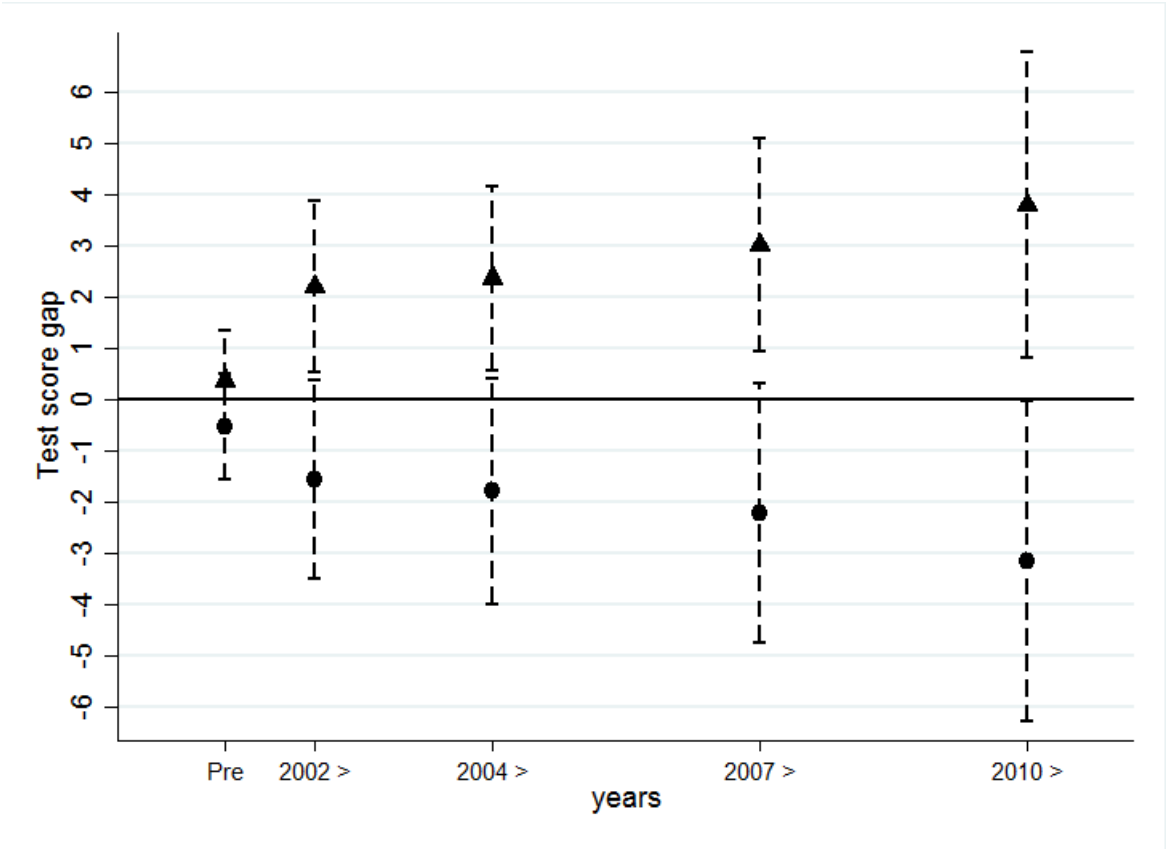
Standard errors clustered by municipality in parentheses
* p<.10 ** p<.05 *** p<.01

6.2 Fixed effects regression results

The results of the fixed effects identification strategy described in Section 5.2 are shown in Table 5, for both the basic model (3) and the specification that allows for effect heterogeneity in development (4). The sample is composed of the 30 municipalities closest to the inhabitant cutoff: 19 non-certified ones with more than 80 thousand inhabitants and 11 certified ones with less than 130 thousand inhabitants. The outcome variables are municipal test score averages in years 2000 to 2012. The first two columns of each panel refer to the basic model, showing OLS and municipality fixed effects estimations. The third and fourth column show OLS and fixed effects estimations of the main specification, using the MDI 2001 as a proxy for municipal development. The average effect of certification is estimated through the basic model as close to zero and statistically not significant. The

Figure 3: Effect of certification over time

[Regression Discontinuity Estimation]



RD estimations of the effect of certification on average Math test scores, for certified municipalities in the top 25% and the bottom 25% development range (triangles and circles series respectively). Capped spikes indicate 95% confidence intervals on point estimates.

model allowing for heterogeneity across development levels unveils the same pattern that was detected through the regression discontinuity identification: the point-estimated effect of autonomy goes from being between one and two points negative at low levels of development, crosses the zero threshold at a MDI level of around 45 and grows to reach a positive value of around two points at MDI levels of 70, as illustrated graphically in Figure 4. These magnitudes are very similar to the ones estimated with the RD technique for the lowest and highest development quartile[32].

[32]Even though they preserve the qualitative pattern, results for Language are again less imposing, as in the RD strategy.

Table 5: Effect of certification on Saber 11 test scores - Main FE estimation

	Mathematics				Spanish language			
	(1) OLS	(2) FE	(3) OLS	(4) FE	(5) OLS	(6) FE	(7) OLS	(8) FE
Certified	0.06	0.02	-3.02*	-3.67***	-0.05	-0.01	-1.93	-0.92
	(0.76)	(0.53)	(1.64)	(1.26)	(0.80)	(0.26)	(1.64)	(0.55)
Certif*MDI'01			0.07*	0.08**			0.04	0.02*
			(0.04)	(0.03)			(0.04)	(0.01)
MDI '01			0.09***				0.13***	
			(0.02)				(0.02)	
Time dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	389	389	389	389	389	389	389	389
N groups		30		30		30		30
R-sq.	0.39	0.67	0.58	0.68	0.38	0.77	0.63	0.77

Standard errors clustered by municipality in parentheses; * p<.10 ** p<.05 *** p<.01

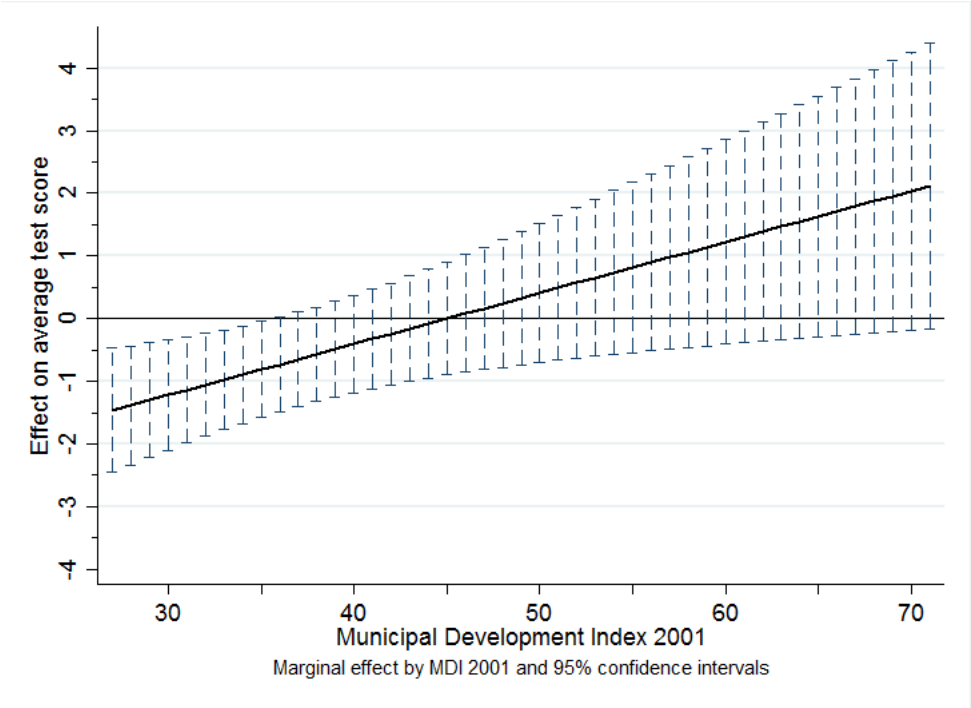


Figure 4: Effects of certification at different development levels (FE estimation)

6.3 Discussion of main results

Two different strategies have been employed to isolate the effect of local autonomy in the education service on quality of education, as measured by student test scores, at different levels of municipal development. Both techniques yielded the conclusion that cities at the upper end of the development range benefitted from the acquired autonomy while those at the lower end did not, and rather took loss on it, giving rise to a test score gap between the two groups that appears to be widening over time.

This section concludes with the reflection that the two identification strategies adopted rely on different assumptions but are shown to yield effect estimations which are qualitatively alike and quantitatively very similar, which represents a rewarding result on its own. Moreover, the regression discontinuity design estimates the effect of autonomy *for city sizes around 100 thousand*, while the fixe d effects methodology estimates the *average* effect of higher autonomy *across all certified cities* included in the sample[33]. By comparing the two sets of results, we thus learn that the effect magnitudes seem to be fairly stable across city sizes ranging from around 100 thousand to around 500 thousand inhabitants[34].

In the remaining sections of the paper the focus will lie upon cities in the highest and lowest quartiles of the certified development range, as these are the two groups on which significant reform effects have been identified. The goal will be exploring these effects in further detail and providing explanations for the test score dynamics found.

6.4 Compositional effects, migration and public-private education

This section addresses the question of whether the over-time changes in test scores purely reflect changes in student performance (an 'intensive margin' result), or whether the pool of test takers has also been changing as a result of the reform ('compositional effect' or 'extensive margin'). The pool of test takers may change if we observed responses to the reform such as selective migration (into or away from the newly autonomous municipalities), switching of students between public and private schools, or changes in student school dropout patterns. Tables 6 and 7 show regression discontinuity and fixed effects estimations respectively, of test taker characteristics on the municipal autonomy indicator (certification status). Characteristics being looked at are number of test takers, share of female students, share of students whose mother is low educated or high educated, and the share of students

[33] In the main results, cities between 80 and 130 thousand inhabitants; in the robustness checks, cities between 50 and 250 thousand and between 10 and 500 thousand inhabitants.

[34] Considering also the exercises of variation of the sample range that are performed in the robustness checks (Table 19).

who work while studying. In the RD results, the only statistically significant pattern we are able to spot is an apparent shift of high-educated families away from low-developed municipalities and into high-developed ones, which might suggest some degree of selective migration of the better educated families. Looking at the FE specifications though, at development ranges which are relevant to certified municipalities (MDI = 29 to 70 approx.), the magnitudes of these shifts are estimated close to zero. Patterns on low-educated mothers are never statistically significant and the share of working students does not exhibit changes. Overall it seems prudent to conclude that with the available data I am not able to pin down any clear and robust compositional effects on Saber 11 test candidates, as a consequence of the decentralization reform.

Table 6: Municipal certification and test taker characteristics (RD)

		(1) N. takers	(2) Female	(3) Low ME	(4) High ME	(5) Work
a) All		–127.91	–0.00	0.04	0.00	0.01
		(99.84)	(0.01)	(0.03)	(0.01)	(0.01)
b) Bottom 25%		–262.10	–0.00	0.10	–0.04**	0.01
		(163.87)	(0.00)	0.06	(0.03)	(0.03)
c) Top 25%		81.04	0.01	–0.04	0.10**	0.03
		(460.65)	(0.02)	(0.06)	(0.04)	(0.03)
d) All, interact.		185.83	0.02	0.03	–0.05*	0.06
		(185.86)	(0.03)	(0.06)	(0.02)	(0.04)
Certif*MDI‘01		–6.71	–0.00	0.00	0.00*	–0.00
		(4.79)	(0.00)	(0.00)	(0.00)	(0.00)
Mean Y (All)		321⋄	0.54	0.60	0.06	0.12
N [n. municip.]	a)	7,523 [698]	7,523 [698]	4,821 [697]	4,821 [697]	4,829 [698]
	b)	6,488 [603]	6,488 [603]	4,158 [602]	4,158 [602]	4,166 [603]
	c)	275 [25]	275 [25]	175 [25]	175 [25]	175 [25]
	d)	7,512 [697]	7,512 [697]	4,814 [696]	4,814 [696]	4,821 [697]

RD regressions of different outcome variables (in columns) on certification status and a third degree population polynomial. Cells show coefficient and standard errors on the certification regressor. The rows refer respectively to: a) All municipalities; b) municipalities in the bottom 25% devlopment range; c) municipalities in the top 25% development range; d) All municipalities, with main effect and development interaction term. SEs clustered by municipalities in parentheses. * p<0.10, ** p<0.05, *** p<0.01. ⋄ The mean of certified municipalities is $1,947$ test takers.

The absence of significant compositional changes in the pool of test candidates is in line

Table 7: Municipal certification and test taker characteristics (FE)

		(1) N. takers	(2) Female	(3) Low ME	(4) High ME	(5) Work
a) All		31.50	–0.00	–0.02	0.00	0.00
		(68.72)	(0.01)	(0.02)	(0.01)	(0.02)
b) All, interact.		128.70	–0.00	–0.04	–0.03*	–0.09
		(270.37)	(0.05)	(0.04)	(0.02)	(0.05)
Certif*MDI'01		–2.15	–0.00	0.00	0.00**	0.00
		(5.03)	(0.00)	(0.00)	(0.00)	(0.00)
Mean Y		971	0.55	0.49	0.08	0.11
N [n. municip.]	a)	389 [30]	389 [30]	270 [30]	270 [30]	270 [30]
	b)	389 [30]	389 [30]	270 [30]	270 [30]	270 [30]

Municipal FE regressions of different outcome variables (in columns) on certification status. Cells show coefficient and standard errors on the certification regressor. The rows refer respectively to: a) All municipalities; b) All municipalities, with main effect and development interaction term. SEs clustered by municipalities in parentheses. * p<0.10, ** p<0.05, *** p<0.01

with what one would expect after considering two basic facts about Colombia's school population[35]. The first is the significant and persistent gap between the quality of private and public education. Private schools score substantially and consistently better on standardized tests such as the national ICFES or international PISA results (Cerquera et al. [2000]; Barrera-Osorio et al., 2012; Gamboa and Waltenberg, 2011), have smaller class sizes[36] and are four times more likely to offer full-day school programs with respect to public schools, which instead see double the frequency of morning-only or late hours programs ([Bonilla Mejía, 2011]. In sum, it is fair to say that private education in Colombia, as well as in the rest of Latin America, is still a privilege restricted to well-off families (also see Gamboa and Waltenberg [2011]for a discussion). The implementation of a reform that shifts responsibility over public schools from the regional to the municipality level would not be expected to close the gap between public and private education, or to make public institutions significantly more attractive to well-off families. The second fact to keep in mind relates to the first: the family of the typical public-school student in Colombia is less likely to be informed about a decentralization reform occurring, to form strong predictions about its effects on educational quality, or to have the means and opportunity to migrate

[35]In addition to the fact that the publicity of this regime change on mass media has been very limited.

[36]Approximately 35 students per teacher in public schools and 25 students per teacher in private schools (averaged over the period 1998-2008). Author's own calculations using national statistics office education data (DANE C-600).

to a different municipality.

6.5 Heterogeneity across people

Along with the heterogeneity in effects across municipalities, heterogeneity across people within municipalities is a dimension that ought to be looked at. In this section I look at how autonomy has impacted the dispersion of test scores in cities of the high and low developed group. Moreover, I am interested in investigating whether students from different socioeconomic backgrounds have been differently affected by local autonomy on public education.

Using the self-reported background information on Saber11 test takers, I divide students by social status as proxied by their mother's education (ME): low mother education for compulsory education (9 school years) or less, high mother education is education beyond compulsory. Information on mother education is available for all years excluding 2005, 2006 and 2007[37]. Table 8 shows estimates obtained applying our main regression discontinuity model (1) on test score standard deviations (SD), and then on test scores of students belonging to the two social background categories.

The picture that emerges from these results is that students with higher social background seem to be more susceptible to changes in local autonomy: they gain more in high-developed cities and they lose more in low-developed ones, with respect to lower social background students.

7 Channels

7.1 Expenditure on education

In the pursuit of the reasons behind heterogeneous educational outcomes between high-developed and low-developed autonomous cities, the perhaps most straightforward starting point is expenditure choices. Using detailed balance sheet data[38] of municipalities in the highest and lowest development quartiles, in Table 9 I perform t-tests on the mean expenditures of the two groups over the post-reform period 2002-2012. Central government transfers that municipalities receive to finance education services (*SGP Educación*) are also

[37] Reason for the smaller sample sizes on the mother education (ME) specifications, with respect to the standard deviation (SD) specifications.

[38] "Ejecuciones municipales, formato largo". Reported yearly by municipalities to the government agency DNP (Departamento Nacional de Planeación). Source: Universidad de Los Andes, Bogotá, 2015.

Table 8: Effect of local autonomy on score dispersion and by social background

(a) Mathematics

	Bottom 25% cities			Top 25% cities		
	(1)	(2)	(3)	(4)	(5)	(6)
	score SD	Low ME	High ME	score SD	Low ME	High ME
Certified	-0.45*	-1.56	-3.05*	0.43*	2.28**	2.76***
	(0.24)	(1.26)	(1.80)	(0.23)	(0.97)	(0.91)
.						
F(Population)	Yes	Yes	Yes	Yes	Yes	Yes
N	6,524	4,128	3,932	275	175	175
R-sq.	0.002	0.002	0.002	0.007	0.046	0.029

(b) Spanish Language

	Bottom 25% cities			Top 25% cities		
	(1)	(2)	(3)	(4)	(5)	(6)
	score SD	Low ME	High ME	score SD	Low ME	High ME
Certified	-0.13	-1.34	-2.85**	0.02	1.47	1.66*
	(0.15)	(1.06)	(1.40)	(0.23)	(1.19)	(0.88)
.						
F(Population)	Yes	Yes	Yes	Yes	Yes	Yes
N	6,524	4,128	3,932	275	175	175
R-sq.	0.004	0.002	0.004	0.007	0.033	0.035

compared, as well as the resulting difference between spending and transfers. Recall that spending on education exceeding government transfers is covered by municipalities using their own resources, which are represented mainly by local tax and fees collection, and by capital gains[39].

What emerges from the comparisons performed in Table 9 is that the average per-pupil expenditure by municipalities in the highest development quartile is almost 23% higher than the one of municipalities in the lowest development quartile. Within education expenditure, the difference on personnel salaries is around 13%, up to 30% on school infrastructure and material and as much as 63% higher on other education expenses and programs[40]. The

[39]Examples of local tax and fees are the housing and land ownership tax, tax on gasoline consumption, traffic fines. Examples of capital gains are interests on municipal accounts and rents from municipal-owned infrastructure and land.

[40]Examples of the most frequent balance sheet items in this category are school transport, teacher for-

asymmetry in spending is not matched by any asymmetry in central government resources received. In fact, while the low-developed group appears to be spending on education barely as much as it receives in education transfers, the high-developed group is integrating transfers with own resources, for around 12% of their total education spending[41]. The differences in spending behavior uncovered through analysis of municipal balance data provides at least suggestive evidence towards the explanation of student test score dynamics previously identified.

Table 9: Per-pupil expenditure and transfers received

	All certified	Low 25%	High 25%	Δ H - L	Δ%
A) Education spending	1160.58	1021.83	1282.54	260.71***	22.66%
	(402.76)	(406.27)	(438.28)	(81.28)	
- Salaries	930.44	836.58	954.25	117.67**	13.14%
	(302.62)	(319.52)	(280.06)	(58.12)	
- Infrastr. and material	98.46	103.95	140.94	37.00*	30.21%
	(90.20)	(93.11)	(132.88)	(21.90)	
- Others	82.65	52.65	100.82	48.17***	62.77%
	(93.85)	(73.70)	(97.78)	(16.56)	
B) Education transfers	1153.41	1114.91	1134.53	19.61	1.74%
	(415.96)	(383.34)	(319.94)	(57.13)	
A) - B)	7.17	–93.08	148.01**		
	(34.23)	(66.73)	(54.26)		
N.obs (expenditure)	240	56	50	106	
N.obs (transfers)	345	88	70	158	
N. municipalities	35	9	7	16	

Table of mean annual per-pupil expenditures and central government transfers received (2002-2012, in thousands of Colombian pesos) and t-tests on the mean differences. Standard deviations in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

mation, planning and development of school information systems, investments in efficiency of the municipal education management authority, contracts with private institutions for additional education services.

[41]Table 15 in the Appendix shows that municipalities in the higher developed group have been enjoying higher availability of own resources both before and after the reform. The resource gap between the two groups has been significantly narrowed in post-reform years, mainly through compensatory transfers from the central government.

7.2 Administration indicators

Table 10 summarizes data on municipal evaluation processes that are carried out by the government on a yearly basis[42]. The 'Legal compliance rate' (*Índice de Cumplimiento de Requisitos Legales*) indicates the extent to which the municipality administration is found to adhere to the national norms in its use of government transfers, in all sectors of activity. The frequency of detection of accounting irregularities and illicit use of funds determine the rating received through this index [DNP, 2014]. The 'Managerial capabilities index' (*Índice de Capacidad Administrativa*) measures the extent to which the municipal administration appears suitable and prepared to perform its tasks thoroughly and to promote local development. The elements factoring into the index are the stability of managerial employees, the level of competency of clerks, the availability of suitable IT equipment and the level of automation of administrative processes [DNP, 2011]. The takeaway from Table 10 is that there are significant differences in these administration quality indicators between high-developed and low-developed municipalities, in the direction one would expect. This holds true for both municipalities that were certified (panel a)) and for smaller municipalities (panel b)). Adding to the evidence on expenditure behavior, the striking differences in these quality indicators provide further suggestive evidence towards the channels through which test score dynamics may have come about. Municipalities in the high-development range, characterized by higher availability of financial resources and a more pro-education spending policy, along with higher quality administration capabilities, were able to improve education quality on their territories with respect to the centralized management. The opposite has been true for cities in the low-development range, with fewer local resources and worse management skills.

7.3 Oaxaca-Blinder decomposition of test score differences

Having documented the existence of significant differences in education expenditure and administration indices between highly developed and low developed municipalities, it is desirable to gauge the extent to which such differences are able to explain the gap in student performance between the two groups of cities. Table 11 shows the results of the decomposition technique proposed by Oaxaca [1973] and Blinder [1973], which splits test score gaps into explained and unexplained components[43]. The results show that around

[42]DNP-DDTS (Departamento Nacional de Planeación - Dirección de Desarrollo Territorial Sostenible) is the government agency in charge of the study.

[43]This is sometimes known as the 'twofold' decomposition approach. See the excellent illustration by Jann [2008] for reference. The baseline coefficient vector is obtained by regressing student test scores on per-pupil expenditure, the two administration quality indices and the third-degree population polynomial

Table 10: Municipal administration indices

a) Certified municipalities	All certified	Low 25%	High 25%	Δ H - L
Managerial capabilities	67.99	46.01	79.12	33.11***
	(25.08)	(26.84)	(16.77)	(4.38)
Compliance rate	77.87	69.74	79.18	9.43*
	(22.95)	(29.45)	(21.50)	(5.01)
N.obs	245	63	49	112
N. municipalities		9	7	16
b) All municipalities	All	Low 25%	High 25%	Δ H - L
Managerial capabilities	63.67	61.98	80.39	18.41***
	(25.25)	(25.56)	(17.39)	(1.99)
Compliance rate	74.65	73.73	83.35	9.6***
	(22.23)	(22.66)	(16.52)	(1.77)
N.obs	4841	4187	168	4355
N. municipalities				

Table of evaluation indices (mean 2005-2012, scales 1-100) and t-tests on the mean differences. Standard deviations in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

one third of the gap is attributable to raw differences in expenditure and administration quality, with the latter playing the by far larger role. According to this decomposition, if highly developed cities recorded the same per-pupil education expenditure and the same administration quality indices as we find among low developed cities, their average test score performances would have been one point (or 10% of a student standard deviation) lower.

Differences in expenditure quantities and administration indices do not directly account for the remaining two thirds of the test score gap: it is *returns* to expenditure and especially to administration capacity that appear to differ significantly between highly- and low developed cities. These differences in returns are likely to capture differences between the two groups of cities which are not currently being accounted for.

on the sample of all certified cities. Baseline coefficients are then used to analyze the score differential betwen cities in the highest and the lowest development quartile. The differential is decomposed into a part that is explained by group *differences in predictors* ("Explained" component, or "quantity effect"), and a part that is attributed to *different returns to predictors* across the two groups ("Unexplained" component).

Table 11: Oaxaca decomposition of test score differences

	Mathematics			Spanish Language		
	2005	2007	2010	2005	2007	2010
Differential						
Prediction LD	42.067***	41.959***	41.665***	43.484***	43.659***	43.057***
	(0.42)	(0.46)	(0.64)	(0.37)	(0.36)	(0.45)
Prediction HD	45.600***	45.813***	46.621***	47.166***	47.250***	47.499***
	(0.46)	(0.49)	(0.63)	(0.40)	(0.41)	(0.46)
Difference	-3.533***	-3.854***	-4.956***	-3.682***	-3.591***	-4.442***
	(0.62)	(0.68)	(0.90)	(0.55)	(0.54)	(0.65)
Explained						
Expenditure	-0.174	-0.181	-0.180	-0.069	-0.045	-0.093
	(0.15)	(0.17)	(0.24)	(0.11)	(0.13)	(0.19)
Admin. cap.	-0.827***	-0.922***	-1.377**	-0.553***	-0.558**	-1.040**
	(0.28)	(0.33)	(0.62)	(0.21)	(0.23)	(0.42)
Legal req.	-0.037	-0.032	-0.008	-0.057	-0.041	-0.001
	(0.06)	(0.06)	(0.04)	(0.05)	(0.05)	(0.02)
F(Popul)	Yes	Yes	Yes	Yes	Yes	Yes
Total	-1.010**	-1.094**	-1.697**	-0.997**	-0.966**	-1.564***
	(0.42)	(0.49)	(0.77)	(0.39)	(0.38)	(0.56)
Unexplained						
Expenditure	-0.950	-0.205	2.139	1.520**	1.920	3.300*
	(0.84)	(0.79)	(1.65)	(0.72)	(1.34)	(1.85)
Admin. cap.	-2.564**	-2.491**	-4.728***	-2.907***	-2.803***	-4.120***
	(1.08)	(1.17)	(1.48)	(0.91)	(1.04)	(1.35)
Legal req.	-0.401	-0.370	0.562	-0.284	-0.100	0.709
	(0.68)	(0.83)	(1.49)	(0.54)	(0.50)	(1.04)
F(Popul)	Yes	Yes	Yes	Yes	Yes	Yes
Total	-2.522***	-2.760***	-3.259***	-2.685***	-2.625***	-2.878***
	(0.62)	(0.69)	(0.98)	(0.57)	(0.56)	(0.73)
N	110	94	47	110	94	47

Oaxaca-Blinder decomposition of test score gaps between highly and low developed municipalities, into explained and unexplained components. Columns indicate time periods from 2005, 2007 and 2010 onwards respectively. All models include population controls (third degree polynomial). Standard errors clustered by municipality in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

8 Conclusion

In this paper I have taken advantage of an unusually favorable reform setting to show that cities characterized by different levels of local development have reacted differently to higher

autonomy over the public education service. Levels of municipal development embody the wealth of local population and the amounts of own financial resources available. In the ten years following the handover of responsibilities, cities in the highest development quartile have significantly improved their student's test score performance with respect to non-autonomous counterparts; cities in the lowest development quartile instead display the opposite test score trend. The test score gaps are growing stronger over time. The high-developed group of cities invests in education more than the ad hoc financial transfers it receives from the central government, while cities in the lowest development quartile barely invest their financial allocation. The largest differences in investment shares between the two groups are on "other education programs", which include teacher formation, school transport, planning and maintenance of school information systems and investments into the efficiency of the local education authority itself. Spending differences are also found on the infrastructure and school material investments, and on school staff. Moreover, high developed municipalities perform significantly better on different administration quality indicators with respect to low developed cities, which suggests additional explanations for why, once given autonomy on the delivery of the public service, their results have started drifting apart.

The findings of this study sound a note of caution in the design of decentralization reforms in contexts in which subnational heterogeneity in wealth and development is an issue. When handing responsibilities in public service delivery to the local level, less advantaged localities may need additional training and support in order to avoid regional inequality to grow, and decentralization to backfire for segments of the population.

References

A. Alesina, A. Carrasquilla, and J.J. Echavarría. Decentralization in Colombia. *Fedesarrollo Working Paper Series*, 15, Aug 2000.

J. Angrist and V. Lavy. Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, 114(2):533–75, 1999.

P. Bardhan and D. Mookherjee. Capture and governance at local and national level. *American Economic Review [PaP]*, 90(2):135–39, 2000.

P. Bardhan and D. Mookherjee. Decentralization of governance and development. *Journal of Economic Perspectives*, 16(4):185–205, 2002.

P. Bardhan and D. Mookherjee. Decentralizing antipoverty program delivery in developing countries. *Journal of Public Economics*, 89(4):675–704, Apr 2005.

P. Bardhan and D. Mookherjee. Decentralisation and accountability in infrastructure delivery in developing countries. *The Economic Journal*, 116, Jan 2006.

Felipe Barrera-Osorio, Darío Maldonado, and Catherine Rodríguez. Calidad de la educación básica y media en Colombia: Diagnóstico y propuestas. *Serie Documentos de Trabajo Universidad del Rosario*, 126, October 2012.

T. Besley and R. Burgess. The political economy of government responsiveness: Evidence from India. *The Quarterly Journal of Economics*, 117(4):1415–1451, Nov 2002.

H. Blair. Participation and accountability at the periphery: Democratic local governance in six countries. *World Development*, 28(1):21–39, Jan 2000.

Alan S. Blinder. Wage discrimination: reduced form and structural estimates. *Journal of Human Resources*, pages 436–455, 1973.

Leonardo Bonilla Mejía. Doble jornada escolar y calidad de la educación en colombia. *Documentos de trabajo sobre Economía Regional - Banco de la República*, 143, April 2011.

G.J. Borjas and O.L. Acosta. Education reform in Colombia. *Fedesarrollo Working Paper Series*, 19, Aug 2000.

A. Breton and A. Scott. *The Economic Constitution of Federal States*. University of Toronto Press, Toronto, 1978.

P. Caballero, editor. *Foro "Descentralización y certificación en educación, experiencias y desafíos, el caso colombiano"*, Bogotá, D.C., Mar 2006. PREAL GDyÁ - Conversemos sobre educación.

Daniel Cerquera, Paula Jaramillo, and Natalia Salazar. La educación en Colombia: evolución y diagnostico. Boletines de divulgación económica, Departamento Nacional de Planeación, Bogotá, D.C., 2000.

C. Corte Constitucional. Sentencia su.559/97. Technical report, Corte Constitucional de Colombia, 1997.

D. Cortés. Do more decentralized local governments do better? An evaluation of the 2001 decentralization reform in Colombia. *Serie documentos de trabajo - Universidad del Rosario*, 84, Jun 2010.

R. Crook and J. Manor. *Democracy and Decentralisation in South Asia and West Africa: Participation, Accountability and Performance*. Cambridge University Press, Cambridge, 1998.

DNP DDTS. Evaluacion del sistema general del participaciones 2003. Technical report, Departamento Nacional de Planeación - Dirección de Desarrollo Territorial, Bogotá, D.C., Dec 2004.

DNP. Documento conpes social n.57. Technical report, Departamento Nacional de Planeación, Bogotá, D.C., Jan 2002.

DNP. Documento conpes social n.89. Technical report, Departamento Nacional de Planeación, Bogotá, D.C., Dec 2004a.

DNP. Documento conpes social n.77. Technical report, Departamento Nacional de Planeación, Bogotá,

D.C., Jan 2004b.

DNP. Evaluación del desempeño integral de los municipios. Technical report, Departamento Nacional de Planeación - Dirección de Desarrollo Territorial, Bogotá, D.C., 2011.

DNP. Orientaciones para realizar la evaluación del desempeño integral municipal, Vigencia 2013. Technical report, Departamento Nacional de Planeación - Dirección de Desarrollo Territorial, Bogotá, D.C., Apr 2014.

J.P. Faguet. Does decentralization increase responsiveness to local needs? Evidence from Bolivia. *Journal of Public Economics*, 88:867–94, 2004.

J.P. Faguet and F. Sanchez. Decentralization's effects on educational outcomes in Bolivia and Colombia. *World Development*, 36(7):1294–1316, 2008.

J.P. Faguet and F. Sanchez. Decentralization and access to social services in Colombia. *Public Choice*, 160 (1-2):227–249, Jul 2014.

C. Ferraz and F. Finan. Electoral accountability and corruption: Evidence from the audits of local governments. *American Economic Review*, 101(4):1274–1311, 2011.

S. Galiani, P. Gertler, and E. Schargrodsky. School decentralization: Helping the good get better, but leaving the poor behind. *Journal of Public Economics*, 92:2106–20, 2008.

Luis Fernando Gamboa and Fábio D. Waltenberg. Inequality of opportunity in educational achievement in latin america: Evidence from pisa 2006-2009. *ECINEQ - Society for the Study of Economic Inequality*, August 2011.

Andrew Gelman and Guido Imbens. Why high-order polynomials should not be used in regression discontinuity designs. *NBER Working Paper No. 20405*, Aug 2014.

A.S. Gómez, L.P. Tovar, and C. Alam. *Situación de la educación básica, media y superior en Colombia.* Casa Editorial El Tiempo - Fundación Corona - Fundación Antonio Restrepo Barco, 2001.

G.W. Hammond and M.S. Tosun. The impact of local decentralization on economic growth: Evidence from U.S. counties. *Journal of Regional Science*, 51(1):47–64, 2011.

J.J. Heckman and Richard Robb. Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics*, 30(1):239–267, 1985.

G. Imbens and G. Lemieux. Regression Discontinuity Designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635, Feb 2008.

A.M. Iregui B., L. Melo B., and J. Ramos. Evaluación y análisis de eficencia de la educación en colombia. Technical report, Banco de la Republica, Colombia, Bogotá, D.C., Feb 2006.

Ben Jann. The Blinder-Oaxaca decomposition for linear regression models. *The Stata Journal*, 8(4): 453–479, 2008.

J. Juetting, E. Corsi, C. Kauffmann, I. McDonnell, H. Osterreider, N. Pinaud, and L. Wegner. What makes decentralization in developing countries pro-poor? *The European Journal of Development Research*, 17 (4):626–48, 2005.

K. Kaiser. Decentralization reforms. In A. Coudouel and S. Paternostro, editors, *Analyzing the Distributional Impact of Reforms*, volume 2. World Bank, Washington, 2006.

I. Lonzano, J. Ramos, and H. Rincón. Implicaciones fiscales y sectoriales de la reforma a las transferencias territoriales en Colombia. *Borradores de Economía - Banco de la Republica, Colombia*, 437, Apr 2007.

J. Manor. *The Political Economy of Democratic Decentralization.* World Bank, Washington, Mar 1999.

MEN. Directiva ministerial n.04/2003. Technical report, Ministero de Educación Nacional, Bogotá, D.C., Mar 2003.

R. Musgrave. *The Theory of Public Finance.* McGraw-Hill, New York, 1959.

W. Oates. *Fiscal Federalism.* Harcourt Brace Jovanovich, New York, 1972.

Ronald Oaxaca. Male-female wage differentials in urban labor markets. *International Economic Review*, pages 693–709, 1973.

U. Panizza. What drives fiscal decentralisation? *CESifo DICE Report 1/2004*, 2004.

Per Pettersson-Lidbom. Do parties matter for economic outcomes? A regression-discontinuity approach. *Journal of the European Economic Association*, 6(5):1037–1056, Sep 2008.

R. Reinikka and J. Svensson. Local capture: Evidence from a central government transfer program in Uganda. *The Quarterly Journal of Economics*, 119(2):678–704, May 2004.

Dennis A. Rondinelli, John R. Nellis, and G. Shabbir Cheema. Decentralization in developing countries: A review of recent experience. *World Bank Staff Working Papers - Management and Development Series*, 581(8), Jul 1983.

M. Santa Maria S., N. Millan U., J. Moreno B., and C.F. Reyes. La descentralización y el financiamento de la salud y la educación en los departamentos: Â¿Cuáles son las alternativas? Informe final, Federación nacional de departamentos y Fedesarrollo, Dec 2009.

A. Sarmiento and J.E. Vargas. Descentralización de los servicios de educación y salud en colombia. *Conyuntura Social - Fedesarrollo / Instituto SER de Investigación*, 16:91–136, May 1997.

R. John Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3): 751–754, 1986.

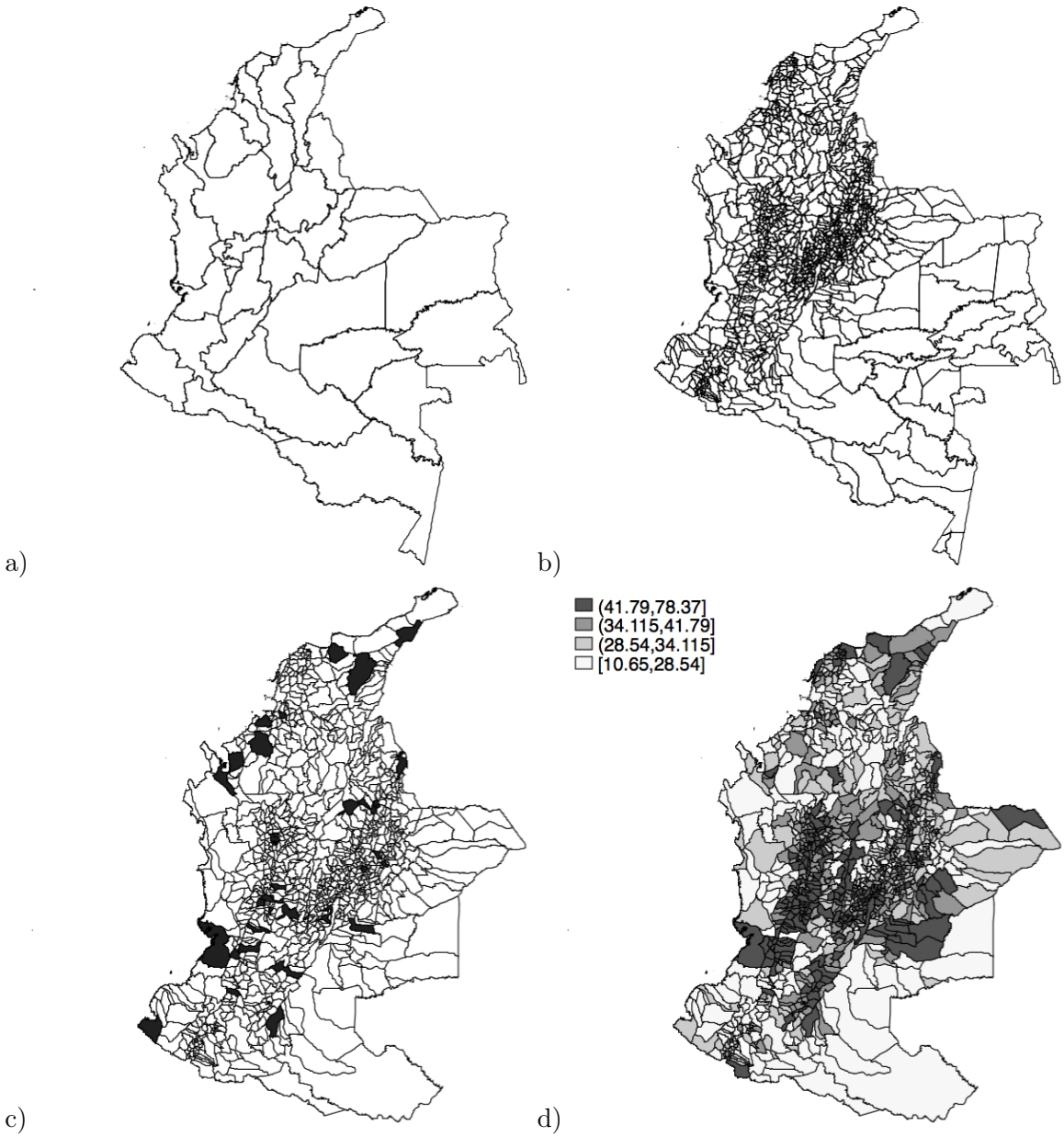
G. Toro. Trabajamos por la descentralizacion, la gobernabilidad y la autonomia local. Presentacion, Federación Colombiana de Municipios (FCM), Jun 2006.

C.H. Vergara and M. Simpson. Evaluación de la descentralización municipal en colombia. Estudio general sobre los antecedentes, diseño, avances y resultados generales del proceso de descentralización territorial en el Sector educativo. *Archivos de Economía - Departamento Nacional de Planeación*, 168, Dec 2001.

X. Zhang. Fiscal decentralization and political centralization in China: Implications for growth and inequality. *Journal of Comparative Economics*, 34(4):713–726, Dec 2006.

A Appendix

A.1 Maps



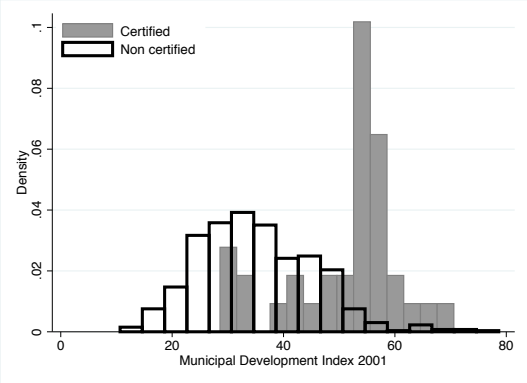
a) Colombia's departments; b) Colombia's municipalities; c) Municipalities which were certified in education in 2002 (in black); d) Distribution of Municipal Development Index in 2001. In maps c) and d) the rural south-east is omitted to allow larger zoom on the densely populated area.

A.2 Population and Municipal Development Index distributions

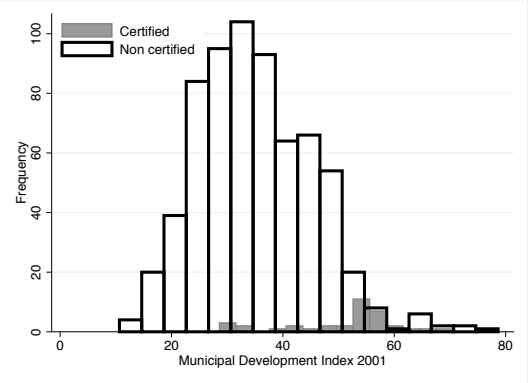
Figure 5 shows the distributions of the municipal development measures by certification status. Figure 6 extends Figure 1, illustrating population and MDI distributions for a wider range of municipalities.

Figure 5: Distribution of MDI by certification

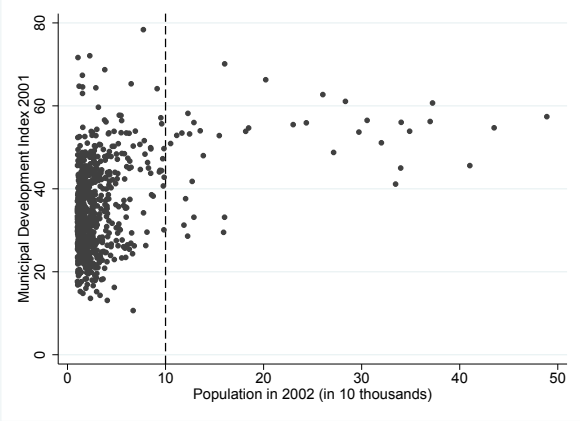
(a) Densities



(b) Frequencies

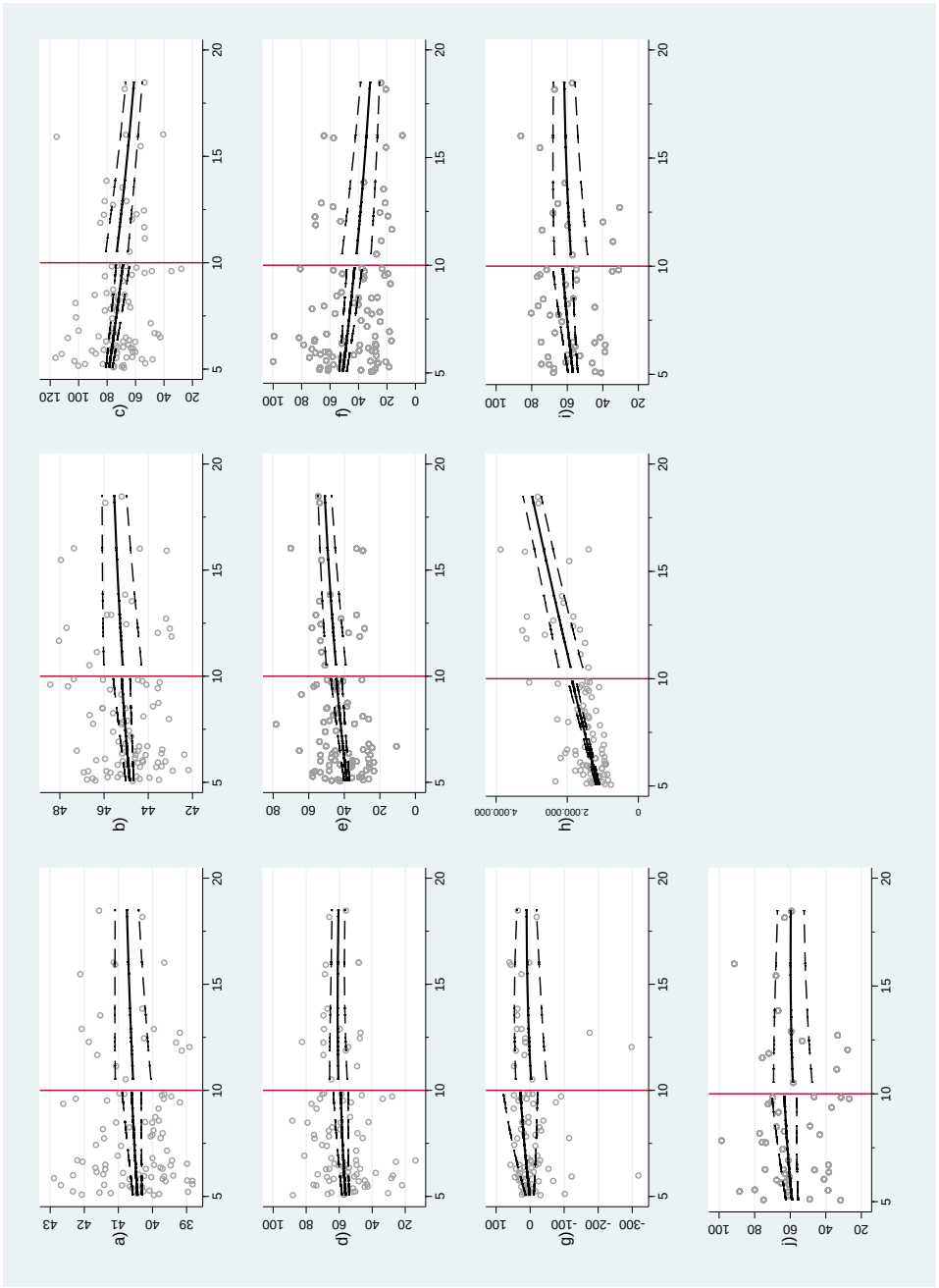


(c) Scatterplot



A.3 Smoothness checks

Figure 7: Smoothness of municipal characteristics across the discontinuity



a) Municipal Maths score average 2001 (value); b) Municipal Spanish Language score average 2001 (value); c) Gross primary school coverage 2001 (percent); d) Gross secondary school coverage 2001 (percent); e) Municipal Development Index 2001 (value); f) Unsatisfied Basic Needs indicator 1993 (value); g) Share of saved municipal current revenues (percent); h) Central Govt. transfers for education 2001 (1,000s of Pesos); i) Transparency index 2005 (value); j) Visibility and accountability index 2005 (value).

A.4 Common pre-reform trend and falsification test

In order to dissolve residual doubts about whether municipalities on the two sides of the certification cutoff may have been evolving in different ways over time also in absence of the reform, I perform a test on the pre-reform trend. Lamentably the available pre-reform years of test score data are only two (2000 and 2001), thus the test will look at the changes between those two years only: the variable $\Delta Y = Y_{2001} - Y_{2000}$ is the outcome variable in panel (a) of Table 12. No discontinuities nor patterns are discovered in the results, neither across the four development subsamples nor through the interaction term specification, confirming our belief that cities above and below the treatment cutoff are not intrinsically different from each other. Panel (b) of the same table shows the results of the RD estimation on pre-reform data, in order to verify that the test score patterns and discontinuities identified in the main results were not already existing before the treatment. Again the subsample analysis does not reveal any particular relationship between scores and development before autonomy, while the interaction term specification does show a pattern qualitatively similar to the post-reform scenario but significantly weaker in magnitude. The overall conclusion I draw from the two panels of Table 12 is that before the 2001 decentralization reform, among municipalities sized around 100thousand inhabitants, there was no evident relationship between development measures and student test scores levels or growth rates. That relationship emerged only once cities were endowed with decisional and financial autonomy over the public education service.

A.5 Progress over time using time bins

Expanding on Section 6.1.2, here I provide a different approach to the analysis of over-time behavior of the certification effect. Instead of looking at progressively later time periods as done in Table 4, Table 13 and Figure 8 show results of RD estimations performed on successive and mutually exclusive 3-year bins. Standard errors are larger with respect to the previous approach, as fewer data points enter each bin; the result patterns remain the same and average effects can be observed growing over time.

Table 12: Common-trend test and falsification test - RD estimation

(a) Pre-reform trend in scores (outcome is $\Delta Y = Y_{2001} - Y_{2000}$)

	Mathematics					Spanish Language				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	B 25%	B 50%	T 50%	T 25%	Int.	B 25%	B 50%	T 50%	T 25%	Int.
Certified	-0.461	-0.272	-0.436	0.468	-0.075	0.178	-0.589	0.393	-0.309	0.364
	(0.42)	(0.27)	(0.56)	(0.59)	(0.52)	(0.59)	(0.40)	(0.57)	(0.60)	(0.79)
Certif.*MDI '01					-0.002					-0.011
					(0.01)					(0.01)
MDI '01					0.011**					-0.010*
					(0.00)					(0.01)
F(Population)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	581	632	43	25	674	581	632	43	25	674
R-sq.	0.00	0.00	0.06	0.10	0.01	0.00	0.00	0.05	0.04	0.01

(b) Pre-reform scores (2000 and 2001)

	Mathematics					Spanish Language				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	B 25%	B 50%	T 50%	T 25%	Int.	B 25%	B 50%	T 50%	T 25%	Int.
Certified	-0.534	-0.166	0.731	0.362	-1.813***	-1.294*	0.544	-0.159	1.760	-1.531*
	(0.53)	(0.40)	(0.43)	(0.50)	(0.62)	(0.76)	(0.82)	(1.21)	(1.25)	(0.86)
Certif.*MDI'01					0.040***					0.032*
					(0.01)					(0.02)
MDI '01					0.010**					0.075***
					(0.00)					(0.01)
F(Population)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,177	1,280	86	50	1,364	1,177	1,280	86	50	1,364
R-sq.	0.00	0.00	0.08	0.06	0.01	0.01	0.01	0.03	0.08	0.22

Standard errors clustered by municipality in parentheses. * p<.10 ** p<.05 *** p<.01

Table 13: Certification on Saber 11 test scores - progress over time

[Regression Discontinuity Estimation]

(a) Top 25% MDI '01

	Mathematics				Spanish Language			
	(1) 2002-04	(2) 2005-07	(3) 2008-10	(4) 2011-12	(5) 2002-04	(6) 2005-07	(7) 2008-10	(8) 2011-12
Certified	1.591**	1.456*	2.209**	4.222**	2.642	1.331	1.134	2.294
	(0.77)	(0.83)	(0.86)	(1.69)	(1.57)	(1.15)	(0.85)	(1.41)
.								
F(Population)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	75	75	75	50	75	75	75	50
R-sq.	0.072	0.119	0.134	0.260	0.054	0.075	0.143	0.164

(b) Bottom 25% MDI '01

	Mathematics				Spanish Language			
	(1) 2002-04	(2) 2005-07	(3) 2008-2010	(4) 2011-12	(5) 2002-04	(6) 2005-07	(7) 2008-10	(8) 2011-12
Certified	-0.714	-0.820	-1.940	-3.350**	-1.523	-1.404	-1.412	-2.195*
	(0.52)	(0.85)	(1.31)	(1.66)	(1.04)	(1.13)	(0.89)	(1.33)
.								
F(Population)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,731	1,796	1,803	1,206	1,731	1,796	1,803	1,206
R-sq.	0.004	0.003	0.005	0.007	0.003	0.004	0.002	0.008

Standard errors clustered by municipality in parentheses
* p<.10 ** p<.05 *** p<.01

A.6 Testing the difference between pre- and post-reform coefficients

In this section I formally test the difference between the RD coefficient estimates obtained using post-reform data (Section 6.1) and the 'falsification' coefficients obtained using pre-reform data (Section A.4), for both the high and the low development groups. Table 14 shows difference estimates and the associated standard errors. Focusing on mathematics test scores, for high-developed cities the differences between pre-and post-reform coefficients are statistically significant on every time span considered. For the low-developed group instead, statistical difference is reached only once we look further away from the reform date, at the years 2010-2012, suggesting a slower emergence of the reform effects on this group. Coefficients on language scores do not reach statistically significant differences between pre- and post-reform years.

Table 14: Differences between pre-and post-reform coefficients

	Mathematics				Spanish Language			
	2002>	2004>	2007>	2010>	2002>	2004>	2007>	2010>
a) Top 25%	1.72*	1.89*	2.52**	3.32**	1.53	1.36	1.08	1.64
	(0.87)	(0.93)	(1.08)	(1.51)	(0.94)	(0.88)	(0.84)	(1.12)
b) Bot. 25%	–1.05	–1.27	–1.70	–2.64*	–0.95	–0.86	–1.00	–1.42
	(0.98)	(1.11)	(1.29)	(1.59)	(0.99)	(1.00)	(1.05)	(1.11)
N a)	325	275	200	125	325	275	200	125
N.mun. a)	25	25	25	25	25	25	25	25
N b)	7,713	6,521	4,786	2,986	7,713	6,521	4,786	2,986
N.mun. b)	603	602	603	603	603	602	603	603

Table of differences between pre-reform and post-reform coefficients of certification on student test scores. Standard errors on differences in parentheses. Pre-reform coefficients are obtained estimating RD model (1) on 2000-2001 student test scores. Post reform coefficients are obtained estimating RD model (1) respectively on 2000-2012, 2004-2012, 2007-2012 and 2010-2012 student test scores. * p<0.10, ** p<0.05, *** p<0.01.

A.7 Financial resources of municipalities

In Table 15 I show how the level of local development embodies significant differences in the amount of financial resources available to the municipal administration of autonomous cities. Central government transfers do implement some redistribution, but differences in local tax collection and capital gains sustain the advantage of high-developed cities with respect to low-developed ones.

Table 15: Percapita resources of certified municipalities in bottom and top development quartile

	Pre-reform (1998-2001)			Post-reform (2002-2012)		
	Low 25%	High 25%	Δ%	Low 25%	High 25%	Δ%
Total	154.63	343.71	75.88%	638.61	716.24	11.46%
	(41.68)	(122.75)		(280.38)	(250.99)	
Transfers	113.72	82.08	–32.32%	480.20	260.36	–59.37%
	(40.33)	(30.98)		(208.34)	(88.40)	
Tax collection	24.38	158.77	146.77%	54.33	254.57	129.65%
	(13.72)	(73.18)		(28.34)	(117.02)	
Capital gains	10.23	29.76	97.67%	83.11	138.39	49.91%
	(12.48)	(19.35)		(91.60)	(87.31)	
N.obs	23	14	37	90	64	154
N. municipalities	8	7	15	9	7	16

Table of mean annual percapita resources reported by municipalities in pre- and post-reform years (in thousands of Colombian pesos), and percentage differences between the two groups. Standard deviations in parentheses. Significance stars refer to t-tests on the mean differences.

A.8 Descriptive evolution of test scores

Table 16 illustrates test score differences between highly developed (first quartile) and low developed (last quartile) municipalities, at several points in time, for both Mathematics and Spanish language. Differences are indicated separately for the group of cities that obtained certification in 2002 and for the group that did not. In both groups, highly developed cities always see higher test scores than low developed ones, even before the decentralization reform. Nevertheless we can observe that for cities that obtained autonomy in 2002, the test score gap between high and low developed members increases in the post-reform period significantly more than what it does in the never-autonomous group of cities. These are descriptive patterns that do not represent estimations of the causal effects of the 2002 reform (refer to the main results for such estimations): instead they inform us about the bare over-time evolution of student performance in one group of cities relative to the other.

Table 16: Over time evolution in test score differences

	<2002	2002>	2004>	2007>	2010>
A) Mathematics					
A1. Certified in 2002	1.09***	2.83***	3.16***	3.79***	4.81***
	(0.28)	(.62)	(.68)	(.82)	(1.11)
A2. Not certified in 2002	0.01	1.40***	1.52***	2.02***	2.74***
	(.12)	(.23)	(.25)	(.29)	(.37)
A1. - A2.	1.08***	1.43**	1.63**	1.77**	2.07*
	(.31)	(.65)	(.71)	(.87)	(1.16)
B) Language					
B1. Certified in 2002	2.50***	3.32***	3.33***	3.45***	4.29***
	(.62)	(3.32)	(.70)	(.68)	(.84)
B2. Not certified in 2002	1.61***	2.05***	2.00***	1.96***	2.64***
	(.30)	(.25)	(.24)	(.24)	(.30)
B1. - B2.	0.89	1.28*	1.32*	1.49**	1.65*
	(.69)	(.75)	(.74)	(.72)	(.88)
N.obs	1,227	6,811	5,569	3,759	1,884
N. municipalities	621	628	628	628	628

Table of differences in test scores between cities in the highest and in the lowest development quartile, all conditional on population (third-degree polynomial). Differences are indicated at progressive points in time, and separately for cities certified in 2002 (rows A1. and B1.) and for cities not certified in 2002 (rows A2. and B2.). Standard errors clustered by municipality in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

A.9 Robustness checks

A.9.1 Regression Discontinuity Estimation

Different polynomials at each side of the cutoff

Table 17 replicates the results shown in Table 3, now allowing for a different polynomial on each side of the certification cutoff. In econometric terms, this table shows the results of fitting the models $Y_i = \alpha + \tau^{RD} C_i + \beta D_i + f(P_i) + f(P_i) \times C_i + \epsilon_i$ and $Y_i = \alpha + \tau_0^{RD} C_i + \tau_1^{RD} C_i * D_i + \beta D_i + f(P_i) + f(P_i) \times C_i + \epsilon_i$, where $f(P_i)$ is a third-order polynomial of population P_i. The results from the main section are robust to these alternative model specifications.

Table 17: Certification on Saber 11 test scores - by MDI '01 (2 polynomials)

(a) Mathematics

	(1) All	(2) Bottom 25%	(3) Bottom 50%	(4) Top 50%	(5) Top 25%	(6) Interaction
Certified	0.176	-2.259*	1.211	-0.307	3.065***	-2.267**
	(1.12)	(1.37)	(1.60)	(1.27)	(0.81)	(0.88)
Certif.*MDI'01 perc.						0.038***
						(0.01)
MDI'01 percentile						0.023***
						(0.00)
.						
F(Population)	Yes	Yes	Yes	Yes	Yes	Yes
N	7,572	6,536	7,100	472	275	7,561
R-sq.	0.013	0.003	0.004	0.028	0.073	0.084

(b) Spanish Language

	(1) All	(2) Bottom 25%	(3) Bottom 50%	(4) Top 50%	(5) Top 25%	(6) Interaction
Certified	0.072	-3.602***	1.274	-0.432	2.379**	-2.043**
	(1.14)	(1.31)	(1.60)	(1.18)	(1.10)	(0.96)
Certif.*MDI'01 perc.						0.035***
						(0.01)
MDI'01 percentile						0.031***
						(0.00)
.						
F(Population)	Yes	Yes	Yes	Yes	Yes	Yes
N	7,572	6,536	7,100	472	275	7,561
R-sq.	0.018	0.003	0.004	0.029	0.073	0.124

Standard errors clustered by municipality in parentheses. * p<.10 ** p<.05 *** p<.01

Correction of statistical significance for multiple testing

If one considers columns (2) to (5) in Table 3 to be testing four separate hypotheses that are related to each other in unknown ways, the α levels used as benchmarks should be appropriately adjusted. Table 18 reports the results for municipalities in the highest and lowest development quartile, with significance stars corrected for multiple testing according to the method illustrated by Simes [1986]. Let $p_{(1)}, p_{(2)}, p_{(3)}, p_{(4)}$ be the p-values, ordered from smallest to largest, for testing hypotheses $H_0 = \{H_1, H_2, H_3, H_4\}$, each corresponding to one of the four development categories used. Then each $H_{(j)}$ is rejected if $p_{(j)} \leq j\alpha/n$ for any $j = 1..4$, and H_0 is rejected if all $H_{(j)}$ are rejected.

Table 18: Certification on Saber 11 test scores - Significance corrected for multiple testing

(a) Top 25% MDI '01

	Mathematics				Spanish Language			
	(1) Post '02	(2) Post '04	(3) Post '07	(4) Post '10	(5) Post '02	(6) Post '04	(7) Post '07	(8) Post '10
Certified	2.20*	2.37*	3.00**	3.80*	1.81	1.64	1.36	1.92
	(0.86)	(0.92)	(1.06)	(1.52)	(1.14)	(1.09)	(1.04)	(1.29)
.								
F(Population)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	275	225	150	75	275	225	150	75
R-sq.	0.050	0.069	0.137	0.199	0.035	0.038	0.126	0.137

(b) Bottom 25% MDI '01

	Mathematics				Spanish Language			
	(1) Post '02	(2) Post '04	(3) Post '07	(4) Post '10	(5) Post '02	(6) Post '04	(7) Post '07	(8) Post '10
Certified	–1.58	–1.80	–2.23	–3.17	–1.55	–1.47	–1.61	–2.03
	(0.99)	(1.12)	(1.29)	(1.60)	(1.00)	(1.02)	(1.06)	(1.12)
.								
F(Population)	Yes	Yes	Yes	Yes	Yes	Yes		
N	6,536	5,344	3,609	1,809	6,536	5,344	3,609	1,809
R-sq.	0.003	0.003	0.005	0.007	0.003	0.003	0.004	0.007

Effect of certification on student test scores. with significance stars corrected for multiple testing according to the Simes (1986) method. * p(j)<j0.10/4, ** p(j)<j0.05/4, *** p(j)<j0.01/4, for any j=1,2,3,4 of the ordered p-values.

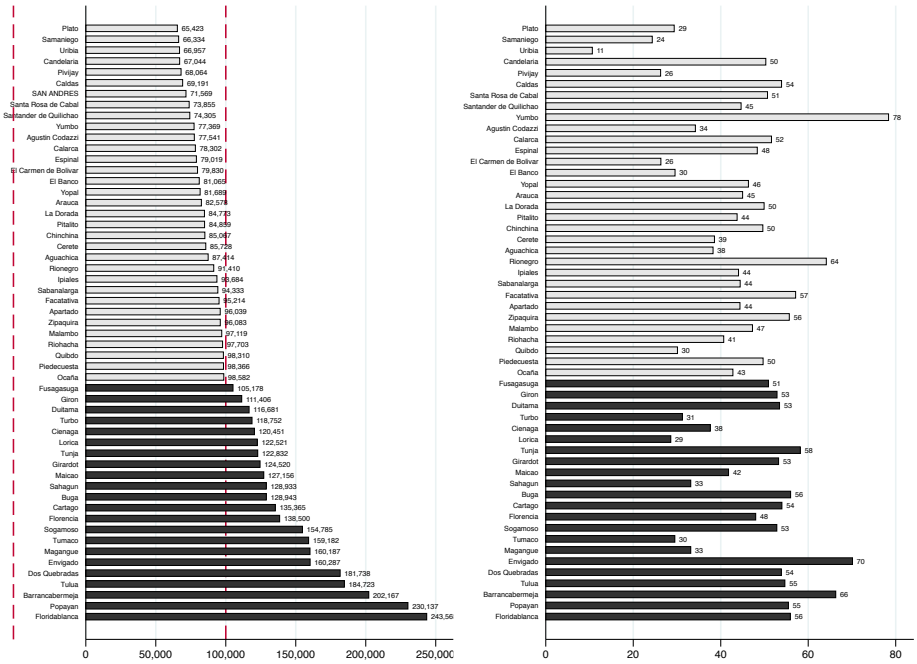
A.9.2 Difference-in-Differences Estimation

Different cutoffs for the discontinuity sample

Table 19 shows the results of Table 5 employing different choices of the discontinuity-sample. Columns (4) to (8) correspond to the same sample as used for the RD estimation.

Figure 6: Population and MDI distributions

(a) 65 to 250 thousand inhabitants



(b) 40 to 500 thousand inhabitants

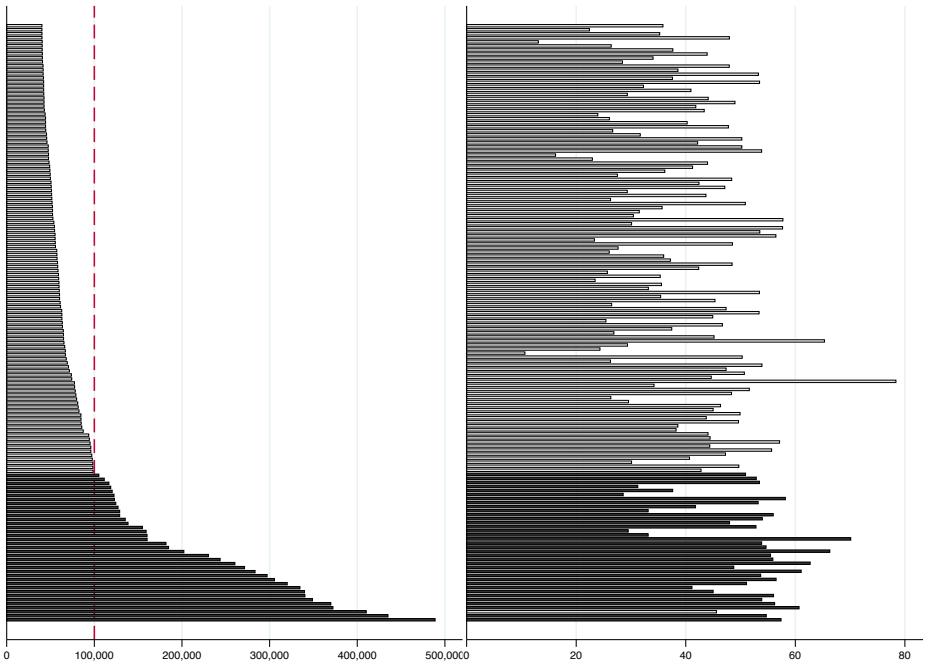
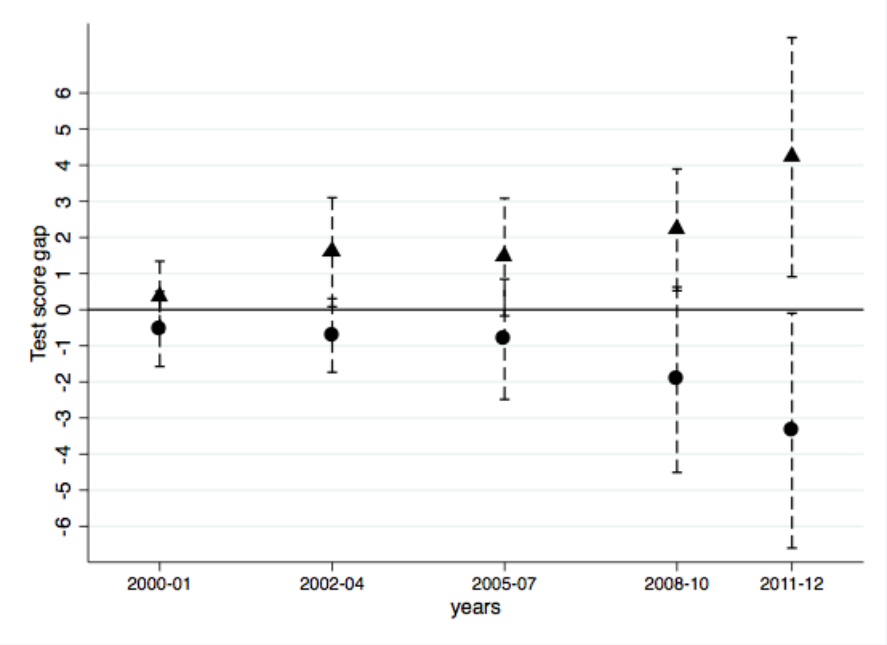


Figure 8: Effect of certification over time

[Regression Discontinuity Estimation]



RD estimations of the effect of certification on average Math test scores, for certified municipalities in the top 25% and the bottom 25% development range (triangles and circles series respectively). Capped spikes indicate 95% confidence intervals on point estimates.

Table 19: Different cutoffs for the discontinuity sample (FE)

	50,000 - 250,000				10,000 - 500,000			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Mate	Mate	Lang	Lang	Mate	Mate	Lang	Lang
Certified	0.526	–3.210***	0.240	–1.346***	0.775***	–2.778***	0.264**	–1.232**
	(0.35)	(0.91)	(0.18)	(0.50)	(0.22)	(0.80)	(0.13)	(0.48)
Certified*MDI'01		0.077***		0.032***		0.070***		0.029***
		(0.02)		(0.01)		(0.02)		(0.01)
.								
Time dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,227	1,214	1,227	1,214	8,938	8,925	8,938	8,925
N groups	95	94	95	94	698	697	698	697
R-sq.	0.60	0.61	0.71	0.71	0.42	0.42	0.62	0.62

Standard errors clustered by municipality in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Chapter 3

New Teachers for Colombia: Is Quality Control Working?

Abstract

In 2002 the career of Colombian public school teachers was significantly reformed through the introduction of a selective entry competition and of further quality incentives. This paper estimates how the new quality-screened teachers impact students' high school performance. We exploit the fact that the novel regulation applied only to newly hired teachers, whereas those already in office in 2002 remained exempt, creating a mix of New-Regulation and Old-Regulation teachers in Colombian schools. Using data at the school-year-subject level, we eliminate any school-level confounders and associate the proportion of New-Regulation teachers to the variation in student test scores. We pin down a positive and significant, although not very large, effect of New Regulation teachers on student performance. New Regulation teachers have decreasing marginal returns, are more effective in larger schools and when surrounded by colleagues holding postgraduate degrees. We also document that the enforcement of the New Regulation has been somewhat unsatisfactory, since in the period 2008-2013 around 30% of all New Regulation teachers are employed in temporary positions without having passed the compulsory entry exam. These teachers have lower and less robust impacts on student performance.

1 Introduction

The opening words of the 2016 OECD report on Colombian education read '*Colombia has made education a main priority to improve the economic and social prosperity of the country and pledged more resources to this sector than any other policy area*' [OECD, 2016]. Indeed, over the past two decades the country has been transforming its education system on several aspects, and in this paper we are focusing on the quality assurance mechanisms that have been introduced to the teaching profession in 2002 in the effort of improving its standards.

Teacher hiring systems that feature competition on skills, such as national entry exams and performance rankings, may by now be common in the Western world but much less so in Latin America, Asia and Africa. Before the reform, in Colombia the career of public teacher was open to any candidate possessing the citizenship and given levels of education. The appointment of candidates into the open vacancies followed opaque evaluation processes, and once entered the career, salary and rank upgrades were based on seniority. This system lent itself to dynamics of corruption and clientelism [Duarte, 2001], beyond not creating any incentives for individuals to invest in their teaching skills and subject knowledge. Sadly, this type of structure is still common in many developing countries around the world[1], making policy advice and careful reform evaluation in this context particularly urgent.

This paper analyzes the consequences of the introduction of a broad set of quality controls to the public teacher profession in Colombia. The quality controls can be summarized by (1) a public entry contest, that establishes the order in which candidates will choose vacancies 2) a probation period of 4 to 12 months, which confirms the appointment to the vacancies upon successful completion 3) continuous evaluation along the career, with yearly performance reports by the school headmaster, that can result in termination of the employment if unsatisfactory 4) salary upgrades being made conditional on skills evaluations, which are carried out through a written examination. We measure the impact of the reform on student performance at high school level, and find that bringing the share of New Regulation teachers from 0 to 1 in a given school year and subject increases the average performance of students by about 6% of a standard deviation.

We further explore heterogeneities in the effect across teacher and school characteristics. We

[1] We found many similarities between the pre-reform Colombian context and the current ones in many Latin American countries, such as Chile, Argentina, Brazil, Venezuela, Peru, Nicaragua, El Salvador and Panama'. For similar contexts in South Asia and the rest of the world, find discussions and details in Murillo et al. [2007], Hanushek [2009] and Hanushek and Woessmann [2012].

find that the marginal effect of New Regulation teachers reaches more than double the average magnitude when the share of these teachers is still low ($< 20\%$) in a given school year and subject: in other words, marginal returns are initially high and then decreasing. The effect of new teachers also depends positively on the number of colleagues in the year and subject, and on the share of postgraduate-holding colleagues. We are not able to pin down any significant relationship between the effect of New Regulation teachers on student performance and the average entry contest score of these teachers.

We also examine teacher retention patterns in the education system, and conclude that the professional incentive framework seems to be working partially, as having a postgraduate degree and scoring higher in the entry test is associated with longer permanence in the system, but the very top-scorers are observed dropping out faster than the median ones, suggesting a lower attractiveness of the profession for such teachers.

Finally, our empirical analysis suggests that loopholes in the new legislation may have allowed a conspicuous fraction of teachers to evade some of the quality assurance provisions, by teaching under the status of "temporaries" without having entered the career through the regular path. These teachers have a lower and less robust effect on student performance.

The next section briefly reviews the closest related works; Section 3 gives a thorough description of the most important aspects of the 2002 reform; Section 4 describes the data we use; Section 5 details the empirical strategy employed; Section 6 illustrates and discusses results, followed by concluding remarks in Section 10.

2 Closely related literature

This paper contributes to the flourishing literature on teacher quality, which divides into several themes: assessing how much of the variation in student outcomes is explained by the teacher effect [2]; understanding which teacher characteristics make teachers better [3]; understanding which type of policies and interventions are effective in making teachers better or attracting and retaining better teachers[4]. The answer to our main exercise, quantifying the impact of New Regulation teachers

[2] See for example Chetty et al. [2014], Rivkin et al. [2005], Rockoff [2005] and, for an excellent review, Hanushek and Rivkin [2012].

[3] For example Rockoff et al. [2011], Kane et al. [2008], Gordon et al. [2006]

[4] For example Hanushek et al. [2004], Figlio [2001]

on student performance, shall provide a reference point about the effectiveness of reforms such as the Colombian one[5] in providing better teachers. The results of our secondary analyses on heterogeneity and retention patterns help towards understanding the channels of the reform effect, and suggests which types of teachers may find the profession appealing under a framework such as the Colombian one.

Closest to this paper both in scope and methodology we found work by Ome(2013, 2012), who also uses a fixed effects methodology to estimate the impact of the Colombian teacher career reform on primary and high school test scores. He uses three years of data on teacher records and student scores (2002, 2008 and 2009); he exploits within-school, across-year variation in the share of new-regulation teachers to identify the effect of the new teachers. While pinning down positive effects at primary school level and on student dropout rates, he finds no effect of the New Regulation teachers on test scores at high school level. We think that the little amount of within-school variation in teacher composition across the three years might be the reason for the negative finding (low testing power).

3 The 2002 reform of the teacher career

3.1 Pre-reform situation and reasons for the reform

Before 2002, the teaching profession in Colombia was regulated by Decree 2277 of year 1979, and successive modifications. The appointment and transfers of teachers were considered administrative acts, and thus the responsibility of department governors and/or mayors[6]. There were requirements on the education levels for teachers at the different school grades. The scheme for career upgrades was based on combinations of years of service, education level and attendance of training workshops. Finally, teachers were guaranteed to remain in service until retirement age except in cases of ascertained severe misconduct[7]. Overall, the legal framework was characterized by very little transparency in procedures, excessive protection of employed teachers, and lack of incentives towards the improvement of skills and teaching performance. Clientelism and politiciza-

[5] in a context similar to pre-reform Colombia, which is characteristic still today of many countries in Latin America, see footnote 1.

[6] See Art. 106 Ley 115 /1994

[7] Only in 1994 the law started to mention public contests as a desirable method of appointment of teachers, but these never took place until the 2002 reform.

tion of teacher appointments were substantial and well-known issues; far too often public schools were used as 'placement pools' for relatives and connections of influent personalities (see excellent descriptions and discussions in Duarte [2001] and Duarte [2003]).

3.2 The new public contest procedure

Teachers that have been entering the profession from January 1st 2002 onwards have to go through a public contest procedure in order to be assigned a vacancy. Contests are called separately for each education authority (department or certified municipality) and specify the vacancies available on their territories, and candidates must choose the one education authority they wish to apply to in that year. The contest is based on a score system and serves the purpose of establishing a rank among applicants, which determines the order in which successful candidates will be allowed to choose their preferred vacancies. The stages of the contest and their corresponding weights are summarized in Table 1; the subsequent subsections detail each stage further.

Table 1: Stages of the entry contest

	Purpose	Use of score	Minimum threshold	Weight in contest		Responsibility
				T*	H*	
Exam	Teaching aptitude, subject knowledge	Eliminatory and ranking	60% (T) 70% (H)	55%	45%	ICFES[21]
Exam	Psychometric test	Ranking	-	10%	10%	ICFES
CV	Credentials evaluation	Ranking	-	20%	30%	CNSC[10] or delegate
Interview	Behavioral evaluation	Ranking	-	15%	15%	CNSC or delegate

Note: * T = teachers; H = headmasters
Source: MEN [2006]; MEN [2012]; GEARD [2013]

The exam

At the first stage of the contest, candidates sit an exam which is identical across the country and is administered and evaluated centrally by governmental agencies[8]; it is structured into three modules

[8]ICFES (Instituto Colombiano para la Evaluación de la Educación) and CNSC (Comisión Nacional del Servicio Civil). The exam registration fee is below 9 USD (2012-2013 contest). The exam questions are elaborated by the National University, the largest public university in Colombia.

testing teaching aptitude, subject knowledge and psychometric values[9]. Candidates who do not score a minimum of 60 /100 points on each of the three modules must exit the contest. For surviving candidates, the exam stage will represent 65% of their global score (55% for school directors).

The evaluation of credentials

Scores are assigned to academic degrees, additional training courses, academic productions and publications, past experience, past teaching evaluations (where present) and career awards, according to official tables set by CNSC[10], the body in charge of this stage of the context. The credentials score represents 20% of the total (30% for aspiring school directors).

The interview

The interview of candidates is also responsibility of CNSC, who may nominate local delegate bodies to decentralize the process. Typically universities and other certified higher education institutions are delegated and form ad hoc interview committees under the supervision of CNSC. The committee questions each candidate in person and the evaluation accounts for the remaining 15% of the global contest score.

3.3 The probation period

Once having completed the public contest and chosen one of the available vacancies according to his or her priority rank, the aspiring teacher starts a probation period that lasts up to the end of the ongoing academic year (minimum four months). At the end of probation, the candidate's performance is evaluated by the school headmaster, and conditional on a positive evaluation the new teacher takes permanent possession of the chair. Anecdotal evidence suggests that, in practice, very little further selection is happening at this stage, i.e. that virtually all teachers reaching the probation period are then appointed to the chair; in Table 5 we show statistical evidence supporting

[9]The 'aptitudes' module aims at assessing the candidate's ability to appropriately deal with language and numbers, and his knowledge of basic pedagogical concepts. The second module evaluates proficiency and skills of the candidate in his subject speciality. The psychometric test evaluates the candidate's hypothetical responses when facing pedagogical or institutional issues arising in the everyday teaching life.

[10]Comisión Nacional del Servicio Civil. It is an autonomous and independent body, located at the highest level in the structure of the Colombian State. It has legal identity and administrative, financial and managerial independence, and it is not part of any sector of the government authority. *(Description translated from the institutional webpage of CNSC, http://www.cnsc.gov.co/index.php/institucional/direccionamiento-estrategico/quienes-somos-cnsc; fetched on 19 Jan 2015)*

this claim.

3.4 Permanent evaluation and incentives

The New Regulation introduced permanent evaluation practices, aiming at ensuring a continued satisfactory performance by teachers, as well as providing them with incentives to improve over time.

The first form of permanent evaluation consists of yearly assessments compiled by school headmasters and reported to the local education authority, in which the headmaster comments on the teacher's performance following standardized criteria[11]. Two consecutive years of negative evaluations lead to discontinuation of the employment as a teacher. This change is important especially if compared to the over-protected status that teachers were enjoying under the old career regulation.

The second form of permanent evaluation and incentives brought in by the 2002 reform was making career upgrades conditional on passing public examinations that evaluate teachers' subject knowledge and teaching skills (*Evaluación de Competencias*). The new examination was added to the existing requirements of possessing the level of education required for the upper level, and of having spent 3 years at the current one (MEN [2013b]). The career structure and the corresponding pay scale of New Regulation teachers is different with respect to Old Regulation ones, and the two are illustrated in Table A.2, along with 2008 salary levels. Noticeable changes between the Old and the New Regulation are the introduction of a postsecondary specialization as the minimum education level for teachers[12], and the reward of higher education degrees through higher salaries for teachers holding them.

4 Data

Data on teachers

Data on teachers is available thanks to a the yearly information reporting procedure that is being enforced by the Ministry of Education across all public schools of the country[13]. Schools are required

[11] *Evaluación anual de desempeño laboral.* The current evaluation procedures are regulated in detail through Decree 3782 /2007 by the Ministry of Education.

[12] Bringing the minimum years of education from 11 (secondary education) to 13 or 14 (postsecondary specialization).

[13] Resolución 166 /2003 and following versions.

to give details on their pupils, staff and infrastructure, through the standardized formats set by the Ministry every year[14]. We have individual data on teachers at all public schools of the country for the years 2008 to 2013; each record includes the teacher's demographics, education level, school of placement, teaching subject, teaching level, type of contract, date first hired in the public education system, and salary level.

Teacher test scores, and two types of New Regulation teachers

Using their unique national ID document, we managed to match 81.15% of all New Regulation high school teachers[15] to their entry exam scores. In the analysis we will use the most recent test score for each teacher, as the most up-to-date measure of his or her skills; Figure 1 graphs the density of these scores. Surprisingly, the mass of scores is below the minimum requirement of 60 points is very large. A worrying 34% of all new-regulation teachers who were holding a teaching position over the period 2008-2013[16] did not meet the minimum test score requirement on their most recent attempt[17]. Table 2 shows that a mere minority of these teachers have obtained a sufficient score at *some* point in the past[18], leaving almost 13 thousand subjects in teaching positions without having ever passed the compulsory entry exam[19]. The law allows such individuals to fill teaching vacancies temporarily, and only in the absence of legitimate candidates willing to fill the positions. In fact, Table 2 also reveals that 89% of teachers that have never passed the exam are holding temporary positions and 6% are registered as being on probation, while the remaining 5% has remarkably managed to land a permanent position.

These statistics split the group of New Regulation teachers into two potentially very different subgroups: New Regulation teachers who have at some point passed the entry exam and New

[14]The reasons behind the collection of this administrative data is for the Ministry to be able to monitor the status of the public education sector and to identify needs and priorities. School headmasters are in charge of ensuring the correct collection and reporting of the information, and of passing the data onto the local education authority ("education secretariat" of the department or certified municipality), which in turn reports to the Ministry of Education.

[15]That is 43,197 out of 53,234 New Regulation high school teachers. Overall we were able to match to their test scores 115,462 out of 145,724 New Regulation teachers (79,23%). The unmatched part is largely due to a large number of missing IDs in the 2006 edition of the entry examination.

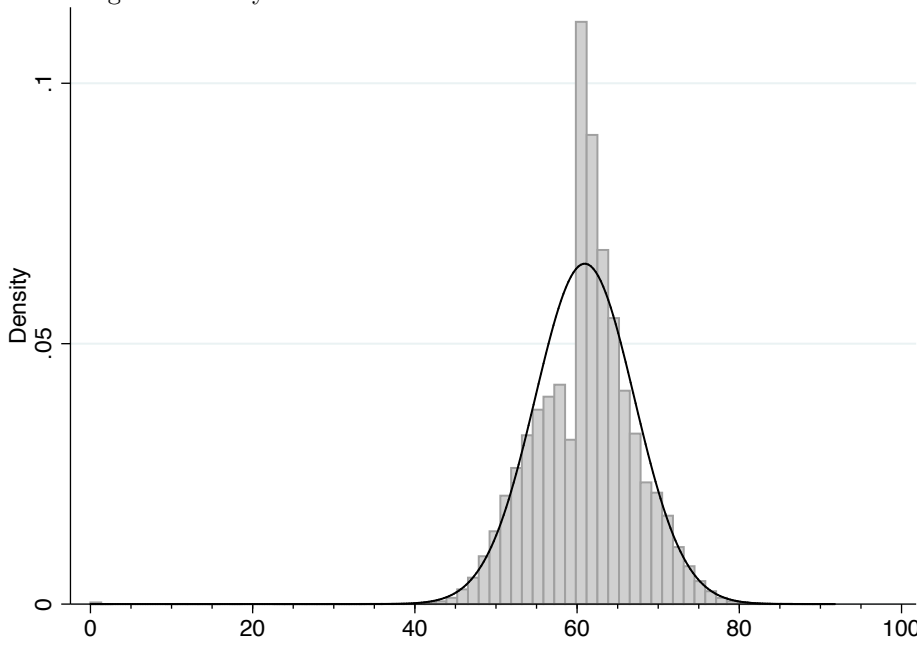
[16](and whom we were able to match with their scores)

[17]The figure could be reduced by matching the 2012 entry contest data, that was not available to us at the time of this project. If all teachers hired in 2012 and 2013 had regularly passed the exam, the percentage of those not having passed it would reduce to 28.36%.

[18]And have been retaking the exam, for example, in order to move to a different education authority.

[19]And almost 37 thousand teachers overall, including the school levels other than secondary. These figures are a lower bound to the true ones, as we are missing the test scores of about 20% of our sample.

Figure 1: Entry exam scores of teachers observed over 2008-2013



Density of entry test scores (most recent score of each teacher); bin width = 1.33 points; normal curve is overlaid.

Table 2: Teachers with most recent score below 60

Type of position	Never above 60	Above 60 in past	Total
Permanent	676	429	1,105
	5.24%	23.04%	7.49%
Temporary	11,442	1,261	12,703
	88.73%	67.72%	86.08%
Probation	777	172	949
	6.03%	9.24%	6.43%
Total	12,895	1,862	14,757
	100.00%	100.00%	100.00%

Regulation teachers who have not, and we will refer to these two groups as "New Regulation Passed" and "New Regulation Not Passed" in the remainder of the paper.

Data on student outcomes

We will focus on high school test outcomes. The Colombian high school test is called Saber11 and is sat by students after 11 years of schooling, at completion of secondary school and before

university[20]. Saber11 test data is collected by ICFES[21] and it is the most complete among the standardized tests being conducted at different school grades; it is widely accepted as the reference examination to evaluate the quality of Colombian secondary education. Saber11 evaluates a range of seven school subjects, which are the same in all secondary schools[22]; test scores range from 0 to 100 in each subject and are standardized by subject at the national level, so that each student's score is informative about his/her position relative to the national average in that subject. Individual-level Saber11 test scores are made available by ICFES for the years 2000 to 2012, with information about the school and municipality to which each student belongs. We will use years 2008-2013 and associate them with the available teacher data.

Matching students and teachers

Unfortunately we do not have information on class groups, and thus we are unable to achieve a perfect match between teachers and the students they actually taught. The closest match we can obtain is between teachers teaching at secondary school level (years 6 to 11) in school i, year t and subject s, and the Saber11 (year 11) student test scores in the same school i, year t and subject s. Therefore our unit of observation is a school-year-subject cell, and we collapse and average the individual data to that level. Our main outcome variable is "Average test score", our regressor of interest is "Share of New Regulation teachers", which is given by the number of teachers hired under the New Regulation over the total number of teachers teaching the given subject at secondary level, in the given school, the given year.

5 Empirical strategy

We are interested in measuring the impact that the introduction of New Regulation teachers has had on student test scores. In our years of reference, 2008-2013, both New Regulation and Old Regulation teachers were teaching in Colombian secondary schools, and we will exploit this fact in

[20]Schooling years 10 and 11 are not compulsory and are attended by around 41% of the eligible school population (2012 data: Sistema Nacional de Indicadores - Tasa de cobertura neta - Ministerio de Educación Nacional - http://menweb.mineducacion.gov.co)

[21]Instituto Colombiano para la Evaluación de la Educación. The same agency also administers the national tests taken by students at different stages of their career. It is a governmental agency with social scope within the sector of public education; a national, decentralized public entity of special nature, with own legal identity, administrative independence and own assets; it is bound to the Ministry of Education. *(Description translated from the institutional webpage of ICFES, http://www.icfes.gov.co/informacion-institucional/marco-legal; fetched on 19 Jan 2015)*

[22]Mathematics; Spanish Language and Literature; Biology; Chemistry; Physics; Philosophy and English.

order to identify the effect of the former. We also exploit the fact that within each school, students sit the exam on different subjects. We use heterogeneity in student performances and in the New Regulation teacher shares *across subject, within the same school*. Our main specification would be:

$$y_{its} = \beta_0 + \beta_1 SNP_{its} + \beta_2 SNNP_{its} + \boldsymbol{\beta_k X_{kits}} + \alpha_{it} + \alpha_s + e_{its} \quad (1)$$

where the unit of observation is at the school(i)-year(t)-subject(s) level, α_{it} is a school-year fixed effect, α_s is the subject fixed effect and e_{its} is the residual. Our coefficient of interest is mainly β_1, on the share of New Regulation teachers that have passed the entry exam (SNP), as this is the type of teachers the reform has been aiming at producing. Secondarily, we are also interested in coefficient β_2, on the share of New Regulation teachers that have not passed the exam ($SNNP$), as it can provide an interesting comparison to β_1 and help us understanding the channels of the effect. This specification enables us to rule out any school-level factor or shocks which may induce selection of teachers into schools, or cause correlation between the share of New teachers and the student scores, such as school characteristics of both time-invariant and time-varying nature.

Similar regression models which do not feature school fixed effects are exposed to the bias induced by non-random selection of new-regulation teachers into schools. Those who feature school- but not school-year fixed effects (as in Ome [2012a]) are able to account for time-invariant sources of bias but still exposed to the spurious correlation between y_{it} and SNP_{it} or $SNNP_{it}$ deriving from time-varying factors or shocks, such as changes in school management or shocks to school resources.

The next section presents our main results. Importantly, we find significant differences between the output of the more naïve school-fixed-effects model and that of the school-year-fixed-effects strategy we use.

6 Main results

Table 3 shows our main results, obtained estimating model (1) on the data previously described. We estimate a positive and significant effect of New Regulation - Passed teachers on student Saber11 scores, in a magnitude of about 0.20 points increase in the average score for that subject when the share of new-regulation teachers goes from 0 to 1 in that subject (6% of a subject standard

deviation; 2% of a student standard deviation). Our preferred school-year-fixed effects specification in column (5) estimates an effect of New Regulation - Passed teachers by two thirds higher than the one we obtain through a regular school-fixed-effects model (column (3)). Estimations obtained without any within-school variation are very far off, reflecting strong sorting of teachers into schools (columns (1 and 2)).

Table 3: The effect of New Regulation teachers on student performance

	(1)	(2)	(3)	(4)	(5)
Share New Regulation Passed	–0.35***	0.89***	0.12***	0.20***	0.20***
	(0.05)	(0.07)	(0.03)	(0.04)	(0.04)
Share New Regulation Not Passed	–1.83***	–0.36***	–0.03	0.14***	0.14***
	(0.06)	(0.08)	(0.04)	(0.04)	(0.04)
Age		0.03*	0.02**	0.03***	0.03***
		(0.01)	(0.01)	(0.01)	(0.01)
Age^2		–0.00*	–0.00***	–0.00***	–0.00***
		(0.00)	(0.00)	(0.00)	(0.00)
Experience		0.12***	0.02***	0.03***	0.03***
		(0.01)	(0.00)	(0.00)	(0.00)
Experience^2		–0.00***	–0.00***	–0.00***	–0.00***
		(0.00)	(0.00)	(0.00)	(0.00)
Share postgrad degree		0.94***	0.01	0.02	0.02
		(0.06)	(0.02)	(0.02)	(0.02)
Subject FE	✓	✓	✓	✓	✓
School FE			✓	✓	✓
Year FE	✓	✓	✓	✓	✓
School-year FE				✓	✓
Subject-specific trends					✓
Mean(y)	43.33	43.33	43.33	43.33	43.33
sd(y)	3.29	3.29	3.29	3.29	3.29
N.obs	151,178	151,178	151,178	151,178	151,178
N.groups	.	.	5,969	29,609	29,609
R-squared	0.19	0.20	0.68	0.79	0.79

Note: SE clustered by school in parentheses. Each observation is subject 's' in school 'i' in year 'y'. No fixed effects in columns (1) and (2), school fixed effects in column (3), school-year fixed effects in columns (4) and (5). * p<0.05, ** p<0.01, *** p<0.001

We estimate a positive and significant effect on student test scores also by New Regulation teachers who have not passed their entry exam, in a magnitude of about 0.14 points (4% of a subject- or 1.4% of a student standard deviation). The effect of this category of teachers will prove

less robust to the inclusion of additional time and experience controls with respect to the effect of fully accredited teachers (see Table 7), but its magnitude is nevertheless non trivial. We devote section 7.4 to the discussion of this category, as we reflect upon the possible channels through which the reform operated. In the Appendix we also show estimation results obtained considering New Regulation teachers as a single group, without distinguishing between those who have passed the entry exam and those who have not (Table 3).

6.1 Falsification test

Our main identification assumption is the absence of sorting of teachers across subjects, within each school and school year. This assumption would be threatened if we were facing dynamics such as subject-wise selective hiring, i.e. the targeted hiring of New Regulation teachers into a specific subject (within a school and school year). The legislative framework regulating Colombian public education is not suited to such dynamics, given the very limited freedom and decisional power that is left in the hands of single schools and principals - and hiring and firing of personnel is not among their rights[23]. Nevertheless, in order to lift any remaining doubts on these matters, we perform a falsification test that aims at detecting subject-level correlation between pre-reform student test scores and the post-reform shares of New Regulation teachers in that subject. In other words, we estimate our main Model 1 using 2000 and 2001 test scores on the left-hand side, instead of post-reform scores. If New Regulation teachers were selectively entering subjects that 'needed' them because of bad performance, or that 'attracted' them because of good performance, a negative or positive correlation ought to emerge. Table 4 reports the coefficients on *SNP* and *SNNP* for each combination of pre-reform scores and post-reform teacher plant. As we can see, no correlation is detected in any of these combinations, bringing further support to our claim of absence of selection of teachers at the subject level.

6.2 Nonlinear and interaction effects

In this section we explore nonlinearities in the effect of New Regulation teachers on student subject performance. Firstly, we are interested in whether the marginal return to New Regulation teachers is constant across their share, or varying with it. We augment our main model (1) with the quadratic

[23] Except the evaluation post-probation period from 2002 onwards (see Section 7.2)

Table 4: Falsification: Share of New Regulation teachers on pre-reform student test scores

Student test scores 2000	2008	2009	2010	2011	2012	2013
Share New Regulation Passed	0.06	0.01	0.01	–0.04	–0.01	–0.02
	(0.09)	(0.08)	(0.08)	(0.08)	(0.08)	(0.07)
Share New Regulation Not Passed	0.02	0.03	0.03	–0.05	0.01	0.00
	(0.11)	(0.10)	(0.10)	(0.10)	(0.09)	(0.08)
N.obs	11,715	12,395	12,395	13,366	14,188	15,204
N.groups	2,595	2,697	2,697	2,828	2,945	3,057
Student test scores 2001	2008	2009	2010	2011	2012	2013
Share New Regulation Passed	–0.05	–0.01	–0.03	–0.08	–0.09	–0.01
	(0.08)	(0.07)	(0.07)	(0.07)	(0.07)	(0.06)
Share New Regulation Not Passed	–0.01	0.03	–0.02	–0.03	–0.08	–0.02
	(0.10)	(0.09)	(0.10)	(0.09)	(0.09)	(0.08)
N.obs	12,103	13,077	12,831	13,818	14,694	15,769
N.groups	2,696	2,783	2,798	2,932	3,054	3,171

Note: Pre-reform student test scores regressed on each post-reform year's share of New Regulation teachers. SE clustered by school in parentheses. Each observation is subject 's' in school 'i' in year 2000 or 2001. School and subject fixed effects, and all controls of Table 3 - model (5) are also included. * p<0.05, ** p<0.01, *** p<0.001

and cubic terms SNP^2_{its} and SNP^3_{its}, allowing for a nonlinear response to increasing SNP, the share of New Regulation teachers who have passed their exam; Figure 2(a) shows the results graphically and Table A.8 numerically[24]. We find the marginal effect to be highest at low shares, starting as high as 0.75 points (22.8% of a school-year-subject standard deviation) and declining with the share, becoming statistically insignificant when *SNP* reaches values of around 40%.

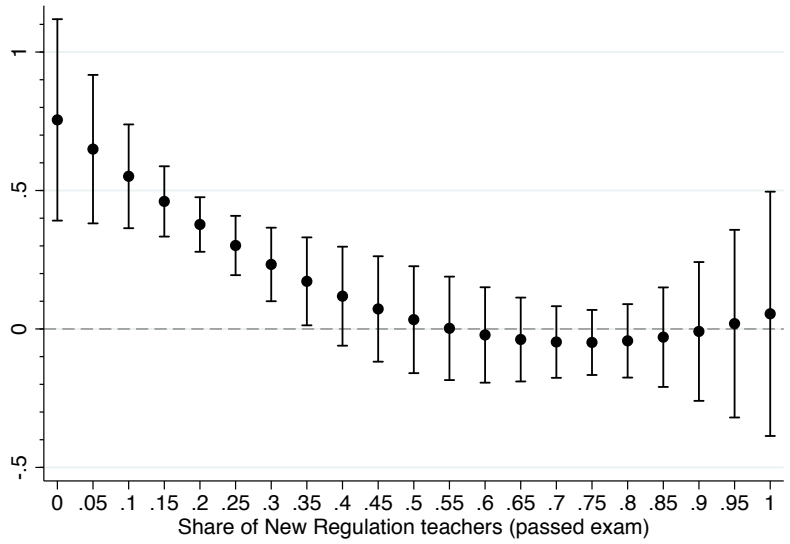
In Figures 2(b) and (d) and the corresponding Tables A.7 and A.10 we can see how the effectiveness of New Regulation teachers appears to increase with the number of colleagues in the year-subject, maybe suggesting positive network effects, and with the share of postgraduate-holding colleagues. Figure 2(c) and Table A.9 show an attempt at pinning down heterogeneous effects by quintiles of average entry tests scores of teachers: the Table reveals positive associations between student test scores and higher test performance of teachers, but the interactions with SNP do not display interesting patterns. Finally, Figure 2(e) and Table A.11 show us that there are some differences in the effect of New Regulation teachers by subject taught, with the weakest effects on

[24]In all panels of Figure 2 and the corresponding Tables, the sample has been limited to subject-years with 11 or less teachers (5 school-subject-year cells dropped).

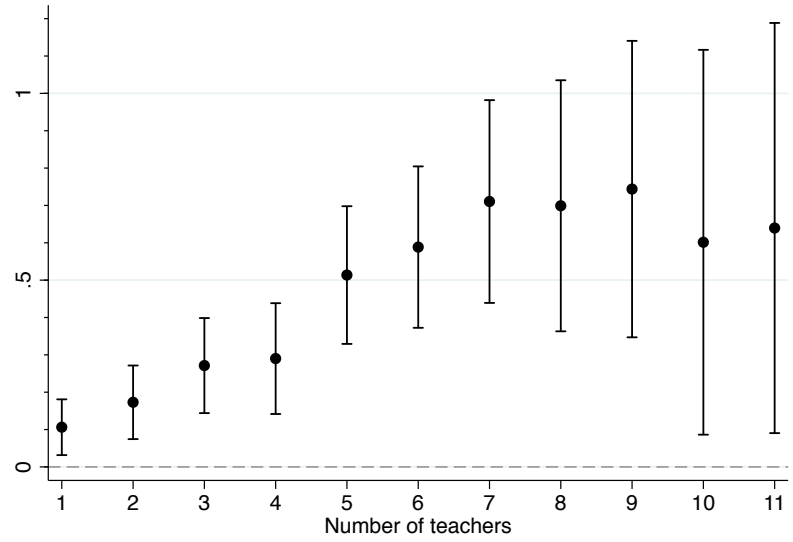
English foreign language and Philosophy, and the strongest on Mathematics.

Figure 2: Marginal effects by different subject-year characteristics

The following panels plot the marginal effect of 'Share New Passed' on student test scores when the variable is interacted with other characteristics of the same subject-year: the share of New Passed itself and its square (own interactions), the number of total teachers, the average entry scores of New Regulation teachers. Capped lines indicate 95% point-wise confidence intervals.



(a) Marginal effect at different levels of '*Share New Passed*'



(b) Marginal effect at different numbers of colleagues

7 Channels of the effect

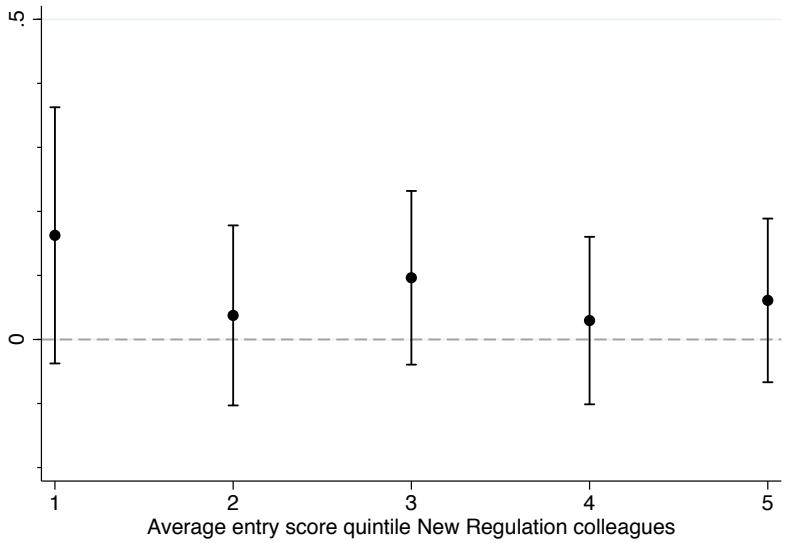
7.1 Selection on skills at entry

The public contest at entry is certainly the most prominent among the novelties that were introduced by the 2002 reform of the teacher career. Its purpose is allowing the most suitable candidates to become school teachers, while keeping the less desirable ones out of the profession - and seeking to measure suitability with transparent and meritocratic criteria, as opposed to the politicized selections and placements that were common before the reform. The first stage of the contest, the written exam that is meant to evaluate knowledge and teaching aptitudes, accounts for a major share of the total score for a candidate (see the contest structure in Table 1). The minimum score threshold required in the exam seems to represent a tough hurdle for many candidates, and the first four contests saw exam-passing rates of only about 30% on average (Table A.1). Keeping in mind that we found indication of a positive correlation between the average entry test scores of New Regulation teaches and student performance (Table A.9), the data does suggest that the entry exam, by cutting out an important fraction of candidates with low test scores, is one of the channels through which the positive impact of New Regulation teachers is materializing.

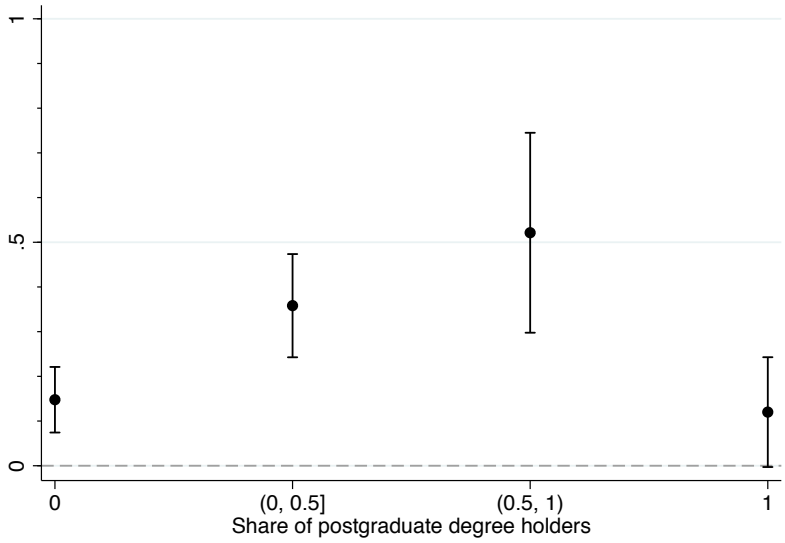
7.2 Selection on the probation period

The probation period that teachers begin after having successfully passed the entry contest and selected their preferred vacancy, is a further selection mechanism that the 2002 reform put in place. School headmasters observe and evaluate the new teachers during their first months of employment, and have the power to end their employment if deemed unfit for the job. Anecdotal evidence suggests that, in practice, the firing of teachers after their probation period is a rare event, due to the fact that headmasters do not wish to incur in the hassle of potential appeals and legal disputes that fired teachers would recur to. We test this claim using our six years of school censuses (2008 to 2013) and constructing an individual-level panel that records the type of position that each New Regulation teacher holds each year: permanent, temporary[25] and in probation. We then estimate a model $Y_{it} = \beta_0 + \beta_1 Temp_{it} + \beta_2 Prob_{it} + \epsilon_{it}$, where $Y_{it} = 1$ if the teacher is still

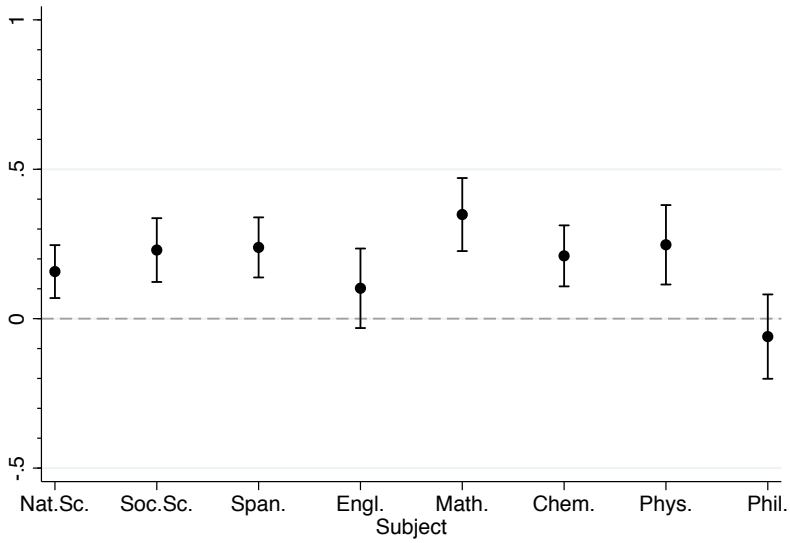
[25] There are two types of temporary statuses: 'Temporary in permanent vacancy', meaning that the vacancy is waiting to be filled with a *new* permanent occupant, and 'Temporary in temporary vacancy' means that the vacancy has a permanent occupant who is only temporarily away. We have grouped the two categories into a single one for our analysis.



(c) Marginal effect at different quintiles of average entry scores



(d) Marginal effect at different shares of postgraduate degree holders



(e) Marginal effect at subjects

in the panel in period $t+1$ and 0 otherwise[26]; $Temp$ and $Prob$ are mutually exclusive dummies indicating the type of position held in year t (omitted category is $Perm$, permanent position). We then also add the Age_{it} controls to the baseline (columns (2) and (4)). In Columns (3) and (4) we adopt a random effects specification, adding individual teacher effects α_i to our model. All columns report the odds ratios from our maximum-likelihood logit estimates. The results of Table 5 may tell a different story with respect to the previous anecdotal evidence: teachers on probation are only about 80% as likely as permanent position teachers to still be recorded system the year after. Nevertheless, we do not know the reasons behind the sample attrition and we are thus unable to distinguish between teachers who have voluntarily quit their job and those sacked. It may be that the first year of employment regularly sees higher dropout rates with respect to the following ones, due to initial adjustment or unmet expectations of new entrants. In conclusion, even though we do find considerably higher teacher dropout rates associated with their probation period, we are unable to conclusively say whether this provision of the reform is implementing selection on new teachers.

[26]We do not use $Y_{i,2013}$, which is equal to 0 for all subjects, since the panel ends in 2013.

Table 5: Panel retention by type of position held (New Regulation teachers)

	Logit		RE Logit	
	(1)	(2)	(3)	(4)
Temporary position	0.25***	0.25***	0.19***	0.19***
	(0.00)	(0.00)	(0.00)	(0.00)
Probation period	0.77***	0.77***	0.78***	0.78***
	(0.03)	(0.03)	(0.03)	(0.03)
Age bins	No	Yes	No	Yes
N.obs	138,869	138,865	138,869	138,865
N.groups			48,172	48,171

Note: Odds ratios displayed. Outcome variable: Y=1 if the teacher is still in the panel the following year, 0 otherwise. Year 2013 excluded. Columns (1) and (2): SE clustered by individual in parentheses. Columns (3) and (4): Observed Information Matrix SE in parentheses. * p<0.05, ** p<0.01, *** p<0.001

7.3 Turnover and discontinuation of employment

We can distinguish discontinuation of employment for Colombian public teachers into forced and voluntary. It is forced when the teacher is dismissed from his or her employment against his or her will, and it is voluntary when the teacher leaves the profession by own decision.

In section 3.4 we discussed how new-regulation teachers, contrary to their old-regulation colleagues, face the threat of seeing their employment discontinued for reasons other than severe misconduct. As an incentive to effort and good teaching, discontinuation of employment may happen due to two consecutive years of unsatisfactory scores on the performance evaluations which are carried out by school headmasters.

Unfortunately we do not have access to data on these yearly performance evaluations, but we can start exploring the question of whether new regulation teachers are experiencing significant screening even once their careers are underway by examining in-career dropout patterns. Table 6 shows maximum likelihood logit estimations of the model $Y_i = \beta_0 + \beta_1 NewReg_i + \epsilon_i$, where $Y_{it} = 1$ if the teacher is still in the panel in period $t+1$ and 0 otherwise; and $NewReg_i$ is a dummy taking value 1 for new-regulation teachers and 0 for old-regulation ones; again we also add the Age_i control to the baseline. In Columns (3)-(4) we adopt a random effects specification, adding individual teacher effects α_i to our model[27]. Conditional on having reached the permanent-position

[27] In this case fixed effects estimation is not feasible, as the new-regulation or old-regulation status is time-invariant.

stage, New Regulation teachers actually display higher panel retention rates than their traditional-regulation colleagues, even controlling for age[28]. In conclusion, we are unable to find any evidence for high rates of forced employment termination occurring among teacher subject to the reformed rules, and we are inclined towards ruling out this channel as an important selection mechanism.

Table 6: Panel retention per type of teacher regulation (permanent-position teachers)

	Logit			RE Logit		
	(1)	(2)	(3)	(4)	(5)	(6)
New Regulation	1.978***	1.516***	1.229***	2.409***	1.759***	1.313***
	(0.030)	(0.028)	(0.022)	(0.044)	(0.038)	(0.026)
Age		0.979***	0.972***		0.975***	0.971***
		(0.001)	(0.001)		(0.001)	(0.001)
Year dummies	No	No	Yes	No	No	Yes
N.obs	309,383	309,383	309,383	309,383	309,383	309,383
N.groups				94,285	94,285	94,285

Note: Outcome variable: Y=1 if the teacher is still in the panel the following year, 0 otherwise. Columns (1)-(3): SE clustered by individual in parentheses. Columns (4)-(6): Observed Information Matrix SE in parentheses. * p<0.05, ** p<0.01, *** p<0.001

7.4 New-Regulation teachers that have not passed the entry exam

Let us now turn to discuss the results we obtained on the second category of New Regulation teachers, those who have not passed their entry exam (the *SNNP* variable, for "Share New Not Passed"). In the main results of Table 3, column (5), we see that the positive impact of these teachers on student outcomes is estimated at around 70% of the impact of teachers who did pass the entry exam. Teachers who have not passed the exam can be employed temporarily, in absence of fully accredited candidates. Nevertheless, they have not officially entered the teaching career and are thus not eligible for any career upgrades: their salary is locked to the first step (step A) of the level to which they would belong if they passed the exam[29] (see Table A.2 for an idea about numbers). These teachers, the vast majority of which occupies temporary positions (see Table 2), is also at constant risk of being replaced by candidates who do pass the exam, as only the latter are entitled to take permanent possession of vacancies. The positive impact that this

[28] Various other flexible forms of controlling for age have been tested without significant changes in the results, and have thus been omitted in the output.

[29] Decree 624/2008.

category of teachers has on student test scores, thus still improving on Old Regulation teachers, can probably be attributed to the effort that they are exerting towards obtaining their full accreditation and overcoming their precarious status in the system. In this regard, it is important to keep in mind that even though they have obtained a score below 60, these teachers have attempted the entry exam at least once, and almost 60% of them have attempted it multiple times[30], and are thus likely to have gone through preparation and quality assurance procedures similarly to their successful colleagues. Moreover, they do possess the preliminary requirements needed to access the entry contest, including the higher education level requirements introduced by the reform. These considerations may explain quite well our finding of a still positive but about 30% lower impact on student test scores that these non-accredited teachers are bringing about, and are consistent with our previous deductions about a positive selection being implemented through the entry exam and possibly through the probation period.

8 Exploring time patterns

We further pursue the understanding of how the two regulations translate into career differences by analyzing survival patterns in Old and New public teachers. A public teacher becomes "at risk of failing" the year in which he/she enters the teacher profession, and "fails" the year in which he/she exits it for whichever reason. The survival analysis performed in this section is based on the individual-level information recorded over the direct observation period 2008-2013[31], and on the retrospective information about each teacher's first hiring year[32]. The survival time of teachers hired before the start of our observation window in 2008 is treated as conditional on having survived already up to that year[33]. In the whole analysis, the sample has been limited to exclude teachers who voluntarily switched from the Old to the New Regulation (by taking the exam) and the other cases in which the regulation recorded is inconsistent with the year of hiring (3,902 teachers), as well as the teachers who are recorded as not exercising in an educational structure (1,899 teachers). The remaining sample consists of 118,117 teachers teaching one of the Saber 11 test subjects at

[30] 7,392 out of 12,895 teachers have attempted the exam 2, 3 or 4 times between 2002 and 2009

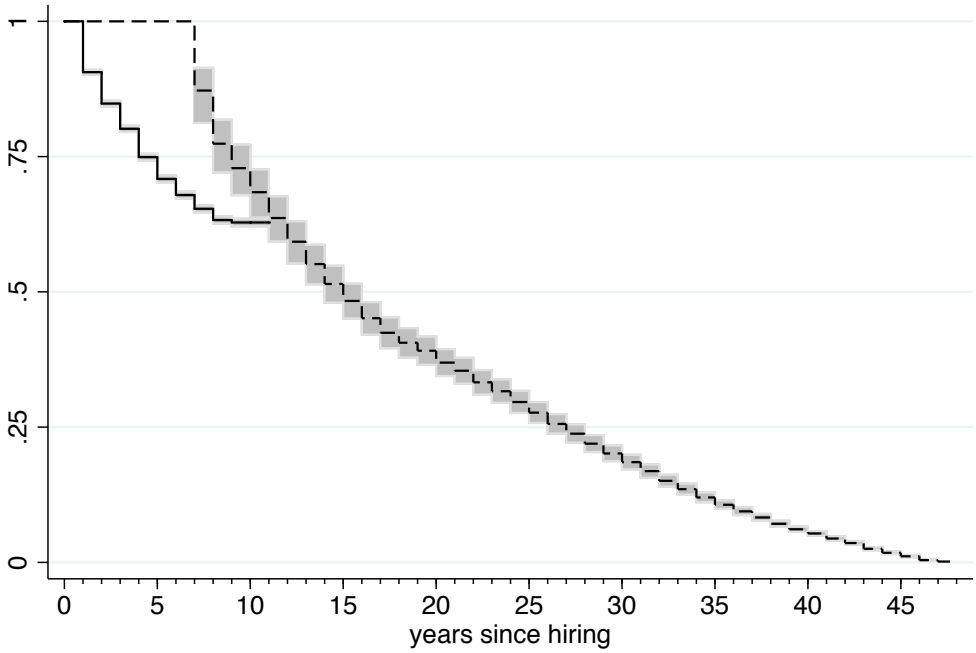
[31] Or a subset of those years, for teachers hired during that time span.

[32] The first hiring date is the date in which the teacher was hired for the first time as a public school teacher.

[33] i.e. we are in the presence of "left censoring". See the excellent discussion in Wooldridge [2010], ch. 22.3.3.

secondary level [34], among which 64,883 belong to the Old regulation and 53,234 to the New one (43,197 with matched test scores).

Figure 1: Kaplan-Meier survivor functions by regulation



Kaplan-Meier survivor functions for teachers belonging to the new regulation (solid line) and the old regulation (dashed line), and 95% confidence intervals.

As a first exercise, we estimate the survival functions corresponding to old regulation and new regulation teachers non-parametrically, and plot them in Figure A.2. Notice that since our observation window spans the years 2008-2013 and the new teacher regulation was implemented in 2002, any old-regulation teacher we observe has already spent at least 6 years in the public education system (which explains the 'shifted' starting point of the dashed curve in Figure A.2). On the other hand, the most senior new-regulation teachers we observe have spent not more than 11 years in the system (which explains the 'early' end of the solid curve in the figure). The two survival functions are therefore shifted with respect to each other, with few years of time overlap. This situation makes it difficult to compare the two survival patterns directly, since the most interesting comparisons

[34] In the Appendix, we also show the results of the analysis on the 360,644 teachers teaching at any preschool, primary and secondary level in the country. The patterns identified in the full sample follow closely those found in the limited one.

between groups are made looking at the same time from origin, or years since hiring in this case[35].

We further explore how observable characteristics of teachers are associated with differences in survival patterns. We show graphical results in Figure 2 and 3 for teachers belonging to the old and the new regulation respectively. Overall we do not find striking differences in the way teacher characteristics affect survival patterns under the two regulations. Holding a postgraduate degree is associated with significantly higher survival rates with respect to not holding one; being female and being located in a rural area associate with somewhat lower permanence in the system, especially under the new regulation; as it is natural, survival decreases as the age at which the teachers first entered the system. In the following subsection we will give special consideration to the patterns on teacher entry scores, depicted panel (e) of Figure 3.

8.1 Entry exam scores

A very interesting dimension to look at for the case of New Regulation teachers is their entry test score. We would expect teachers that score higher on the entry test to survive longer in the system, under the assumptions that 1) the new regulation is effective in rewarding skills and keeping highly skilled teachers in the system, and effective in eliminating unsatisfactory teachers and 2) the entry exam is a good measure of the desirable skills that the reform is aiming at. To the support of aspect 2), we shall recall the positive correlation we found between teacher entry scores and student performance in our heterogeneity analysis (see Table A.9). Regarding aspect 1), previous work by Ome [2012b, 2013] has evidenced how the 2002 reform has made the teaching profession more attractive to high skilled individuals, due to the change in the salary structure and the potentially quicker ascent to top salary levels[36].

We have divided teachers into quartiles[37] according to their performance in the most recent entry exam they took, and we plot the survival rates of the four groups in Panel (e) of Figure A.4. We notice how the first (lowest-scoring) quartile exits from the teacher profession at a dramatically faster rate with respect to the other three. All teachers in the first quartile and part of those in

[35]For example, looking at Figure A.2 one might conclude that new-regulation teachers experience a quite steep dropout pattern during their early years of career (say the first five), but one cannot tell whether the same applies to old-regulation teachers, as we do not observe them in such early years

[36]If we consider education level as a proxy for skills, our own data collection summarized in Table A.2 confirms this view for both individuals holding postsecondary degrees and those holding university (undergraduate or postgraduate) degrees: compared to the Old Regulation, the former are now enabled to reach salary levels double as high, and the latter face a potentially much quicker ascent to the top salaries.

[37]The score ranges defining quartiles are [0, 58.35], (58.35, 61.4], (61.4, 64.55] and (64.55,100].

Figure 2: Kaplan Meier survival functions by teacher characteristics (Old regulation)

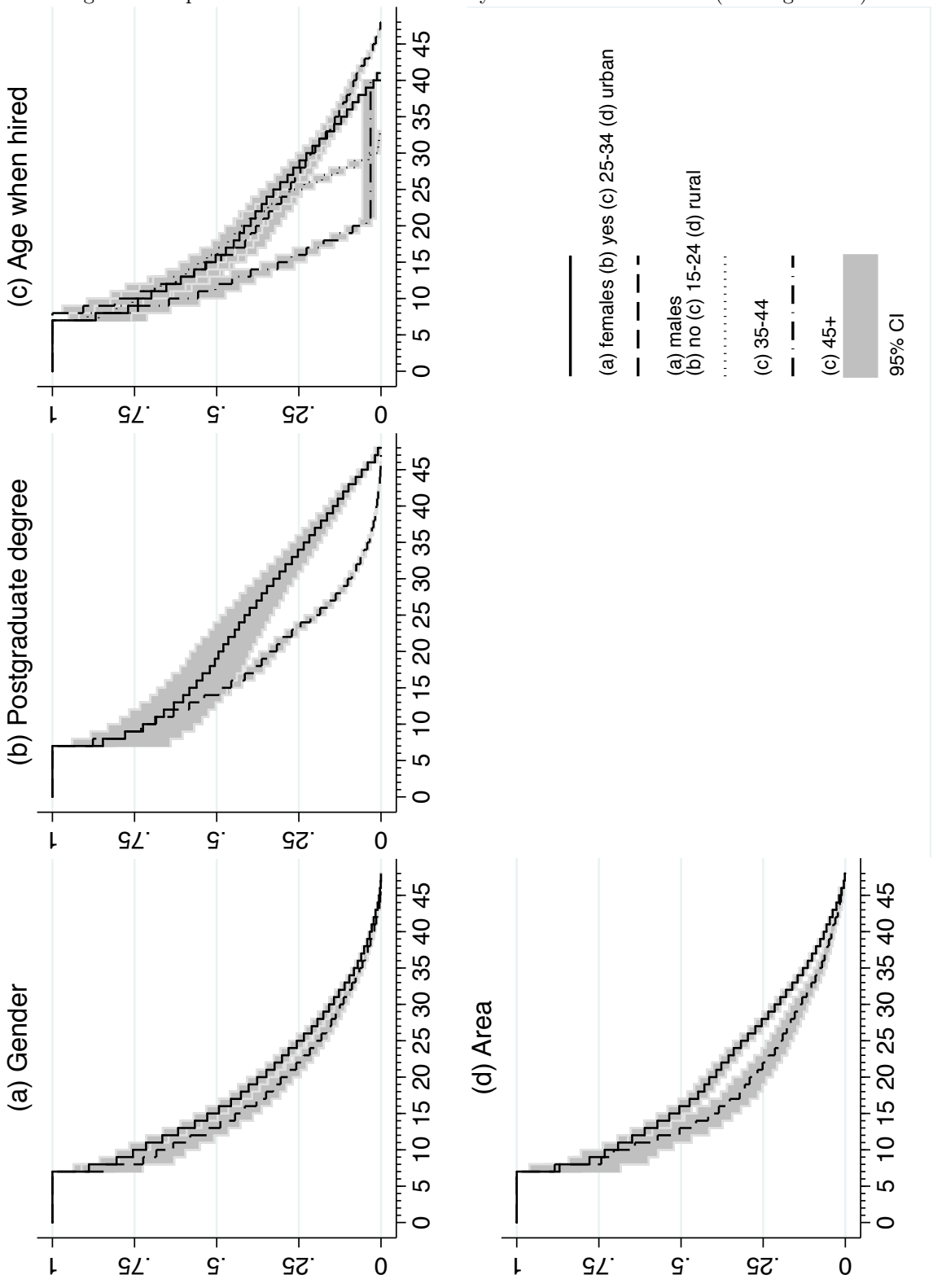
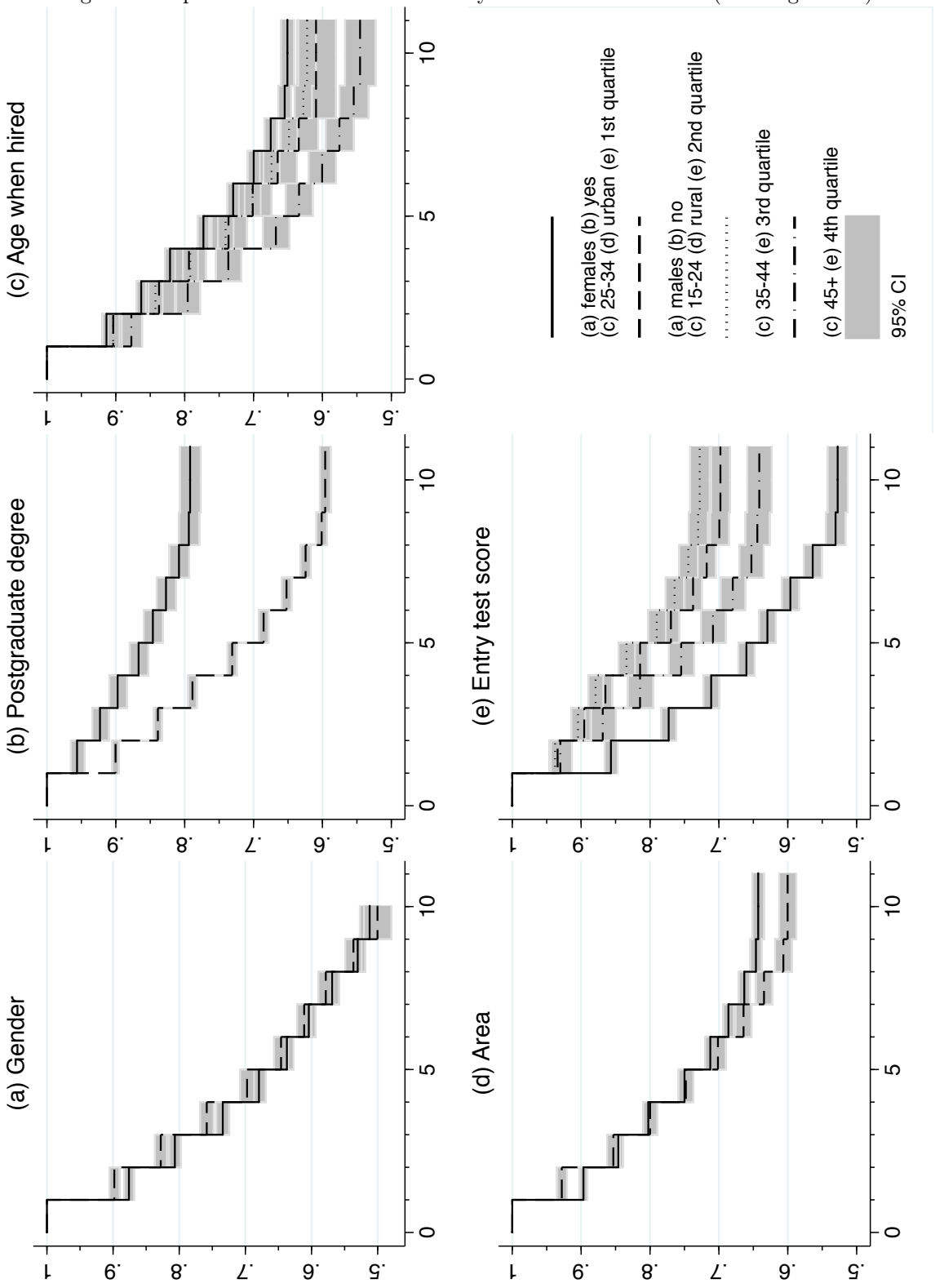


Figure 3: Kaplan Meier survival functions by teacher characteristics (New regulation)



the second quartile have scored below the 60 points required to pass the exam and proceed to the following stages of the entry context, on their most recent attempt. Back in Table 2 and in Section 7.4 described how a majority of these individuals has actually never passed the entry exam, expressing our concern about the significant share of teaching positions they occupy. The survival patterns we describe are consistent with those facts and provide us with some food for thought: on one hand, we find confirmation of the rather unstable nature of the employment histories of non-accredited New Regulation teachers, with respect to their better performing colleagues. On the other hand though, we notice that the instability may be lower than it ought to be, given that the survival chance of these individuals in the public school system is around 50% even after 10 years of employment.

A final interesting aspect to notice is that the relationship between exam score quartiles and expected permanence in the profession is non-monotonic, with the highest-scoring teachers showing lower survival rates with respect to their colleagues in the two middle quartiles. The analysis on score deciles in the Appendix (Table A.5) confirms this finding. It thus appears that even the more skill-rewarding career structure offered by the New Regulation has not yet succeeded at making the teaching profession as attractive to top-performers as for their more average colleagues.

9 Robustness checks

9.1 Additional time controls, and one teacher per subject

In this table we repeat our main estimation (Table 3) with additional cohort and experience. In particular, we add two different sets of hiring cohort dummies (Columns 1 and 3) and limit the sample to school-year-subject cells who do not contain any teachers with less than 5 or more than 40 years of experience (Columns 2 and 4). In these specifications, the effect of New-Regulation Passed teachers reduces between 20% and 43% with respect to our main results, remaining statistically significant throughout. The effect of New-Regulation Non-Passed teachers instead proves less robust and cannot be distinguished from zero. The final exercise (Column 5) uses only school-year-subject cells with only one teacher.

Table 7: Additional time controls, limiting experience, 1 teacher per subject

	Cohorts I	Cohorts I + limit exper.	Cohorts II	Cohorts II + limit exper.	1 teacher
Share New Reg. Passed	0.113*	0.127**	0.145*	0.162*	0.103
	(0.044)	(0.042)	(0.071)	(0.070)	(0.068)
Share New Reg. Not Passed	0.056	0.070	0.134	0.157	0.029
	(0.051)	(0.050)	(0.095)	(0.094)	(0.081)
Experience	0.032***	0.034***	0.057***	0.045***	0.017*
	(0.005)	(0.005)	(0.012)	(0.011)	(0.008)
Experience^2	–0.001***	–0.001***	–0.001***	–0.001***	–0.000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Age	0.028***	0.028***	0.021	0.022	0.013
	(0.009)	(0.009)	(0.020)	(0.020)	(0.013)
Age^2	–0.000***	–0.000***	–0.000	–0.000	–0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Share Postgrad degree	0.024	0.023	0.006	0.002	0.069
	(0.024)	(0.024)	(0.035)	(0.035)	(0.042)
Share hired post 1980	–0.119*		–0.105		
	(0.054)		(0.076)		
Share hired post 1990	–0.111*		–0.023		
	(0.053)		(0.081)		
Share hired post 2000	0.064		0.180*		
	(0.053)		(0.083)		
Share hired post 1985		–0.146*		–0.134	
		(0.057)		(0.081)	
Share hired post 1995		0.006		0.003	
		(0.037)		(0.051)	
Share hired post 2005		0.071		0.124	
		(0.043)		(0.067)	
Subject dummies	✓	✓	✓	✓	✓
Subject-specific trends	✓	✓	✓	✓	✓
Mean(y)	43.33	43.33	43.35	43.35	42.58
sd(y)	3.292	3.292	3.346	3.346	3.324
N.obs	151,178	151,178	74,114	74,114	62,621
N.groups	29,609	29,609	25,055	25,055	25,860
R-squared	0.79	0.79	0.83	0.83	0.80

Note: SE clustered by school in parentheses. Each observation is subject 's' in school 'i' in year 'y'. School-year fixed effects in all columns. * p<0.05, ** p<0.01, *** p<0.001

10 Conclusion

In this paper we estimate the effect that the 2002 reform of the public teacher career has had on the performance of Colombian high school students. The reform introduced exam-based selection at entry and a set of further quality incentives for teachers. We find positive and significant effects of New Regulation teachers on student test scores, in a magnitude of around 6% of a subject standard deviation increase in test scores when the share of these teachers goes from 0 to 1 in a given subject in a given school and year. When the share of New Regulation teachers in a subject is still low, increasing it yields a marginal effect up to three times as high the average one, and higher than average gains are also to be found in larger and more educated teacher groups.

After having explored heterogeneities in the effect and survival patterns of teachers in the education system, we are induced to conclude that the main channel through which the reform has brought about its positive results on student performance is the selection of teacher candidates at entry, which has been quite tight as around two thirds of initial candidates do not reach sufficiency at the exam stage. Selection at the initial probation period and selection on tenured teachers may also be contributing to raise the quality of surviving educators, but we are currently unable to quantify these contributions due to the absence of data on reasons for exit from teacher records.

Our analysis has also exposed a less successful side of the post-reform setting, namely the fact that around 30% of the New Regulation teachers recorded in the system over the 2008-2013 period are not fully accredited as they have not successfully passed the entry exam. They are employed in temporary positions but seem to persist in this status sometimes over several years. The intention of the law in allowing these types of temporary employments was to deal with exceptional circumstances in which vacancies need to be filled but no eligible candidates are at hand, but the dimension of the phenomenon appears to be larger than what would be justified by exceptional circumstances.

In terms of future research, we believe that it shall be highly interesting to look at New Regulation teachers in some year's time, when their employment histories will be longer and more informative on the incentives they face over their careers. The availability of data on the reasons for dismissal or voluntary departure, as well as data on voluntary formation courses attended by teachers, would be of immense value towards understanding to which extent the permanent evaluation aspect of the reform is effectively operating.

References

Raj Chetty, John N. Friedman, and Jonah E. Rockoff. Measuring the impacts of teacher i: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9):2633–2679, 2014.

Jesus Duarte. Política y educación: Tentaciones particularistas en la educación latinoamericana. *Economía Política de las Reformas Educativas en América Latina*, 2001.

Jesus Duarte. *Educación pública y clientelismo en Colombia*. Universidad de Antioquia, Medellín, 2003.

David N. Figlio. Can public schools buy better-qualified teachers. *Industrial and Labor Relations Review*, 55: 686, 2001. URL `http://heinonline.org/HOL/Page?handle=hein.journals/ialrr55&id=688&div=&collection=journals`.

Grupo GEARD. Explicación de las etapas del concurso para ingreso a la carrera docente 2013, 2013. URL `https://www.youtube.com/watch?v=r7iyLO3EvZA`.

Robert James Gordon, Thomas J. Kane, and Douglas Staiger. *Identifying effective teachers using performance on the job*. Brookings Institution Washington, DC, 2006. URL `https://books.google.com/books?hl=en&lr=&id=nOntgoDT4tkC&oi=fnd&pg=PA189&dq=identifying+effective+teachers+using+performance+on+the+job&ots=l73g7bXFYN&sig=nLCKr8AZkGWYAZJz3E_RVrQEmoo`.

Eric A. Hanushek. School policy: implications of recent research for human capital investments in south asia and other developing countries. *Education Economics*, 17(3):291–313, 2009. doi: 10.1080/09645290903142585. URL `http://www.tandfonline.com/doi/pdf/10.1080/09645290903142585`.

Eric A. Hanushek and Steven G. Rivkin. The distribution of teacher quality and implications for policy. SSRN Scholarly Paper ID 2139257, Social Science Research Network, Rochester, NY, July 2012. URL `http://hanushek.stanford.edu/sites/default/files/publications/Hanushek%2BRivkin%202012%20AnnRevEcon%204.pdf`.

Eric A. Hanushek and Ludger Woessmann. Do better schools lead to more growth? cognitive skills, economic outcomes, and causation. *Journal of Economic Growth*, 17(4):267–321, July 2012. ISSN 1381-4338, 1573-7020. doi: 10.1007/s10887-012-9081-x. URL `http://link.springer.com/article/10.1007/s10887-012-9081-x`.

Eric A. Hanushek, John F. Kain, and Steven G. Rivkin. Why public schools lose teachers. *Journal of human resources*, 39(2):326–354, 2004. URL `http://jhr.uwpress.org/content/XXXIX/2/326.short`.

Thomas J. Kane, Jonah E. Rockoff, and Douglas O. Staiger. What does certification tell us about teacher effectiveness? evidence from new york city. *Economics of Education Review*, 27(6):615–631, December 2008. ISSN 0272-7757. doi: 10.1016/j.econedurev.2007.05.005. URL `http://www.sciencedirect.com/science/article/pii/S0272775707000775`.

MEN. Decreto 3982 de 2006. *Ministerio de Educación Nacional; Diario Oficial*, 46449, November 2006.

MEN. Estatuto de profesionalizacion docente - informacion general y avances - abril de 2008, April 2008. URL `http://www.colombiaaprende.edu.co/html/productos/1685/articles-161031_archivo_3.pdf`.

MEN. Por meritocracia se seleccionarán 20.610 docentes, directivos docentes y etnoeducadores. Technical report, Ministerio de Educación Nacional; http://www.mineducacion.gov.co/cvn/1665/w3-article-312669.html [21 Jan 2015], October 2012. URL `http://www.mineducacion.gov.co/cvn/1665/w3-article-312669.html`.

MEN. Estadísticas concursos 2004 - 2009. Technical report, Ministerio de Educación Nacional;

http://www.mineducacion.gov.co/1621/w3-propertyvalue-48463.html [20 Jan 2015], May 2013a. URL http://www.mineducacion.gov.co/1621/w3-propertyvalue-48463.html.

MEN. ¿quiénes se reubican salarialmente o ascienden? Technical report, Ministerio de Educación Nacional; http://www.mineducacion.gov.co/proyectos/1737/w3-article-309814.html [22 Jan 2015], June 2013b. URL http://www.mineducacion.gov.co/proyectos/1737/w3-article-309814.html.

F. J. Murillo, A. González, and M. Rizo. Evaluación del desempeño y carrera profesional docente. una panorámica de américa y europa. *Santiago de Chile: UNESCO.[2ª Ed. Revisada]*, 2007.

OECD. *Education in Colombia*. Reviews of National Policies for Education. OECD Publishing, Paris, 2016.

Alejandro Ome. The effects of meritocracy for teachers in colombia. INFORMES DE INVESTIGACIÓN 010260, FEDESARROLLO, 2012a.

Alejandro Ome. Salarios de los docentes públicos en colombia 1995-2010. *COYUNTURA ECONÓMICA*, 2012b. URL http://econpapers.repec.org/article/col000438/011643.htm.

Alejandro Ome. El estatuto de profesionalización docente: una primera evaluación. CUADERNOS DE FEDESARROLLO, FEDESARROLLO, May 2013. URL http://econpapers.repec.org/paper/col000439/011553.htm.

Steven G. Rivkin, Eric A. Hanushek, and John F. Kain. Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458, 2005. ISSN 1468-0262. doi: 10.1111/j.1468-0262.2005.00584.x. URL http://www.econ.ucsb.edu/~jon/Econ230C/HanushekRivkin.pdf.

Jonah E. Rockoff. The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 5(1):247–252, 2005. URL https://www0.gsb.columbia.edu/faculty/jrockoff/rockoff_teachers_march_04.pdf.

Jonah E. Rockoff, Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger. Can you recognize an effective teacher when you recruit one? *Education Finance and Policy*, 6(1):43–74, January 2011. ISSN 1557-3060. doi: 10.1162/EDFP_a_00022. URL http://dx.doi.org/10.1162/EDFP_a_00022.

Jeffrey Wooldridge. *Econometric analysis of cross section and panel data*. MIT Press, 2010.

A Appendix

A.1 Data on the past entry contests

Table A.1: Selectivity of the entry contests

CONTESTS >	1st (2004)	2nd (2005)	3rd (2006)	4th (2009)
N. of local authorities	69	66	49	66
Vacancies	50.947	23.355	14.579	25.392
Candidates to exam	140.541	134.090	109.487	228.985
Passed exam stage	60.078 (43%)	32.720 (24%)	27.931 (26%)	66.687 (29%)
Assigned to vacancy	30.568 (22%)	14.092 (11%)	13.620 (12%)	39.468 (17%)

Note: all percentages are relative to 'Candidates to exam'
Source: MEN [2013]

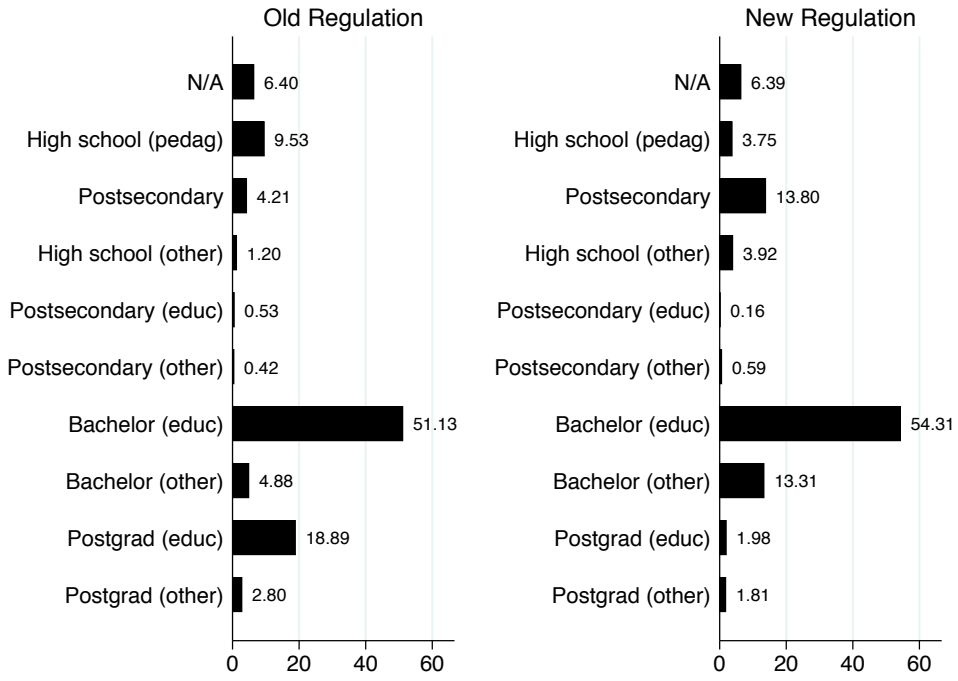
A.2 Career structure, salaries and education of teachers

Table A.2: Career structure of public school teachers, and 2008 pay scales

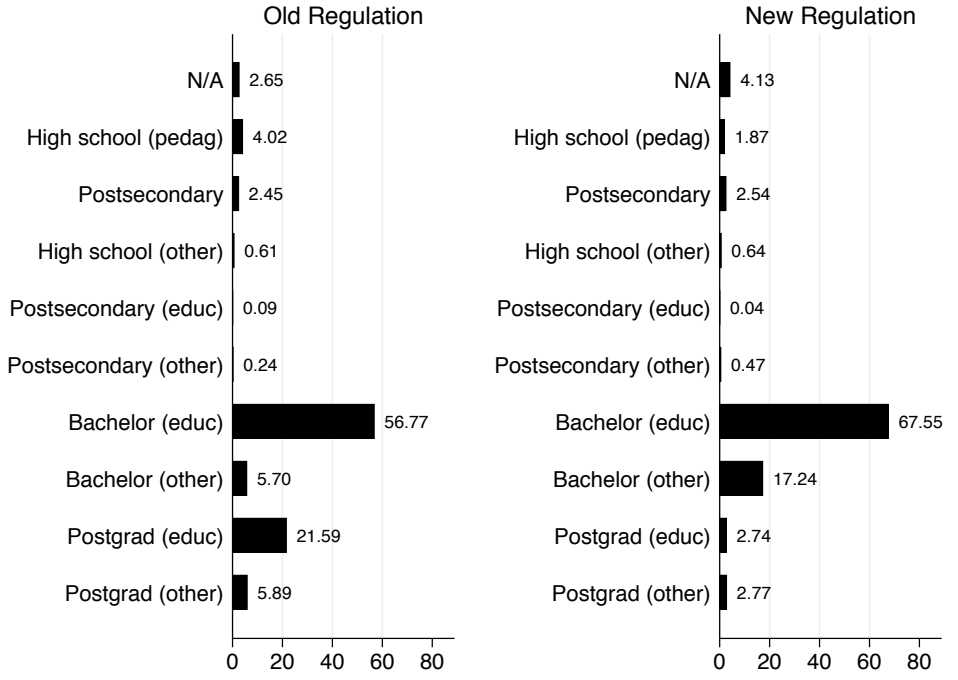
Old Regulation (Dec. 2277 / 1979)			**New Regulation (Dec. 1278 / 2002)**				
Step	Education level required	2008 salary	Level	Step	Education level required	2008 salary	
A	Secondary school	525.240	1	A	Postsecondary specialization	745.624	
B		581.850		B		1.014.611	
1		652.079		C		1.531.186	
2	Postsecondary specialization	675.922		D		1.759.188	
3		717.284	2	A	Undergraduate degree	938.340	
4		745.600		B		1.421.428	
5		792.628		C		1.834.801	
6	Undergraduate degree	838.439		D		1.980.454	
7		938.315				Master deg.	PhD deg.
8		1.030.680	3	A	Postgraduate degree	1.415.933	1.721.798
9		1.141.779		B		1.772.111	2.154.919
10		1.250.166		C		2.017.127	2.452.864
11		1.427.513		D		2.140.784	2.603.232
12		1.698.112					
13		1.879.682					
14	Postgraduate degree	2.140.766					

Source: compiled by the authors based on Decree 2277 / 1979, Decree 259 / 1981, Decree 626 / 2008, Decree 624 / 2008, MEN [2008]. Salaries in 2008 Colombian Pesos. The shaded steps are the possible entry steps for first-time teachers.

Figure A.1: Education level of public school teachers



(a) All teachers



(b) Secondary school teachers only

A.3 Descriptive statistics

Table A.3: Descriptive statistics at school-year-subject level

Mean student score	43.33
	(3.29)
Share New Regulation	0.47
	(0.43)
Share New Regulation Passed	0.29
	(0.38)
Share New Regulation Not Passed	0.18
	(0.34)
Share Old Regulation	0.53
	(0.43)
Mean age	44.20
	(8.31)
Mean experience	12.49
	(9.20)
Share postgraduate degree	0.21
	(0.34)
N	151,178

Note: Variable means and (standard deviations).

Table A.4: Individual-level teacher descriptives

	All teachers	Old Regulation	All New Regulation	New Regul. Passed	New Regul. Not Passed
Age	45.80	50.22	37.26	37.09	37.82
	(10.06)	(7.59)	(8.65)	(8.74)	(8.35)
Experience	15.57	21.92	3.30	3.53	2.55
	(11.75)	(9.38)	(2.50)	(2.51)	(2.31)
Female	0.66	0.67	0.63	0.61	0.70
	(0.47)	(0.47)	(0.48)	(0.49)	(0.46)
Postgrad degree	0.20	0.26	0.09	0.10	0.04
	(0.40)	(0.44)	(0.28)	(0.30)	(0.19)
Experience 5-40yrs	0.75	0.98	0.31	0.35	0.20
	(0.43)	(0.15)	(0.46)	(0.48)	(0.40)
Age when hired	30.23	28.30	33.96	33.55	35.28
	(8.12)	(7.19)	(8.49)	(8.49)	(8.37)
Rural area	0.30	0.23	0.43	0.39	0.54
	(0.46)	(0.42)	(0.49)	(0.49)	(0.50)
Permanent position	0.85	0.99	0.58	0.70	0.19
	(0.36)	(0.09)	(0.49)	(0.46)	(0.39)
Temporary position	0.11	0.01	0.32	0.19	0.74
	(0.32)	(0.08)	(0.47)	(0.39)	(0.44)
Probation position	0.03	0.00	0.09	0.10	0.07
	(0.18)	(0.04)	(0.29)	(0.30)	(0.25)
Most recent test score				63.69	55.32
				(3.87)	(3.52)
N	1743,339	1149,239	594,100	452,493	141,607
N teachers	360,644	214,920	145,724	108,735	36,989

Note: Variable means and (standard deviations).

Table A.5: Individual-level teacher descriptives - Secondary school teachers only

	All teachers	Old Regulation	All New Regulation	New Regul. Passed	New Regul. Not Passed
Age	45.45	50.90	37.47	37.34	37.92
	(9.99)	(7.20)	(7.93)	(7.81)	(8.29)
Experience	14.13	21.61	3.18	3.44	2.33
	(11.43)	(8.83)	(2.50)	(2.51)	(2.23)
Female	0.55	0.57	0.54	0.51	0.62
	(0.50)	(0.50)	(0.50)	(0.50)	(0.49)
Postgrad degree	0.23	0.32	0.10	0.11	0.04
	(0.42)	(0.47)	(0.30)	(0.32)	(0.20)
Experience 5-40yrs	0.71	0.98	0.30	0.34	0.17
	(0.46)	(0.13)	(0.46)	(0.47)	(0.38)
Age when hired	31.32	29.29	34.29	33.90	35.60
	(7.55)	(6.78)	(7.64)	(7.40)	(8.27)
Rural area	0.21	0.14	0.33	0.29	0.44
	(0.41)	(0.34)	(0.47)	(0.46)	(0.50)
Permanent position	0.83	1.00	0.58	0.72	0.13
	(0.38)	(0.07)	(0.49)	(0.45)	(0.33)
Temporary position	0.14	0.00	0.34	0.20	0.83
	(0.35)	(0.07)	(0.47)	(0.40)	(0.38)
Probation position	0.03	0.00	0.07	0.08	0.04
	(0.17)	(0.01)	(0.26)	(0.27)	(0.20)
Most recent test score				64.11	55.38
				(3.96)	(3.59)
N	437,570	259,850	177,720	136,974	40,746
N teachers	118,117	64,883	53,234	40,339	12,895

A.4 New Regulation teachers as a single group

In Table A.6 we repeat our main estimations considering New Regulation teachers as a single group, without distinguishing between those who have passed the entry contest and those who have not.

Table A.6: The effect of New Regulation teachers on student performance

	(1)	(2)	(3)	(4)	(5)
Share New Regulation	–0.95***	0.63***	0.09**	0.19***	0.19***
	(0.05)	(0.07)	(0.03)	(0.04)	(0.04)
Age		0.05***	0.03***	0.03***	0.03***
		(0.01)	(0.01)	(0.01)	(0.01)
Age^2		–0.00***	–0.00***	–0.00***	–0.00***
		(0.00)	(0.00)	(0.00)	(0.00)
Experience		0.15***	0.03***	0.03***	0.03***
		(0.01)	(0.00)	(0.00)	(0.00)
Experience^2		–0.00***	–0.00***	–0.00***	–0.00***
		(0.00)	(0.00)	(0.00)	(0.00)
Share postgrad degree		1.00***	0.01	0.02	0.02
		(0.06)	(0.02)	(0.02)	(0.02)
Subject FE	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
School FE			✓	✓	✓
School-year FE				✓	✓
Subject-specific trends					✓
Mean(y)	43.33	43.33	43.33	43.33	43.33
sd(y)	3.29	3.29	3.29	3.29	3.29
N.obs	151,178	151,178	151,178	151,178	151,178
N.groups	.	.	5,969	29,609	29,609
R-squared	0.17	0.19	0.68	0.79	0.79

Note: SE clustered by school in parentheses. Each observation is subject 's' in school 'i' in year 'y'. No fixed effects in columns (1) and (2), school fixed effects in column (3), school-year fixed effects in columns (4) and (5). * p<0.05, ** p<0.01, *** p<0.001

A.5 Heterogeneity Tables

These tables are behind the graphical results of panels (a)-(e) in Figure 2 in Section 6.2.

Table A.7: Interaction with number of teachers in the year-subject

	Student test scores	
Share New-R. Passed (SNP)	0.106**	(0.038)
Share New-R. Not Passed	0.071	(0.044)
2 teachers	0.017	(0.023)
3 teachers	0.113***	(0.032)
4 teachers	0.192***	(0.036)
5 teachers	0.205***	(0.041)
6 teachers	0.265***	(0.049)
7 teachers	0.296***	(0.055)
8 teachers	0.321***	(0.066)
9 teachers	0.462***	(0.085)
10 teachers	0.447***	(0.090)
11 teachers	0.474***	(0.105)
2 teachers * SNP	0.067	(0.045)
3 teachers * SNP	0.165**	(0.062)
4 teachers * SNP	0.184*	(0.073)
5 teachers * SNP	0.407***	(0.091)
6 teachers * SNP	0.482***	(0.108)
7 teachers * SNP	0.604***	(0.137)
8 teachers * SNP	0.593***	(0.170)
9 teachers * SNP	0.638**	(0.201)
10 teachers * SNP	0.495	(0.262)
11 teachers * SNP	0.533	(0.279)
Age, experience, postgrad	✓	
Subject FE	✓	
School FE	✓	
Year FE	✓	
School-year FE	✓	
Subject-specific trends	✓	
Mean(y)	43.30	
sd(y)	3.29	
N.obs	149,183	
N.groups	29,604	
R-squared	0.79	

Note: SE clustered by school in parentheses. Each observation is subject 's' in school 'i' in year 'y'. * p<0.05, ** p<0.01, *** p<0.001

Table A.8: Own interactions of 'Share New-Regulation Passed' in the year-subject

	Student test scores	
Share New-R. Passed (SNP)	0.755***	(0.186)
Share New-R. Not Passed	0.111*	(0.044)
SNP ^2	–1.093*	(0.546)
SNP ^3	0.495	(0.372)
Age, experience, postgrad	✓	
Subject FE	✓	
School FE	✓	
Year FE	✓	
School-year FE	✓	
Subject-specific trends	✓	
Mean(y)	43.30	
sd(y)	3.29	
N.obs	149,183	
N.groups	29,604	
R-squared	0.78	

Note: SE clustered by school in parentheses. Each observation is subject 's' in school 'i' in year 'y'. * p<0.05, ** p<0.01, *** p<0.001

Table A.9: Interaction with average teacher score in the year-subject

	Student test scores	
Share New-R. Passed (SNP)	0.163	(0.102)
Share New-R. Not Passed	0.105	(0.065)
2nd quintile	0.085*	(0.039)
3rd quintile	0.105*	(0.044)
4th quintile	0.181***	(0.046)
5th quintile	0.168***	(0.046)
2nd quintile * SNP	–0.125	(0.101)
3rd quintile * SNP	–0.066	(0.102)
4th quintile * SNP	–0.133	(0.105)
5th quintile * SNP	–0.102	(0.106)
Age, experience, postgrad	✓	
Subject FE	✓	
School FE	✓	
Year FE	✓	
School-year FE	✓	
Subject-specific trends	✓	
Mean(y)	43.40	
sd(y)	3.24	
N.obs	99,108	
N.groups	27,435	
R-squared	0.81	

Note: SE clustered by school in parentheses. Each observation is subject 's' in school 'i' in year 'y'. * p<0.05, ** p<0.01, *** p<0.001

Table A.10: Interaction with share of postgraduate degree holders in the year-subject

	Student test scores	
Share New-R. Passed (SNP)	0.148***	(0.037)
Share New-R. Not Passed	0.103*	(0.044)
Postgrad share (0, 0.5]	0.064**	(0.025)
Postgrad share (0.5, 1)	0.165***	(0.036)
Postgrad share = 1	0.001	(0.028)
Postgrad share (0, 0.5] * SNP	0.210***	(0.053)
Postgrad share (0.5, 1) * SNP	0.374***	(0.112)
Postgrad share = 1 * SNP	–0.028	(0.060)
Age, experience	✓	
Subject FE	✓	
School FE	✓	
Year FE	✓	
School-year FE	✓	
Subject-specific trends	✓	
Mean(y)	43.30	
sd(y)	3.29	
N.obs	149,183	
N.groups	29,604	
R-squared	0.78	

Note: SE clustered by school in parentheses. Each observation is subject 's' in school 'i' in year 'y'. * p<0.05, ** p<0.01, *** p<0.001

Table A.11: Interaction with subject dummies

	Student test scores	
Share New-R. Passed (SNP)	0.157***	(0.045)
Share New-R. Not Passed	0.143**	(0.044)
Social Sciences	–0.592***	(0.022)
Spanish	0.936***	(0.022)
English	–2.766***	(0.037)
Mathematics	–0.884***	(0.031)
Chemistry	0.151***	(0.026)
Physics	–1.092***	(0.038)
Philosophy	–4.271***	(0.037)
Social Sciences * SNP	0.073	(0.054)
Spanish * SNP	0.081	(0.052)
English * SNP	–0.055	(0.069)
Mathematics * SNP	0.191**	(0.062)
Chemistry * SNP	0.053	(0.053)
Physics * SNP	0.090	(0.069)
Philosophy * SNP	–0.217**	(0.073)
Age, experience, postgrad	✓	
School FE	✓	
Year FE	✓	
School-year FE	✓	
Subject-specific trends	✓	
Mean(y)	43.30	
sd(y)	3.29	
N.obs	149,183	
N.groups	29,604	
R-squared	0.78	

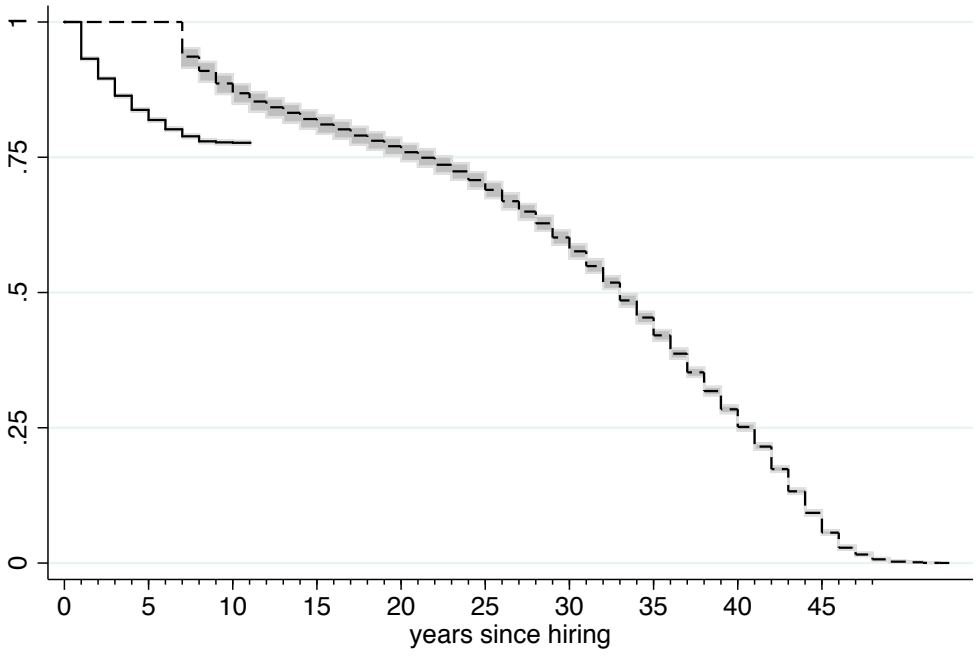
Note: Baseline subject is Natural Sciences. SE clustered by school in parentheses. Each observation is subject 's' in school 'i' in year 'y'. * p<0.05, ** p<0.01, *** p<0.001

A.6 Survival Analysis Extensions

A.6.1 All teachers

In this section we show the results of our survival analysis extending the sample to all 360,644 teachers recorded teaching in any public preschool, primary or secondary school in the country, among which 214,920 belong to the Old regulation and 145,724 to the New one. As in the previous analysis on secondary school teachers, we have excluded teachers who voluntarily switched from the Old to the New Regulation (by taking the exam) and the other cases in which the regulation recorded is inconsistent with the year of hiring, as well as the teachers who are recorded as not exercising in an educational structure.

Figure A.2: Kaplan-Meier survivor functions by regulation



Kaplan-Meier survivor functions for teachers belonging to the new regulation (solid line) and the old regulation (dashed line), and 95% confidence intervals.

A.6.2 Survival functions by entry score deciles

Given the interesting results described in Section 8.1 about survival patterns by entry exam scores, we expand the analysis by looking at score deciles instead of quartiles, in order to detect potentially

finer patterns. The analysis on deciles reflects the conclusions of the one on quartiles: teachers in the lowest three score deciles show the worst survival patterns, but at higher score levels the relationship between scores and expected survival is not monotonic, with survival improving at first, peaking at the mid of the distribution (5th decile) and then lowering somewhat, with the highest scores showing on average lower survival rates than mid-range scorers.

Figure A.3: Kaplan Meier survival functions by teacher characteristics (Old regulation)

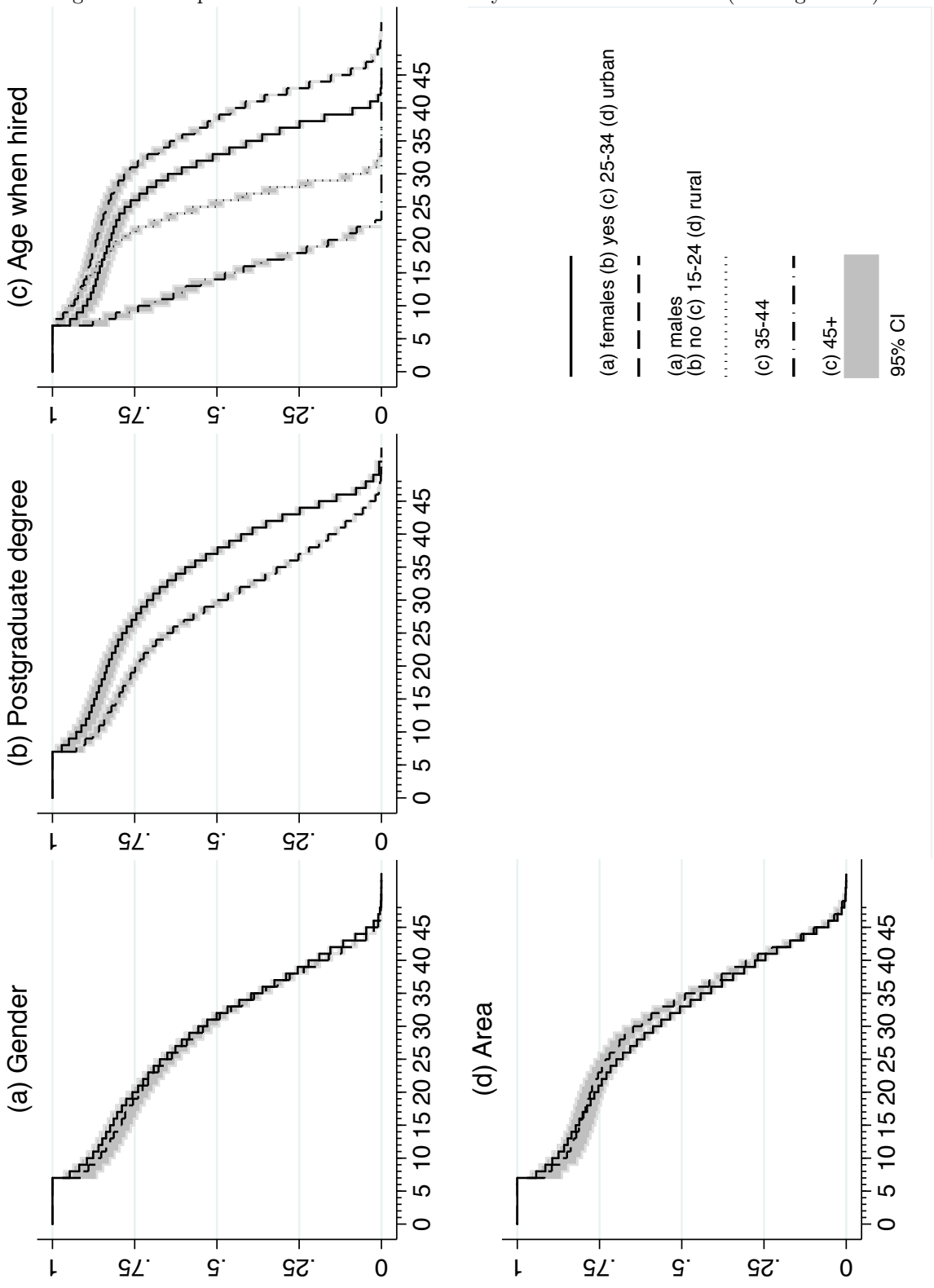


Figure A.4: Kaplan Meier survival functions by teacher characteristics (New regulation)

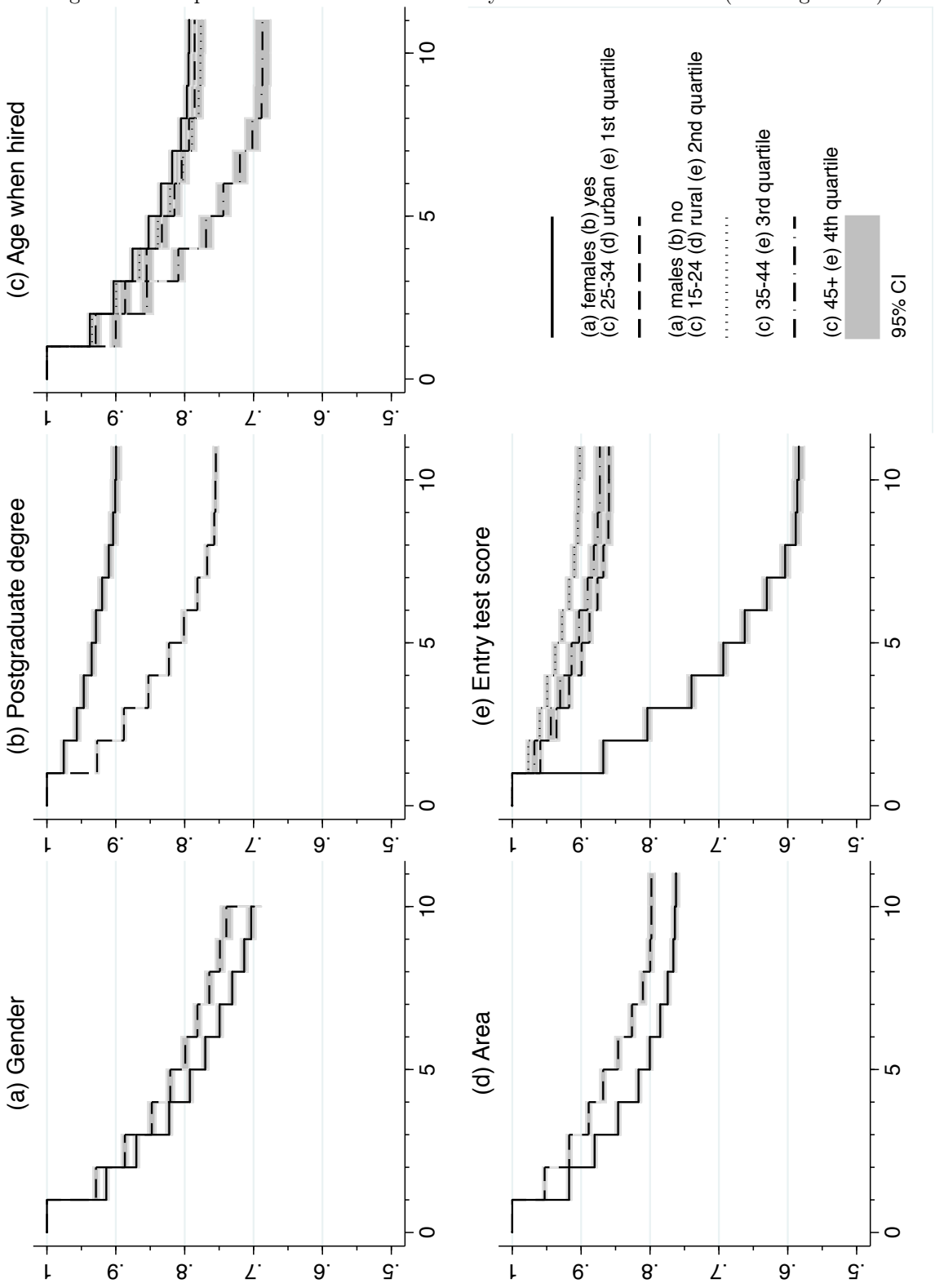
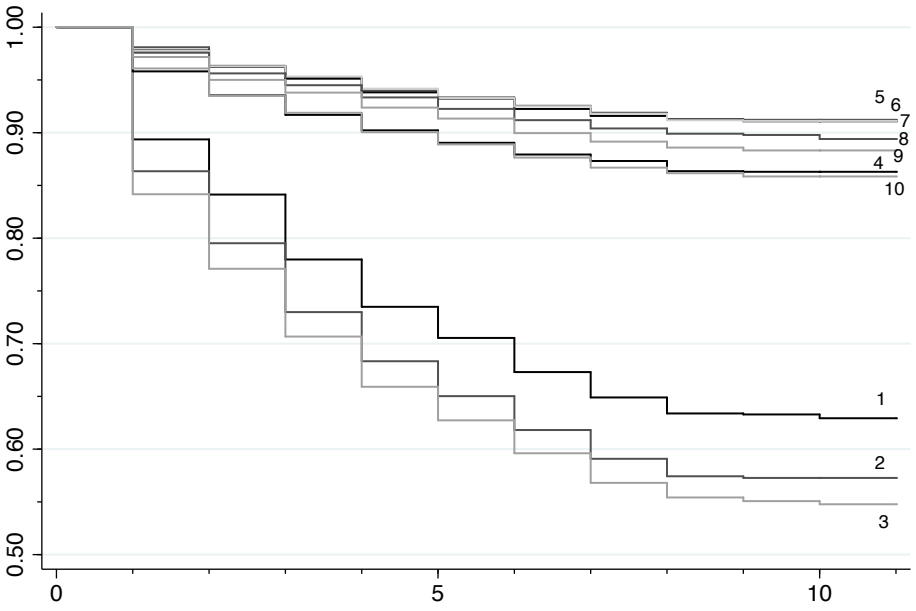
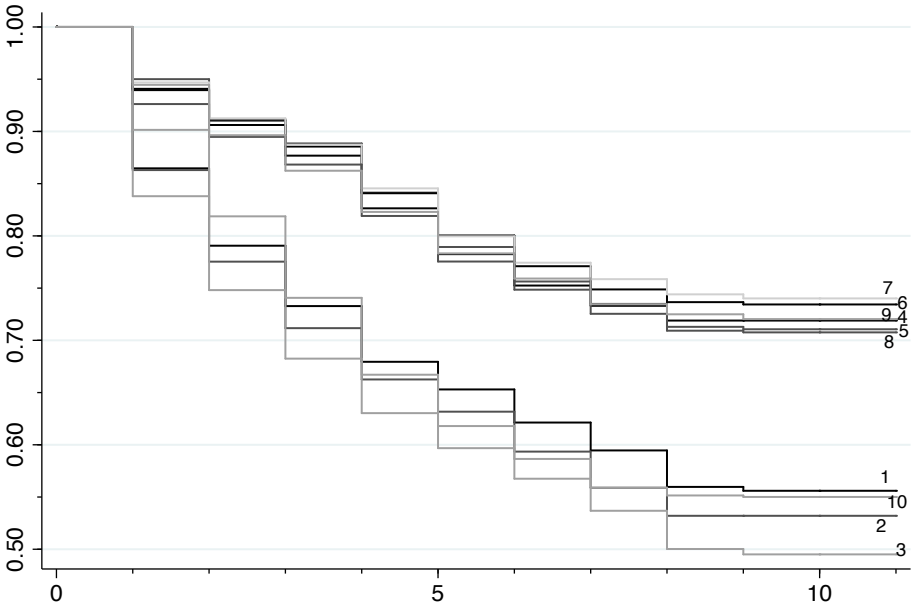


Figure A.5: Kaplan Meier survival functions by entry test score deciles



(a) All teachers



(b) Saber 11 teachers